SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 1, pp. 1–8

CHARACTERIZATIONS AND DECOMPOSITIONS OF ALMOST STRICTLY POSITIVE MATRICES*

M. GASCA† AND J. M. PEÑA†

Abstract. A nonsingular matrix is called almost strictly totally positive when all its minors are nonnegative, and furthermore these minors are positive if and only if their diagonal entries are positive. In this paper we give a characterization of these matrices in terms of the positivity of a very reduced number of their minors (which are called boundary minors), improving previous characterizations that have appeared in the literature. We show the role of boundary minors in accurate computations with almost strictly totally positive matrices. Moreover, we analyze the QR factorization of these matrices, showing the differences and analogies with that of totally positive matrices.

Key words. total positivity, QR factorization, almost strictly totally positive matrices

AMS subject classifications. 65F25, 65F40, 15A48

DOI. 10.1137/050631203

1. Introduction and basic notation.

Matrices with all minors nonnegative (in particular, all positive) have attracted much interest in several branches of mathematics and their applications, including computer aided geometric design, combinatorics, and economics. Unfortunately there is not an agreement to use a unified terminology for them. On one hand, the American school, following Schoenberg and especially Karlin and his book [16], used to call totally positive matrices those matrices with all minors nonnegative and strictly totally positive matrices the ones with all minors positive. Many authors, including Ando, de Boor, and Pinkus, have followed these names in the second half of the last century, and so have we in our papers on this subject. On the other hand, the German school used the terms "totally nonnegative" and "totally positive" matrices instead of the above ones, respectively. These last terms have become more accepted in the recent literature. Due to all of this, the term "totally positive matrix" has become ambiguous because it is used in two slightly (but significantly) different senses.

As we have said above, in the last decade we have used Karlin's terminology in our papers, and so it would be even more confusing to change to the other terminology in the present paper because, as we explain below, it improves some of our previous results and we make frequent references to these results. Consequently, in this paper, we continue calling those matrices with all minors nonnegative and those with all minors positive and hope this will cause no confusion to the reader. In any case, it would be good to unify terminology in the future.

For some important applications, for example, B-splines [2], interpolation [4], Hurwitz matrices [7, 17], or interval mathematics [8], the most important class of TP matrices is that which we called in [9] (referred to as ASTP matrices in the rest of this paper). This class is formed by TP

*Received by the editors May 11, 2005; accepted for publication (in revised form) by R. Bhatia July 18, 2005; published electronically March 17, 2006. This research was partially supported by the Spanish Research Grant BFM2003-03510 and by Gobierno de Aragón and Fondo Social Europeo.
http://www.siam.org/journals/simax/28-1/63120.html
†Departamento de Matemática Aplicada, Universidad de Zaragoza, 50009 Zaragoza, Spain (gasca@unizar.es, jmpena@unizar.es).

matrices whose minors are positive if and only if they do not contain a zero element in the diagonal. This class is intermediate between TP and STP matrices. The most interesting ASTP matrices are the nonsingular ones, and therefore, in the rest of the paper, we deal with these matrices. They have been called in [15] [illegible]. Matrices called in some papers [1] Δ-STP matrices are examples of triangular ASTP matrices.

From the beginning of the study of TP matrices it has been known that it is not necessary to check the sign of all minors of a matrix to decide whether or not it is totally positive and analogously for strict total positivity. Some of our efforts in the last decade have been devoted to getting criteria which decrease the number of minors to be checked and other characterizations in terms of bidiagonal factorizations [10, 11, 12, 14]. So we did this with ASTP matrices too. The nonzero pattern of these matrices [9, 13, 15] always has a staircase form. Roughly speaking (it will be explained more precisely in section 2) we proved [13, Theorem 3.1] that for a nonnegative matrix to be nonsingular ASTP we have to check only that minors formed with consecutive rows and columns, with the first row or column of the minor being the first row or column of a stair (of the nonzero pattern), are nonnegative and that they are positive if and only if the diagonal entries of the minor are all positive. These minors form a subclass of those called in [15, Theorem 2.1] inner minors with consecutive rows and columns, which are the minors to be checked in that paper. See also Theorem 3.1 of [9].

In this paper we improve our characterization of nonsingular ASTP matrices of [13] in the sense that the number of minors to be checked can be decreased. We introduce in section 2 the concept of boundary minor, which has special interest in matrices with staircase nonzero pattern, and prove that only these minors should be checked. Since they are a subclass of the ones used in [13], we decrease considerably the number with respect to [9, 15]. Moreover, we show how boundary minors can play a role in accurate computations with nonsingular ASTP matrices.

In the process of proving these results we have realized that in Theorem 3.1 of [13] the assumption of nonnegativity of the matrix can be suppressed: it is a consequence of any of the two equivalent properties of the theorem. So we have taken into account this fact in Theorem 2.4 of section 2 which is the new, improved version of that theorem.

After getting some results on the LU factorization of TP matrices we studied their QR factorization in [11]. In [13] we provided a bidiagonal factorization of nonsingular ASTP matrices and also the result that a nonsingular matrix A is ASTP if and only if it can be factorized LU with L and U ASTP matrices. It seems natural to study now the QR factorization of ASTP matrices to know if it has some peculiarities with respect to the general class of TP matrices. In section 3 we show the differences and analogies of the QR factorization of nonsingular ASTP matrices with respect to that of nonsingular TP and STP matrices. Boundary minors play again a crucial role in the proofs of that section.

2. Boundary submatrices of ASTP matrices. For k, n positive integers, $1 \leq k \leq n$, $Q_{k,n}$ will denote the set of all increasing sequences of k natural numbers less than or equal to n. For $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$, $\beta = (\beta_1, \beta_2, \ldots, \beta_k) \in Q_{k,n}$, and A an $n \times n$ real matrix, we denote by $A[\alpha|\beta]$ the $k \times k$ submatrix of A containing rows $\alpha_1, \ldots, \alpha_k$ and columns $\beta_1, \ldots, \beta_k$ of A. $Q^0_{k,n}$ will denote the set of sequences of k [illegible] natural numbers less than or equal to n.

By the [illegible] (see [3, Lemma A]), a nonsingular ASTP matrix $A =$

$(a_{ij})_{1\le i,j\le n}$ satisfies

$$\begin{aligned} &a_{ij}=0, i>j \Rightarrow a_{hk}=0 \quad \forall h\ge i, k\le j,\\ &a_{ij}=0, i<j \Rightarrow a_{hk}=0 \quad \forall h\le i, k\ge j. \end{aligned} \tag{2.1}$$

Moreover, it cannot have zero diagonal entries due to its nonsingularity (cf. [1, Corollary 3.8]):

$$a_{ii}\neq 0, \quad i=1,\dots,n. \tag{2.2}$$

Properties (2.1) and (2.2) produce a staircase form for the zero pattern of A, which will be made precise in the following notation, as in [13].

For an $n\times n$ matrix A let us denote

$$\begin{aligned} &i_0=1, \quad j_0=1;\\ &\text{for} \quad t=1,\dots,l:\\ &\qquad i_t=\max\{i|a_{i,j_{t-1}}\neq 0\}+1 \quad (\le n+1),\\ &\qquad j_t=\max\{j|a_{i_t,j}=0\}+1 \quad (\le n+1), \end{aligned}$$

where l is given in this recurrent definition by $i_l=n+1$. Analogously we denote

$$\begin{aligned} &\hat{j}_0=1, \quad \hat{i}_0=1;\\ &\text{for} \quad t=1,\dots,r:\\ &\qquad \hat{j}_t=\max\{j|a_{\hat{i}_{t-1},j}\neq 0\}+1,\\ &\qquad \hat{i}_t=\max\{i|a_{i,\hat{j}_t}=0\}+1, \end{aligned}$$

where $\hat{j}_r=n+1$. In other words, the entries below the places (i_1-1,j) with $j_0\le j<j_1$, (i_2-1,j) with $j_1\le j<j_2$, …, $(i_{l-1}-1,j)$ with $j_{l-2}\le j<j_{l-1}$ are zero. So are the entries to the right of the places $(i,\hat{j}_1-1)$ with $\hat{i}_0\le i<\hat{i}_1$, $(i,\hat{j}_2-1)$ with $\hat{i}_1\le i<\hat{i}_2$, …, $(i,\hat{j}_{r-1}-1)$ with $\hat{i}_{r-2}\le i<\hat{i}_{r-1}$.

When the matrix A is nonsingular ASTP, by (2.1), the remaining elements of A are nonzero. We shall express this by saying that the matrix A has a zero pattern given by $I=\{i_0,i_1,\dots,i_l\}$, $J=\{j_0,j_1,\dots,j_l\}$, $\hat{I}=\{\hat{i}_0,\hat{i}_1,\dots,\hat{i}_r\}$, and $\hat{J}=\{\hat{j}_0,\hat{j}_1,\dots,\hat{j}_r\}$. Only matrices with these patterns of zeros and all the other entries positive can be nonsingular ASTP.

Observe that, for a nonsingular ASTP matrix, by (2.2) we have necessarily

$$\begin{aligned} &i_t\ge j_t, \quad t=1,\dots,l-1,\\ &\hat{j}_t\ge \hat{i}_t, \quad t=1,\dots,r-1. \end{aligned} \tag{2.3}$$

In formula (3.2) of [13], the previous inequalities appeared strict, but in fact the equalities can also appear.

Lemma 2.1. Given any matrix $A=(a_{ij})_{1\le i,j\le n}$, it is easy to deduce that the following properties are equivalent:

(i) A satisfies (2.1) and (2.2).

(ii) A has a zero pattern given by $I,J,\hat{I},\hat{J}$ as above satisfying (2.3).

The submatrices introduced in the following definition are relevant in the context of matrices with a staircase zero pattern and will play a key role in this paper.

Definition 2.2. [illegible] $n\times n$ [illegible] A [illegible] $B:=A[\alpha|\beta]$ [illegible] $\alpha,\beta\in Q^0_{k,n}$ [illegible] $a_{\alpha_1,\beta_1}\cdots a_{\alpha_k,\beta_k}\neq 0$ [illegible] B [illegible] column boundary [illegible] $\beta_1=1$ [illegible]

$\beta_1 > 1$ $A[\alpha|\beta_1 - 1] = 0$ B row boundary
$\alpha_1 = 1$, $\alpha_1 > 1$ $A[\alpha_1 - 1|\beta] = 0$

Minors corresponding to column or row boundary submatrices are called, respectively, column or row boundary minors.

2.3. Using staircase notation, we can easily identify the boundary submatrices for matrices satisfying the zero pattern described above. Let $A = (a_{ij})_{1\le i,j\le n}$ be an $n\times n$ matrix with a zero pattern given by $I, J, \hat{I}, \hat{J}$ satisfying (2.3). Let $B := A[\alpha|\beta]$ with $\alpha, \beta \in Q^0_{k,n}$ and $a_{\alpha_1,\beta_1}\cdots a_{\alpha_k,\beta_k} \neq 0$. Then B is a column boundary submatrix if there exists $k \ge 1$ such that $\beta_1 = j_k$ and $\alpha_1 \ge i_k$. B is a row boundary submatrix if there exists $k \ge 1$ such that $\alpha_1 = \hat{j}_k$ and $\beta_1 \ge \hat{i}_k$. The leading principal minors of A are column and row boundary minors of it.

Let us consider an example of a 5×5 matrix A with $l = 2$, $r = 1$, $\{i_0, i_1, i_2\} = \{1, 4, 6\}$, $\{j_0, j_1, j_2\} = \{1, 3, 6\}$, $\{\hat{j}_0, \hat{j}_1\} = \{1, 6\}$, and $\{\hat{i}_0, \hat{i}_1\} = \{1, 6\}$. Entries represented by the symbol * are nonzero. The row boundary minors of the matrix

$$A = \begin{pmatrix} * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \end{pmatrix} \tag{2.4}$$

are the minors using initial consecutive rows and consecutive columns. The column boundary minors of A are its leading principal minors, the entries $a_{21}, a_{31}, a_{43}, a_{53}$, the minors

$$\det A[2, 3|1, 2], \quad \det A[4, 5|3, 4], \tag{2.5}$$

and the following minors which can be obtained from the previous ones: the minor $\det A[2, 3, 4|1, 2, 3]$ (which is equal to $a_{43} \det A[2, 3|1, 2]$) and $\det A[2, 3, 4, 5|1, 2, 3, 4]$ (which coincides with $\det A[2, 3|1, 2] \det A[4, 5|3, 4]$).

Now we shall prove that, for a matrix A, being nonsingular ASTP depends only on the sign of the boundary minors, improving the characterization of Theorem 3.1 of [13]. In addition, as said in section 1, we point out that the hypothesis of nonnegativity of A used in that theorem is not necessary because it is a consequence of any of the two equivalent properties of the theorem.

THEOREM 2.4. $A = (a_{ij})_{1\le i,j\le n}$ (2.1) (2.2)

(i) A
(ii) A

. By definition of nonsingular ASTP matrices, (i) implies (ii). For the converse, take into account that, by definition, A is a (trivial) boundary submatrix of itself, and consequently it is nonsingular. Now, the arguments of the proof of the converse part of Theorem 3.1 of [13] can be applied. Let us sketch the main points of that proof. It consists of showing that the Neville elimination of A and A^T can be performed without row or column exchanges and with nonnegative pivots which are zero if and only if they lie in the zero pattern of A, which by Remark 2.1 is given by $I, J, \hat{I}, \hat{J}$ as above. If we take a column j with $j_{t-1} \le j < j_t$, the crucial point of the proof of Theorem 3.1 of [13] is to show the positivity of the quotients

$$\frac{\det A[i - j + j_k, \ldots, i - 1, i|j_k, \ldots, j - 1, j]}{\det A[i - j + j_k, \ldots, i - 1|j_k, \ldots, j - 1]}, \quad i = j, j + 1, \ldots, i_t - 1, \tag{2.6}$$

where

$$j_k = \max\{j_s \leq j | 0 \leq s \leq t-1, j - j_s \leq i - i_s\}. \tag{2.7}$$

In fact, in that proof it is shown that the numerator and denominator of (2.6) are positive. Observe that for $j = j_{t-1}$ and $i = i_{t-1}, \ldots, i_t - 1$ we have $j_k = j_{t-1}$, and the quotient above becomes simply a_{ij}. Let us also point out that, by (2.3), in (2.6) one has $i_t \geq j_t$.

Coming back to our present theorem, the same arguments lead us to consider the quotients (2.6). Now, taking into account that, by (2.7), $j - j_k \leq i - i_k$, we have

$$i - j + j_k = i - (j - j_k) \geq i - (i - i_k) = i_k.$$

So, the submatrices of the numerator and denominator of (2.6) are of the form $A[\alpha|\beta]$ with $\alpha, \beta \in Q^0_{k,n}$, $\beta_1 = j_k$, and $\alpha_1 \geq i_k$, and, consequently, they are column boundary submatrices by Remark 2.3. Then (ii) implies that these minors are positive and the arguments of Theorem 3.1 of [13] to prove their positivity are not needed.

Since similar reasoning can be applied to A^T, the positivity of the row boundary minors is also involved.

In summary, the proof of Theorem 3.1 of [13] has been simplified, pointing out that the positivity of all boundary minors of A implies that A is a nonsingular ASTP matrix. □

If we apply the previous theorem to the matrix A of (2.4) in order to know if it is nonsingular ASTP, we have to check the positivity of the minors using initial consecutive rows and consecutive columns (row boundary minors), the elements $a_{21}, a_{31}, a_{43}, a_{53}$, and the two minors given by (2.5). If we apply Theorem 3.1 of [13], we should also check, in addition to all the above minors, the positivity of the entries a_{13}, a_{23}, a_{33} and of the following four minors:

$$\det A[2,3|3,4],\ \det A[3,4],\ \det A[2,3,4|3,4,5],\ \det A[3,4,5].$$

Finally, if we apply the characterization given in [15, Theorem 2.1] and [9], we should check the positivity of the remaining nonzero entries of A and of the following six minors, in addition to all of the previous ones: $\det A[2,3]$, $\det A[2,3,4]$, $\det A[2,3,4,5|2,3,4,5]$, $\det A[2,3|4,5]$, $\det A[3,4|4,5]$, $\det A[4,5]$. In larger matrices, the differences in the number of minors to be checked easily increase.

Given an algebraic expression defined by additions, subtractions, multiplications, and divisions and assuming that each initial real datum is known to high relative accuracy (see p. 52 of [5]), then it is well known that the algebraic expression can be computed accurately if it is defined by sums of numbers of the same sign, products, and quotients. In other words, the only "forbidden" operation is true subtraction, due to possible cancellation in leading digits. From now on, we will use the word [illegible] to mean [illegible]. Let us recall that a nonsingular TP matrix admits a unique factorization as a product of nonnegative bidiagonal, unit diagonal matrices and a diagonal matrix (see [12] or [14]). This factorization has been called recently in [6] and [18] [illegible] of A and is denoted by $\mathcal{BD}(A)$. Moreover, the property of A being nonsingular ASTP or not can be decided by $\mathcal{BD}(A)$ as can be seen in Theorem 4.1 of [13].

In [18] it is shown that an accurate bidiagonal decomposition of a nonsingular TP matrix A allows us to determine its eigenvalues and singular value decomposition to high relative accuracy. The following result proves that the accurate computation

of the boundary minors of A guarantees an accurate bidiagonal decomposition of a nonsingular ASTP matrix A. For the sake of brevity, we refer to [12] and [14] instead of introducing all notation related to Neville elimination.

PROPOSITION 2.5. A A $\mathcal{BD}(A)$. As can be seen in [12] or in section 2 of [14], the diagonal entries of the diagonal factor of $\mathcal{BD}(A)$ are the diagonal pivots of the Neville elimination of A. The nonzero off-diagonal entries of the bidiagonal factors of $\mathcal{BD}(A)$ are the multipliers of the Neville elimination of A or of A^T (see p. 116 of [14]) and, by formula (2.7) of [12], they are quotients of pivots of the Neville elimination of A or A^T. Since the pivots of the Neville elimination of A are given by (2.6) (see (2.3) of [12]), they are quotients of column boundary minors of A, and, analogously, the pivots of the Neville elimination of A^T are quotients of row boundary minors of A. Then we conclude that all pivots and multipliers can be computed accurately and the result follows. ∎

3. *QR* factorization of nonsingular ASTP matrices. In [11], nonsingular TP matrices and STP matrices were characterized in terms of their QR factorization. Now we are going to study that factorization for nonsingular ASTP matrices and show its peculiarity with respect to the other classes.

In this section, L (resp., U) represents a lower (upper) triangular, unit diagonal matrix, and D represents a diagonal matrix. Let us recall that, by Corollary 4.2 of [13], a nonsingular matrix A is ASTP if and only if it can be factorized as $A = LDU$ with L, U ASTP matrices and D a diagonal matrix with positive diagonal entries. Now we define a new class of matrices containing ASTP matrices.

DEFINITION 3.1. A lowerly ASTP $A = LDU$ LD

The following proposition characterizes lowerly ASTP matrices.

PROPOSITION 3.2. $n \times n$ A. If A is lowerly ASTP, then A can be factorized as $A = LDU$ with LD ASTP. Hence, all column boundary minors of LD are positive. Since U is an upper triangular matrix with unit diagonal, it is easy to see that rows and columns involved in the column boundary submatrices of A are the same as those of the column boundary submatrices of LD and that the column boundary minors of A have the same value as the corresponding column boundary minors of LD. So, all column boundary minors of A are positive.

For the converse, if all column boundary minors of A are positive, in particular, the leading principal minors of A are positive. So A can be decomposed as $A = LDU$. Again the column boundary minors of LD have the same value as those of A, and so they are positive. The row boundary minors of the lower triangular matrix LD are principal minors of LD using consecutive rows and columns, that is, of the form $(LD)[k, k+1, \ldots, k+r]$ $(1 \le k \le n,\ 0 \le r \le n-k)$. Using Schur complements, we have

$$\det(LD)[k, k+1, \ldots, k+r] = \frac{\det A[1, 2, \ldots, k+r]}{\det A[1, 2, \ldots, k]}. \tag{3.1}$$

Since the numerator and the denominator of (3.1) are column boundary minors of A, they are positive, and so the row boundary minors of LD are positive. Then, by Theorem 2.4, LD is ASTP and A is lowerly ASTP. ∎

The following definition will be used in the QR decomposition of nonsingular ASTP matrices.

DEFINITION 3.3. [illegible] A [illegible] γ [illegible] $A = LDU$ U^{-1} [illegible]

PROPOSITION 3.4. [illegible] A [illegible] $(A^T)^{-1}$ [illegible] A [illegible] γ [illegible]. The proof is completely analogous to that of Proposition 4.6 of [11]. The only difference is that, in the factorization $A = LDU$, in order to see that the upper triangular matrix $(U^T)^{-1}$ is ASTP, we have to use the same reasoning as in the proof of the converse of Proposition 3.2 to show the almost strict total positivity of LD. ∎

The following theorem characterizes ASTP matrices by means of their QR decompositions. This characterization is slightly different from those of nonsingular TP matrices and STP matrices given in Theorem 4.7 of [11], as we shall explain later.

THEOREM 3.5. [illegible] A [illegible] A [illegible] γ [illegible] Q_1, Q_2 [illegible] R_1, R_2 [illegible]

$$A = Q_1 R_1, \qquad A^T = Q_2 R_2. \tag{3.2}$$

The proof is analogous to that of Theorem 4.7 of [11], replacing TP by ASTP until the step when we use that the product of TP matrices $A^T A$ is also TP, because the product of ASTP matrices is [illegible] necessarily ASTP. So, the reasoning leading to the total positivity of R_1 (R_2) in the proof of Theorem 4.7 of [11] does not lead to the almost strict total positivity of them but only to their total positivity.

In fact, the following counterexample shows that in the above theorem we cannot replace the total positivity of R_1, R_2 by almost strict total positivity. The ASTP matrix

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

can be decomposed as $A = Q_1 R_1$, where

$$Q_1 = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad R_1 = \begin{pmatrix} \sqrt{2} & 1/\sqrt{2} & 1/\sqrt{2} \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 0 & 0 & 1 \end{pmatrix}.$$

The matrix R_1 is TP but not ASTP due to the minor $\det R_1[1,2|2,3] = 0$ in spite of the positivity of its diagonal elements. The essential uniqueness of the QR factorization implies that it is not possible to decompose $A = QR$ with Q orthogonal and R ASTP. Moreover, $A^T A$ illustrates that the property of being ASTP is not inherited under the product of matrices. In fact, $A^T A$ is not ASTP due to the minor $\det(A^T A)[1,2|2,3]$, which is zero and has positive diagonal elements.

Finally, let us recall that, in the particular case of A being STP, Theorem 4.7 of [11] shows that Q_1 and Q_2 are strict γ-matrices and R_1 and R_2 are Δ-STP matrices.

In summary, a matrix A is STP or nonsingular ASTP or nonsingular TP if and only if A and A^T can be decomposed as in (3.2) with Q_1, Q_2 orthogonal and R_1, R_2

nonsingular upper triangular, according to the following table.

A	Q_1, Q_2	R_1, R_2
STP	strict γ-matrices	Δ-STP
nonsingular ASTP	almost strict γ-matrices	TP
nonsingular TP	γ-matrices	TP

REFERENCES

[1] T. Ando, *Totally positive matrices*, Linear Algebra Appl., 90 (1987), pp. 165–219.

[2] C. de Boor, *Total positivity of the spline collocation matrix*, Indiana Univ. Math. J., 25 (1976), pp. 541–551.

[3] C. de Boor and A. Pinkus, *The approximation of a totally positive band matrix by a strictly positive one*, Linear Algebra Appl., 42 (1982), pp. 81–98.

[4] J. M. Carnicer and J. M. Peña, *Spaces with almost strictly totally positive bases*, Math. Nachr., 169 (1994), pp. 69–79.

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

[6] J. Demmel and P. Koev, *The accurate and efficient solution of a totally positive generalized Vandermonde linear system*, SIAM J. Matrix Anal. Appl., 27 (2005), pp. 142–152.

[7] D. Dimitrov and J. M. Peña, *Almost strict total positivity and a class of Hurwitz polynomials*, J. Approx. Theory, 132 (2005), pp. 212–223.

[8] J. Garloff, *Intervals of almost totally positive matrices*, Linear Algebra Appl., 363 (2003), pp. 103–108.

[9] M. Gasca, C. A. Micchelli, and J. M. Peña, *Almost strictly totally positive matrices*, Numer. Algorithms, 2 (1992), pp. 225–236.

[10] M. Gasca and J. M. Peña, *Total positivity and Neville elimination*, Linear Algebra Appl., 165 (1992), pp. 25–44.

[11] M. Gasca and J. M. Peña, *Total positivity, QR factorization, and Neville elimination*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1132–1140.

[12] M. Gasca and J. M. Peña, *A matricial description of Neville elimination with applications to total positivity*, Linear Algebra Appl., 202 (1994), pp. 33–54.

[13] M. Gasca and J. M. Peña, *On the characterization of almost strictly totally positive matrices*, Adv. Comput. Math., 3 (1995), pp. 239–250.

[14] M. Gasca and J. M. Peña, *On factorizations of totally positive matrices*, in Total Positivity and Its Applications, M. Gasca and C. A. Micchelli, eds., Kluwer Academic Press, Dordrecht, The Netherlands, 1996, pp. 109–130.

[15] G. M. L. Gladwell, *Inner total positivity*, Linear Algebra Appl., 393 (2004), pp. 179–195.

[16] S. Karlin, *Total Positivity, Vol.* I, Stanford University Press, Stanford, CA, 1968.

[17] J. H. B. Kemperman, *A Hurwitz matrix is totally positive*, SIAM J. Math. Anal., 13 (1982), pp. 331–341.

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

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 1, pp. 9–25

$\{k\}$-GROUP PERIODIC MATRICES*

JULIO BENÍTEZ† AND NÉSTOR THOME†

Abstract. In this paper we deal with two problems related to $\{k\}$-group periodic matrices (i.e., $A^{\#} = A^{k-1}$, where $A^{\#}$ is the group inverse of a matrix A). First, we give different characterizations of $\{k\}$-group periodic matrices. Later, we present characterizations of the $\{k\}$-group periodic matrices for linear combinations of projectors. This work extends some well-known results in the literature.

Key words. periodic matrices, involutive matrices, projectors, group inverse

AMS subject classifications. 15A09, 15A24

DOI. 10.1137/S0895479803437384

1. Background and notation. It is well known that projectors and their generalizations [10] have been widely used in different mathematical areas and applications. Moreover, applications to statistics [1], [2], [3], [13] reveal the importance of oblique projectors as well as their applications for developing the theory of perturbations of generalized inverses [7], [8], [16] and for studying iterative numerical methods [12]. In [14] and [15] the problem of characterizing $\{k\}$-potent matrices was studied from the viewpoint of sign-patterns.

The symbol $\mathbb{C}^{m\times n}$ is used to denote the set of $m \times n$ complex matrices. The transpose and the conjugate transpose of $A \in \mathbb{C}^{m\times n}$ are, respectively, denoted by A^{T} and A^*, and, if $m = n$, the spectrum of A (i.e., the set of all the eigenvalues of A) is denoted by $\sigma(A)$. We denote the direct sum of A and B by $A \oplus B$. Moreover, for a scalar $\alpha \in \mathbb{C}$ and a set S, we denote by αS the set of all elements αs, where $s \in S$.

For a given matrix $A \in \mathbb{C}^{n\times n}$, a matrix $X \in \mathbb{C}^{n\times n}$ satisfying $AXA = A$, $XAX = X$, and $AX = XA$ is a [illegible] of A. It is well known that the group inverse exists if and only if A and A^2 have the same rank, and that if it exists, then it is unique [4]. It is customary to denote the group inverse of A by $A^{\#}$. It is also well known [6] that the group inverse of a matrix $A \in \mathbb{C}^{n\times n}$ can be represented as a polynomial in A, whenever it exists. A natural question is, When can the matrix $A^{\#}$ be represented as a monomial in A? Fix $k \geq 2$. A matrix $A \in \mathbb{C}^{n\times n}$ satisfying $A^{\#} = A^{k-1}$ is called a $\{k\}$[illegible]. A characterization of this kind of matrix [5] is $A^{\#} = A^{k-1}$ if and only if $A^{k+1} = A$ for $k = 2, 3, \ldots$. We denote the set of all $\{k\}$-group periodic matrices by

$$\mathcal{G}^n(k) := \{A \in \mathbb{C}^{n\times n} : A^{k+1} = A\}, \quad k = 1, 2, 3, \ldots.$$

Let Ω_k denote the set of roots of unity of order k. We recall that if $\omega_k := \exp(2\pi i/k)$, then $\Omega_k = \{\omega_k^0, \omega_k^1, \ldots, \omega_k^{k-1}\}$.

In the study presented here, we consider only natural powers of matrices. The notation $r|k$ indicates that r divides to k, and the greatest common divisor and least common multiple of r and s are denoted by $\gcd(r, s)$ and $\mathrm{lcm}(r, s)$, respectively.

*Received by the editors November 4, 2003; accepted for publication (in revised form) by H. J. Werner August 22, 2005; published electronically March 17, 2006. This work was partially supported by the Generalitat Valenciana under Project Grupos03/062.

http://www.siam.org/journals/simax/28-1/43738.html

†Departamento de Matemática Aplicada, Universidad Politécnica de Valencia, Valencia 46022, Spain (jbenitez@mat.upv.es, njthome@mat.upv.es).

Throughout this paper it is assumed that c_1 and c_2 are nonzero elements of $\mathbb{C}$ and P_1, P_2 are nonzero different projectors of the same order over the field $\mathbb{C}$, that is, $P_1, P_2 \in \mathbb{C}^{n\times n} \setminus \{O\}$ and $P_1^2 = P_1 \neq P_2 = P_2^2$. Idempotent matrices $A \in \mathbb{C}^{n\times n}$ are also called ( ) , and if A is a Hermitian matrix (i.e., $A^* = A$), it is called an .

In this paper we first obtain some characterizations of $\{k\}$-group periodic matrices and later study the problem of finding linear combinations of projectors that are $\{k\}$-group periodic matrices; that is, we will describe the set

$$\mathcal{S}(P_1, P_2, k) := \{(c_1, c_2) \in \mathbb{C}^2 : (c_1P_1 + c_2P_2)^{k+1} = c_1P_1 + c_2P_2\},$$
$$k = 1, 2, 3, \ldots.$$

The obtained results here extend the results in [2] and [9]. The technique used for solving these particular cases ($k = 1$ in [2] and $k = 2$ in [9]) is tedious for $k \geq 3$, and we introduce a new technique for the general case solved here.

This paper is organized as follows. Section 2 gives some algebraic characterizations of the set $\mathcal{G}^n(k)$ and also some geometrical and topological aspects. Section 3 provides all the elements of the set $\mathcal{S}(P_1, P_2, k)$ for the case $P_1P_2 = P_2P_1$ by means of simultaneous diagonalization of projectors P_1 and P_2. Section 4 describes the set $\mathcal{S}(P_1, P_2, k)$ for the case $P_1P_2 \neq P_2P_1$, splitting this study into two cases: if k is not a multiple of 6 and if k is a multiple of 6.

Next, we quote some known definitions and results for further references.

A , $\mathcal{F} \subset \mathbb{C}^{n\times n}$ is an arbitrary (finite or infinite) set of matrices in which each pair in the set commutes under multiplication. A family of projectors $\{A_\alpha\}_{\alpha\in\mathcal{A}}$ is said to be if $A_\alpha A_\beta = O$ for all $\alpha, \beta \in \mathcal{A}$ and $\alpha \neq \beta$.

THEOREM 1.1 (see Thm. 1.3.19 of [11]). $\mathcal{F} \subset \mathbb{C}^{n\times n}$ $\mathcal{F}$ $S \in \mathbb{C}^{n\times n}$ $S^{-1}AS$ $A \in \mathcal{F}$

THEOREM 1.2 (see Thm. 6.31 of [11]). $A \in \mathbb{C}^{n\times n}$ $A = SDS^{-1}$ $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ $E \in \mathbb{C}^{n\times n}$ $\hat{\lambda}$ $A+E$ λ_i A $|\hat{\lambda} - \lambda_i| \leq k_\infty(S)\|E\|_\infty$ k_∞ $\|\cdot\|_\infty$

2. Characterizations of $\{k\}$-group periodic matrices. We start this section by giving a canonical form of the $\{k\}$-group periodic matrices.

THEOREM 2.1. $A \in \mathbb{C}^{n\times n}$

1. $A^{k+1} = A$ A $\{k\}$
2. $A_0, \ldots, A_{k-1}$ $A = \sum_{j=0}^{k-1} \omega_k^j A_j$
3. A $\sigma(A) \subseteq \{0\} \cup \Omega_k$

1. $A = \sum_{j=0}^{k-1} \omega_k^j A_j$ $A = \sum_{i=1}^{m} \lambda_i B_i = \sum_{j=1}^{m'} \mu_j C_j$ $\lambda_1, \ldots, \lambda_m$ $\mu_1, \ldots, \mu_{m'}$ $i \neq j$ $\lambda_i \neq \lambda_j$ $\mu_i \neq \mu_j$ $\{B_i\}_{i=1}^m$ $\{C_j\}_{j=1}^{m'}$ $m = m'$ $\sigma : \{1, \ldots, m\} \to \{1, \ldots, m'\}$ $\lambda_i = \mu_{\sigma(i)}$ $B_i = C_{\sigma(i)}$ $i = 1, 2, \ldots, m$
2. $j = 0, 1, \ldots, k-1$ ω_k^j A $A_j \neq O$

. We shall prove the implications $1 \Rightarrow 3 \Rightarrow 2 \Rightarrow 1$ and later the uniqueness. In fact, we have the following:

$1 \Rightarrow 3$. From $A^{k+1} = A$, the polynomial $q(t) = t^{k+1} - t$ is a multiple of the minimal polynomial $q_A(t)$ of A, and then every root of $q_A(t)$ has multiplicity 1. Hence A is diagonalizable by Corollary 3.3.10 in [11]. Moreover, it is clear that $\sigma(A) \subseteq \{0\} \cup \Omega_k$.

$3 \Rightarrow 2$. By the spectral theorem (see, for example, [4]), the decomposition in condition 2 can be easily obtained. In the case that $\omega_k^j \notin \sigma(A)$ for $j \in \{0, 1, \ldots, k-1\}$ we take $A_j = O$.

$2 \Rightarrow 1$. Since the family is commuting, $A^n = \sum_{j=0}^{k-1} \omega_k^{nj} A_j$.

In order to prove the uniqueness, we suppose

$$\sum_{i=1}^{m} \lambda_i B_i = \sum_{j=1}^{m'} \mu_j C_j \tag{2.1}$$

and without loss of generality $m \leq m'$ (if $m' < m$, it is sufficient to change m by m'). Choose $r \in \{1, \ldots, m\}$ and $s \in \{1, \ldots, m'\}$. Premultiplying (2.1) by B_r and postmultiplying (2.1) by C_s we get $\lambda_r B_r C_s = \mu_s B_r C_s$ and if $\lambda_r \neq \mu_s$, then $B_r C_s = O$. Premultiplying (2.1) again by B_1, we obtain

$$\lambda_1 B_1 = \sum_{j=1}^{m'} \mu_j B_1 C_j. \tag{2.2}$$

Since $\lambda_1 B_1 \neq O$ and $\mu_1, \ldots, \mu_{m'}$ are different, there exists only one $j \in \{1, \ldots, m'\}$ such that $\mu_j = \lambda_1$. Set $\sigma(1) := j$. From (2.2) we get $\lambda_1 B_1 = \mu_{\sigma(1)} B_1 C_{\sigma(1)}$ and thus

$$B_1 = B_1 C_{\sigma(1)}. \tag{2.3}$$

Postmultiplying (2.1) by $C_{\sigma(1)}$, we obtain that $B_1 C_{\sigma(1)} = C_{\sigma(1)}$, that (2.3) yields $B_1 = C_{\sigma(1)}$, and from (2.1) that the equality

$$\sum_{i=2}^{m} \lambda_i B_i = \sum_{j \neq \sigma(1)} \mu_j C_j$$

holds. Similarly, there exists $\sigma(2) \in \{1, \ldots, m'\} \setminus \{\sigma(1)\}$ such that $\lambda_2 = \mu_{\sigma(2)}$ and $B_2 = C_{\sigma(2)}$. Following in the same way, an injective function $\sigma : \{1, \ldots, m\} \to \{1, \ldots, m'\}$ satisfying $\lambda_i = \mu_{\sigma(i)}$ and $B_i = C_{\sigma(i)}$ for $i = 1, \ldots, m$ can be constructed. Let us prove that $m = m'$. Because of $m \leq m'$, if we suppose $m < m'$, the equality

$$\sum_{j \notin \{\sigma(1), \ldots, \sigma(m)\}} \mu_j C_j = O$$

follows from (2.1). By choosing $s \in \{1, \ldots, m'\} \setminus \{\sigma(1), \ldots, \sigma(m)\}$ and postmultiplying the last equality by C_s, we obtain $\mu_s C_s = O$, which is a contradiction. This concludes the proof. □

A useful fact is that a tripotent matrix (i.e., $A^3 = A$) can uniquely be decomposed as a difference of two disjoint projectors [13]. This property can be immediately derived as a simple consequence from Theorem 2.1.

Two important particular cases of Theorem 2.1 are considered in the following results.

COROLLARY 2.2. [illegible] $A \in \mathbb{C}^{n\times n}$ [illegible] $\{k\}$ [illegible]

1. [illegible] A [illegible] $A^3 = A$ [illegible] A [illegible] $A^2 = A$ [illegible] k [illegible]
2. [illegible] A [illegible] $A = O$ [illegible] k [illegible] 4 [illegible] $A^3 = -A$ [illegible] k [illegible] 4

[illegible]. Since normal matrices are diagonalizable, in both cases A is diagonalizable.

1. Since A is Hermitian, $\sigma(A) \subset \mathbb{R}$ and because A is a $\{k\}$-group periodic matrix, $\sigma(A) \subseteq \{0\} \cup \Omega_k$ by Theorem 2.1. Hence $\sigma(A) \subseteq (\{0\} \cup \Omega_k) \cap \mathbb{R}$. If k is even, then $(\{0\} \cup \Omega_k) \cap \mathbb{R} = \{0, 1, -1\} = \{0\} \cup \Omega_2$. Now, from Theorem 2.1, we obtain $A^3 = A$. The argument when k is odd is similar.

2. Since A is skew-Hermitian, $\sigma(A) \subset i\mathbb{R}$. As before, $\sigma(A) \subseteq (\{0\} \cup \Omega_k) \cap i\mathbb{R}$ by Theorem 2.1. If k is not a multiple of 4, then $(\{0\} \cup \Omega_k) \cap i\mathbb{R} = \{0\}$ and hence $A = O$. If $4|k$, then $(\{0\} \cup \Omega_k) \cap i\mathbb{R} = \{0, i, -i\}$ and hence $A^3 + A = O$. ∎

COROLLARY 2.3. [illegible] $A \in \mathbb{C}^{n\times n}$ [illegible] $A = \sum_{i=1}^{m} \lambda_i A_i$ [illegible] 2.1 [illegible] A_i [illegible]

[illegible]. If A is Hermitian, then $\sigma(A) \subset \mathbb{R}$. By the decomposition of A in Theorem 2.1, one has

$$A = A^* = \left(\sum_{i=1}^{m} \lambda_i A_i\right)^* = \sum_{i=1}^{m} \overline{\lambda_i} A_i^* = \sum_{i=1}^{m} \lambda_i A_i^*.$$

The uniqueness of the representation permits us to conclude that $A_i = A_i^*$ holds for $i = 1, \ldots, m$. If A is skew-Hermitian, then $\sigma(A) \subset i\mathbb{R}$. A similar argument as before is also valid for this case. ∎

It is known [10] that the class of projectors is the intersection of the class of tripotent ($A^3 = A$) and the class of quadripotent ($A^4 = A$) matrices. The next theorem extends this result and gives some geometrical and topological aspects of the set $\mathcal{G}^n(k)$ for an arbitrary $k \in \mathbb{N}$. We recall that $\mathcal{G}^n(k) := \{M \in \mathbb{C}^{n\times n} : M^{k+1} = M\}$.

THEOREM 2.4. [illegible] $r, s, k \in \mathbb{N}$ [illegible]

1. $r|k$ [illegible] $\mathcal{G}^n(r) \subseteq \mathcal{G}^n(k)$.
2. $\mathcal{G}^n(r) \cap \mathcal{G}^n(s) = \mathcal{G}^n(\gcd(r, s))$.
3. [illegible] $\omega \in \Omega_r$ [illegible] $\omega\mathcal{G}^n(r) = \mathcal{G}^n(r)$.
4. [illegible] $r|k$ [illegible] $\mathcal{G}^n(r)$ [illegible] $\mathcal{G}^n(k)$

[illegible]. 1. If $r|k$, then $\Omega_r \subseteq \Omega_k$. If $M \in \mathcal{G}^n(r)$, by Theorem 2.1, M is diagonalizable and $\sigma(M) \subseteq \{0\} \cup \Omega_r \subseteq \{0\} \cup \Omega_k$, and hence $M \in \mathcal{G}^n(k)$.

Conversely, since $\omega_r I \in \mathcal{G}^n(r) \subseteq \mathcal{G}^n(k)$, then $1 = \omega_r^k = \exp(2\pi ki/r)$ and from this last equality we get $k|r \in \mathbb{N}$.

2. From elementary algebra, it is known that $\Omega_r \cap \Omega_s = \Omega_{\gcd(r,s)}$. Now the assertion follows from Theorem 2.1.

3. It is obvious.

4. The first item implies that $\mathcal{G}^n(r) \subseteq \mathcal{G}^n(k)$. Let us prove that $\mathcal{G}^n(r)$ is a closed and open subset of $\mathcal{G}^n(k)$.

The function $f : \mathcal{G}^n(k) \to \mathbb{C}^{n\times n}$ given by $f(M) := M^{r+1} - M$ is continuous. Since $\mathcal{G}^n(r) = f^{-1}(\{O\})$, then $\mathcal{G}^n(r)$ is a closed subset of $\mathcal{G}^n(k)$.

Let

$$\epsilon := \min\{|z-w| : z \neq w, z, w \in \{0\} \cup \Omega_k\} > 0.$$

If $\mathcal{G}^n(r)$ were not an open subset of $\mathcal{G}^n(k)$, then there would exist $M \in \mathcal{G}^n(r)$ and $(M_i)_{i=1}^{\infty} \subset \mathcal{G}^n(k) \setminus \mathcal{G}^n(r)$ such that $\lim_{i\to\infty} M_i = M$. Since $M_i \notin \mathcal{G}^n(r)$ for $i \in \mathbb{N}$, there exists $\lambda_i \in \sigma(M_i) \setminus (\{0\} \cup \Omega_r)$ and now we can apply Theorem 1.2 because M is diagonalizable. Let S nonsingular and D diagonal be two matrices such that $M = SDS^{-1}$. If $E_i := M_i - M$, there exists an eigenvalue μ_i of M such that

$$|\lambda_i - \mu_i| \leq k_\infty(S)\|E_i\| = k_\infty(S)\|M_i - M\|. \tag{2.4}$$

Choose $i \in \mathbb{N}$ such that $\|M - M_i\| \leq \epsilon/k_\infty(S)$. From this inequality and (2.4) it follows that $|\lambda_i - \mu_i| < \epsilon$. Since $\lambda_i, \mu_i \in \{0\} \cup \Omega_k$ (because λ_i and μ_i are eigenvalues of two $\{k\}$-group periodic matrices), by definition of ϵ, we get $\lambda_i = \mu_i$. Since $M \in \mathcal{G}^n(r)$, then $\mu_i \in \{0\} \cup \Omega_r$. This contradicts $\lambda_i \notin \{0\} \cup \Omega_r$. □

Theorem 2.4 gives some information about $\mathcal{S}(P_1, P_2, k)$. Recall that

$$\mathcal{S}(P_1, P_2, k) := \{(c_1, c_2) \in \mathbb{C}^2 : (c_1P_1 + c_2P_2)^{k+1} = c_1P_1 + c_2P_2\}.$$

The first item of this theorem implies that each element of $\mathcal{S}(P_1, P_2, r)$ is also an element of $\mathcal{S}(P_1, P_2, k)$ for each divisor r of k. The second item describes the common elements of $\mathcal{S}(P_1, P_2, k)$ and $\mathcal{S}(P_1, P_2, s)$ by means of the greatest common divisor of r and s. The third assertion implies that the subset $\mathcal{S}(P_1, P_2, r)$ is invariant under rotations about the origin through an angle $2\pi/r$. The last item assures that $\mathcal{S}(P_1, P_2, r)$ is the union of connected components of $\mathcal{S}(P_1, P_2, k)$. In fact, if we define $f : \mathcal{S}(P_1, P_2, k) \to \mathcal{G}^n(k)$ given by $f(c_1, c_2) := c_1P_1 + c_2P_2$, then, since $\mathcal{G}^n(r)$ is a closed and open subset of $\mathcal{G}^n(k)$, the continuity of f implies that $\mathcal{S}(P_1, P_2, r) = f^{-1}(\mathcal{G}^n(r))$ is a closed and open subset of $\mathcal{S}(P_1, P_2, k)$; i.e., $\mathcal{S}(P_1, P_2, r)$ is the union of connected components of $\mathcal{S}(P_1, P_2, k)$.

We close this section with a characterization of $\{k\}$-group periodic matrices involving a rank factorization.

THEOREM 2.5. Let $A \in \mathbb{C}^{n\times n}$ be such that $A = FG$, A $F \in \mathbb{C}^{n\times r}$, $G \in \mathbb{C}^{r\times n}$, $r = \operatorname{rank}(A) = \operatorname{rank}(F) = \operatorname{rank}(G)$, A $\{k\}$, $(GF)^k = I$.

Proof. Assume that A is $\{k\}$-group periodic. Then $A^{k+1} = A$. Since $\operatorname{rank}(F) = \operatorname{rank}(G) = r$, there exist two matrices $F^{(1)} \in \mathbb{C}^{r\times n}$ and $G^{(1)} \in \mathbb{C}^{n\times r}$ such that $F^{(1)}F = I$ and $GG^{(1)} = I$ [4]. Premultiplying and postmultiplying the equality $A^{k+1} = A$ by $F^{(1)}$ and $G^{(1)}$, respectively, and using that $A = FG$, we get $(GF)^k = I$.

The other implication follows from $A^{k+1} = F(GF)^kG = FG = A$. □

Setting $k = 2$ in Theorem 2.5, we recover Theorem 2 in [4, p. 163] as a simple consequence.

3. Elements of $\mathcal{S}(P_1, P_2, k)$: Commutative case. This section describes, in an explicit manner, the elements (c_1, c_2) of $\mathcal{S}(P_1, P_2, k)$ whenever $P_1P_2 = P_2P_1$, with $P_1 \neq P_2$ and $c_1, c_2 \in \mathbb{C} \setminus \{0\}$.

THEOREM 3.1. Let P_1, P_2, $P_1P_2 = P_2P_1$, $P_1 \neq P_2$, $(c_1, c_2) \in \mathcal{S}(P_1, P_2, k)$,

1. $c_1 \in \Omega_k$, $c_1 + c_2 = 0$
   (a) k, P_1, P_2
   (b) k, $P_1P_2 = P_2$

2. $c_1 \in \Omega_k$, $c_1 + c_2 \in \Omega_k$
   (a) $c_2 \in \Omega_k$, P_1, P_2 [illegible]
   (b) $c_2 \notin \Omega_k$, $P_1P_2 = P_2$
3. $c_1 \in \Omega_k$, $c_2 \in \Omega_k$
   (a) $c_1 + c_2 \in \{0\} \cup \Omega_k$, P_1, P_2 [illegible]
   (b) $c_1 + c_2 \notin \{0\} \cup \Omega_k$, $P_1P_2 = O$
4. $c_2 \in \Omega_k$, $c_1 + c_2 = 0$
   (a) k [illegible] P_1, P_2 [illegible]
   (b) k [illegible] $P_1P_2 = P_1$
5. $c_2 \in \Omega_k$, $c_1 + c_2 \in \Omega_k$
   (a) $c_1 \in \Omega_k$, P_1, P_2 [illegible]
   (b) $c_1 \notin \Omega_k$, $P_1P_2 = P_1$

[illegible]. Since $P_1P_2 = P_2P_1$ and P_1, P_2 are diagonalizable matrices, by Theorem 1.1 there exist a nonsingular matrix S and two diagonal matrices $D_1 = \operatorname{diag}(\lambda_{11}, \ldots, \lambda_{1n})$ and $D_2 = \operatorname{diag}(\lambda_{21}, \ldots, \lambda_{2n})$ such that $P_i = SD_iS^{-1}$ for $i = 1, 2$. Since λ_{ij} are eigenvalues of projectors, $\lambda_{ij} \in \{0, 1\}$ for $i = 1, 2$ and $j = 1, \ldots, n$. Suppose that $(c_1P_1 + c_2P_2)^{k+1} = c_1P_1 + c_2P_2$. Since $c_1P_1 + c_2P_2 = S(c_1D_1 + c_2D_2)S^{-1}$ and $(c_1P_1 + c_2P_2)^{k+1} = S(c_1D_1 + c_2D_2)^{k+1}S^{-1}$, then $(c_1D_1 + c_2D_2)^{k+1} = c_1D_1 + c_2D_2$. Since $D_i = \operatorname{diag}(\lambda_{i1}, \ldots, \lambda_{in})$, then $(c_1\lambda_{1j}+c_2\lambda_{2j})^{k+1} = c_1\lambda_{1j}+c_2\lambda_{2j}$ for $j = 1, \ldots, n$. Hence

$$c_1\lambda_{1j} + c_2\lambda_{2j} \in \{0\} \cup \Omega_k \qquad \text{for } j = 1, \ldots, n. \tag{3.1}$$

If $\lambda_{1j} = \lambda_{2j}$ for $j \in \{1, \ldots, n\}$, then $D_1 = D_2$, and hence $P_1 = P_2$, which is a contradiction. Thus, there exists $r \in \{1, \ldots, n\}$ such that $\lambda_{1r} \neq \lambda_{2r}$. Since $\lambda_{1r}, \lambda_{2r} \in \{0, 1\}$, there are only two possibilities:

$$\lambda_{1r} = 1,\ \lambda_{2r} = 0 \qquad \text{or} \qquad \lambda_{1r} = 0,\ \lambda_{2r} = 1. \tag{3.2}$$

Using the first possibility and (3.1) we get $c_1 \in \{0\} \cup \Omega_k$, and so $c_1 \in \Omega_k$ because $c_1 \neq 0$. If $\lambda_{2j} = 0$ for all $j \in \{1, \ldots, n\}$, then $P_2 = SD_2S^{-1} = O$, which is again a contradiction, so there exists $s \in \{1, \ldots, n\}$ such that $\lambda_{2s} \neq 0$, and hence $\lambda_{2s} = 1$. Now, there are two possibilities for λ_{1s}. Using (3.1) for these possibilities, we obtain $\lambda_{1s} = 1$, which yields $c_1 + c_2 \in \{0\} \cup \Omega_k$ or $\lambda_{1s} = 0$, which implies $c_2 \in \{0\} \cup \Omega_k$, i.e., $c_2 \in \Omega_k$.

The first three cases of the theorem have been obtained. The second possibility in (3.2) yields the remaining cases of the theorem. Now define $Q := (c_1P_1 + c_2P_2)^{k+1} - (c_1P_1 + c_2P_2)$. Since $P_1P_2 = P_2P_1$,

$$\begin{aligned} Q &= \sum_{m=0}^{k+1} \binom{k+1}{m} (c_1P_1)^m (c_2P_2)^{k+1-m} - (c_1P_1 + c_2P_2) \\ &= c_1^{k+1}P_1 + \left(\sum_{m=1}^{k} \binom{k+1}{m} c_1^m c_2^{k+1-m} \right) P_1P_2 + c_2^{k+1}P_2 - (c_1P_1 + c_2P_2) \\ &= c_1(c_1^k - 1)P_1 + [(c_1 + c_2)^{k+1} - c_1^{k+1} - c_2^{k+1}]P_1P_2 + c_2(c_2^k - 1)P_2. \end{aligned} \tag{3.3}$$

The proof is now split into five cases:

1. Using $c_1 \in \Omega_k$ and $c_1 + c_2 = 0$ with (3.3) yields $Q = c_1(1-(-1)^k)(P_2 - P_1P_2)$. Now the assertion of the theorem for this case is simple to prove.

2. Since $c_1, c_1+c_2 \in \Omega_k$, by (3.3) we obtain $Q = c_2(1-c_2^k)(P_1P_2-P_2)$. Subcases 2(a) and 2(b) of the theorem are easily obtained.
3. From $c_1, c_2 \in \Omega_k$ and (3.3), it is easy to see that $Q = (c_1+c_2)((c_1+c_2)^k - 1)P_1P_2$. Now subcases 3(a) and 3(b) are evident.

Cases 4 and 5 are completely analogous to cases 1 and 2, and their proofs are omitted. The sufficiency is simple by considering the previous computations for the matrix Q. This completes the proof. □

1. If $c_1 + c_2 = 0$ and $c_1, c_2 \in \Omega_k$, then k must be even. In fact, $1 = c_1^k = (-c_2)^k = (-1)^k c_2^k = (-1)^k$.

2. If $c_1, c_2, c_1 + c_2 \in \Omega_k$, then k must be a multiple of 6 and $c_1 = \omega_6^2 c_2$ or $c_2 = \omega_6^2 c_1$. In fact, let us consider the triangle in the complex plane with vertices located at 0, c_1 and $c_1 + c_2$. All sides of this triangle have the same length. So the three angles are equal to $\pi/3$. Now it is easy to conclude the assertion.

As particular cases we obtain the conditions in the main results in [2] and [9].

4. Elements of $\mathcal{S}(P_1, P_2, k)$: Noncommutative case. As in the previous section, we will describe the set $\mathcal{S}(P_1, P_2, k)$ but now in the case that P_1 and P_2 do not commute.

The following results are the basis for our derivations.

LEMMA 4.1. $P_1P_2 \neq P_2P_1$ $(c_1, c_2) \in \mathcal{S}(P_1, P_2, k)$ $\alpha, \beta \in \{0\} \cup \Omega_k$ $c_1 + c_2 = \alpha + \beta$ $\alpha \neq \beta$

. By Theorem 2.1 there exists a nonsingular matrix S such that

$$c_1P_1 + c_2P_2 = SDS^{-1}, \qquad D = \begin{pmatrix} \lambda_1 I & O & \cdots & O \\ O & \lambda_2 I & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & \lambda_m I \end{pmatrix}, \tag{4.1}$$

where all $\lambda_j \in \{0\} \cup \Omega_k$ for $j = 1, \ldots, m$ and $\lambda_i \neq \lambda_j$ if $i \neq j$; that is, $c_1P_1 + c_2P_2$ has m distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_m$ of multiplicities $n_1, n_2, \ldots, n_m$. If we define $Q_1 := S^{-1}P_1S$ and $Q_2 := S^{-1}P_2S$, the following facts are easily obtained:

1. $c_1Q_1 + c_2Q_2 = D$.
2. $Q_i^2 = Q_i$.
3. $Q_1Q_2 - Q_2Q_1 \neq O$.
4. $Q_2D - DQ_2 \neq O$.

Fact 4 follows from facts 1 and 3 since $Q_2D - DQ_2 = Q_2(c_1Q_1 + c_2Q_2) - (c_1Q_1 + c_2Q_2)Q_2 = c_1Q_2Q_1 - c_1Q_1Q_2 = c_1(Q_2Q_1 - Q_1Q_2) \neq O$. From facts 1 and 2 we get the following equalities:

$$c_1Q_1 = D - c_2Q_2, \qquad c_1^2Q_1 = D^2 - c_2(DQ_2 + Q_2D) + c_2^2Q_2. \tag{4.2}$$

The matrices Q_1 and Q_2 are partitioned as

$$Q_1 = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mm} \end{pmatrix}, \qquad Q_2 = \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1m} \\ B_{21} & B_{22} & \cdots & B_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ B_{m1} & B_{m2} & \cdots & B_{mm} \end{pmatrix},$$

where each block in the position (i, j) has the same size as the block (i, j) in the partitioned matrix D in (4.1). Since $DQ_2 \neq Q_2D$, it is easy to check that there exist $i, j \in \{1, \ldots, m\}$ with $i \neq j$ such that $(\lambda_i - \lambda_j)B_{ji} \neq O$. Since $\lambda_i \neq \lambda_j$, then $B_{ji} \neq O$.

By considering the block (j,i) in (4.2) the following equalities are obtained:

$$c_1A_{ji} = -c_2B_{ji} \quad \text{and} \quad c_1^2A_{ji} = -c_2(\lambda_i+\lambda_j)B_{ji} + c_2^2B_{ji}.$$

From these last equalities we get

$$O = -c_1c_2B_{ji} + c_2(\lambda_i+\lambda_j)B_{ji} - c_2^2B_{ji} = c_2[(\lambda_i+\lambda_j) - (c_1+c_2)]B_{ji}. \tag{4.3}$$

Since $c_2 \neq 0$ and $B_{ji} \neq 0$, we get $\lambda_i + \lambda_j = c_1 + c_2$. The proof is completed. □

LEMMA 4.2. [illegible] $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \in \{0\} \cup \Omega_k$ [illegible] $\lambda_1 \neq \lambda_2$ $\lambda_3 \neq \lambda_4$ [illegible] $\lambda_1 + \lambda_2 = \lambda_3 + \lambda_4 \neq 0$ [illegible]

1. [illegible] $\lambda_j \neq 0$ [illegible] $j = 1,2,3,4$ [illegible] $\{\lambda_1, \lambda_2\} = \{\lambda_3, \lambda_4\}$
2. [illegible] $6 \nmid k$ [illegible] $\{\lambda_1, \lambda_2\} = \{\lambda_3, \lambda_4\}$
3. [illegible] $6 \mid k$ [illegible]
   (a) $\{\lambda_1, \lambda_2\} = \{\lambda_3, \lambda_4\}$
   (b) [illegible] $\omega \in \Omega_k$ [illegible] $\{\lambda_1, \lambda_2\} = \{0, \omega\}$ [illegible] $\{\lambda_3, \lambda_4\} = \{\omega_6^{-1}\omega, \omega_6\omega\}$
   (c) [illegible] $\omega \in \Omega_k$ [illegible] $\{\lambda_1, \lambda_2\} = \{\omega_6^{-1}\omega, \omega_6\omega\}$ [illegible] $\{\lambda_3, \lambda_4\} = \{0, \omega\}$

[illegible]. 1. Let $z := (\lambda_1+\lambda_2)/2 = (\lambda_3+\lambda_4)/2$ and $S_{\mathbb{C}} := \{u \in \mathbb{C} : |u| = 1\}$. For $u \in S_{\mathbb{C}}$, we define $r_u := \{\omega \in \mathbb{C} : \omega = z + xu, x \in \mathbb{R}\}$ and if $z + xu \in r_u \cap S_{\mathbb{C}}$, we get

$$1 = (z+xu)(\overline{z+xu}) = (z+xu)(\overline{z}+x\overline{u}) = |z|^2 + x(\overline{z}u + z\overline{u}) + x^2. \tag{4.4}$$

Since z is fixed, if $z + x_1u \in r_u \cap S_{\mathbb{C}}$ and $z + x_2u \in r_u \cap S_{\mathbb{C}}$, then $x_1x_2 = |z|^2 - 1$, because x_1 and x_2 satisfy (4.4).

Set $u := (\lambda_2 - \lambda_1)/|\lambda_2 - \lambda_1| \in S_{\mathbb{C}}$. Since

$$z + \frac{|\lambda_2-\lambda_1|}{2}u = \frac{\lambda_1+\lambda_2}{2} + \frac{\lambda_2-\lambda_1}{2} = \lambda_2 \in r_u \cap S_{\mathbb{C}}$$

and, analogously,

$$z - \frac{|\lambda_2-\lambda_1|}{2}u = \lambda_1 \in r_u \cap S_{\mathbb{C}},$$

we get that

$$|z|^2 - 1 = -\left(\frac{|\lambda_2-\lambda_1|}{2}\right)^2.$$

By using the same reasoning, but now with λ_3 and λ_4, we obtain that $|z|^2 - 1 = -(|\lambda_3 - \lambda_4|/2)^2$. Hence $|\lambda_1 - \lambda_2| = |\lambda_3 - \lambda_4| =: \rho$. Now it is easy to see that

$$\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\} \subset \left\{w \in \mathbb{C} : |w - z| = \frac{\rho}{2}\right\} =: C.$$

Hence $\lambda_j \in C \cap S_{\mathbb{C}}$ for $j = 1,2,3,4$. If the set $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ has at least three distinct elements, then C and $S_{\mathbb{C}}$ are two circumferences with at least three distinct common points. Hence $C = S_{\mathbb{C}}$ and so their centers coincide; i.e., $z = 0$, which is a contradiction. Thus, the set $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ has at most two distinct elements and this completes the proof of the first item.

2. If $0 \notin \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$, then the situation is the same as in the previous item. We suppose that $0 \in \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ and, without loss of generality, we take $0 = \lambda_1$ and hence $\lambda_2 = \lambda_3 + \lambda_4$. If $0 \in \{\lambda_3, \lambda_4\}$, it is easy to obtain the conclusion. If

$0 \notin \{\lambda_3, \lambda_4\}$, from $\lambda_2 = \lambda_3 + \lambda_4$, it is simple to get that $\{\lambda_3, \lambda_4\} = \{\omega_6^{-1}\lambda_2, \omega_6\lambda_2\}$. Since $\lambda_2, \lambda_3 \in \Omega_k$, we get $1 = (\lambda_3)^k = (\omega_6^{\pm 1}\lambda_2)^k = \exp(\pm ki\pi/3)$ and hence k is a multiple of 6, which contradicts the assumption of this item.

3. The proof of the above item works. □

LEMMA 4.3. $\alpha, \beta \in \mathbb{C}$ $P_1P_2 \neq P_2P_1$ $c_1P_1 + c_2P_2$ $X^2 - (\alpha+\beta)X + \alpha\beta I = O$ $c_1 + c_2 = \alpha + \beta$ $c_1c_2(P_1 - P_2)^2 = \alpha\beta I$

. If $c_1P_1 + c_2P_2$ satisfies the equation $X^2 - (\alpha+\beta)X + \alpha\beta I = O$, then

$$c_1^2P_1 + c_2^2P_2 + c_1c_2(P_1P_2 + P_2P_1) - (\alpha+\beta)(c_1P_1 + c_2P_2) + \alpha\beta I = O. \tag{4.5}$$

Premultiplying and postmultiplying (4.5) by P_1, we get, respectively,

$$(c_1^2 - (\alpha+\beta)c_1 + \alpha\beta)P_1 + (c_2^2 + c_1c_2 - (\alpha+\beta)c_2)P_1P_2 + c_1c_2P_1P_2P_1 = O,$$

$$(c_1^2 - (\alpha+\beta)c_1 + \alpha\beta)P_1 + (c_2^2 + c_1c_2 - (\alpha+\beta)c_2)P_2P_1 + c_1c_2P_1P_2P_1 = O.$$

Hence $c_2[c_1 + c_2 - (\alpha+\beta)](P_1P_2 - P_2P_1) = O$. Under the assumptions of the theorem, it is clear that this equation is equivalent to

$$c_1 + c_2 = \alpha + \beta. \tag{4.6}$$

Replacing (4.6) in (4.5) we get

$$\begin{aligned} O &= c_1[c_1 - (\alpha+\beta)]P_1 + c_2[c_2 - (\alpha+\beta)]P_2 + c_1c_2(P_1P_2 + P_2P_1) + \alpha\beta I \\ &= -c_1c_2[P_1 + P_2 - (P_1P_2 + P_2P_1)] + \alpha\beta I \\ &= -c_1c_2(P_1 - P_2)^2 + \alpha\beta I. \end{aligned}$$

Conversely, $(c_1P_1 + c_2P_2)^2 = (\alpha+\beta)(c_1P_1 + c_2P_2) - \alpha\beta I$ holds from the obvious equality $P_1P_2 + P_2P_1 = P_1 + P_2 - (P_1 - P_2)^2$. □

Now, we are ready to state the main result when k is not a multiple of 6.

THEOREM 4.4. k 6 $P_1P_2 \neq P_2P_1$ $c_1 + c_2 \neq 0$ $(c_1, c_2) \in \mathcal{S}(P_1, P_2, k)$ $\lambda_1, \lambda_2 \in \{0\} \cup \Omega_k$ $\alpha \in \mathbb{C}$ $\lambda_1 \neq \lambda_2$ $c_1 + c_2 = \lambda_1 + \lambda_2$ $\alpha c_1c_2 = \lambda_1\lambda_2$, Π $\Pi P_1 = P_1\Pi$ $\Pi P_2 = P_2\Pi$, $\Pi\Delta^2 = \alpha\Pi$ $\Delta := P_1 - P_2$

1. $\Delta^2 = \alpha I$
2. $c_1 \in \Omega_k$ $c_1 + c_2 \in \Omega_k$ $c_2 \notin \Omega_k$ $(I - \Pi)(\Delta^2 - \Delta) = O$
3. $c_1 \in \Omega_k$ $c_2 \in \Omega_k$ $c_1 + c_2 \notin \Omega_k$ $(I - \Pi)P_1P_2 = (I - \Pi)P_2P_1 = O$
4. $c_2 \in \Omega_k$ $c_1 + c_2 \in \Omega_k$ $c_1 \notin \Omega_k$ $(I - \Pi)(\Delta^2 + \Delta) = O$
5. $c_2 \in \Omega_k$ $(I - \Pi)P_1 = O$
6. $c_1 \in \Omega_k$ $(I - \Pi)P_2 = O$
7. $c_1 + c_2 \in \Omega_k$ $(I - \Pi)\Delta = O$

. Following the notation of Lemma 4.1, there exist $i, j \in \{1, \ldots, m\}$ such that $B_{ji} \neq O$ with $i \neq j$. In addition, $c_1 + c_2 = \lambda_i + \lambda_j$ holds.

By Lemma 4.2, we get $\lambda_i + \lambda_j \neq \lambda_s + \lambda_t$ for all $\{s, t\} \neq \{i, j\}$ and $s \neq t$. By (4.3), rearranging the eigenvalues of D and the blocks of Q_2 by some suitable permutation of rows and columns, we can suppose that

$$Q_2 = \begin{bmatrix} B_{11} & B_{12} & O & \cdots & O \\ B_{21} & B_{22} & O & \cdots & O \\ O & O & B_{33} & \cdots & O \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ O & O & O & \cdots & B_{mm} \end{bmatrix}.$$

Note that blocks $B_{33},\dots,B_{mm}$ may be absent. By this rearrangement, $c_1+c_2=\lambda_1+\lambda_2$ holds. From (4.2) we get $c_1D+c_2(DQ_2+Q_2D)=D^2+(c_1+c_2)c_2Q_2$ and its block (r,r) gives

$$(c_1-\lambda_r)\lambda_r I=c_2(c_1+c_2-2\lambda_r)B_{rr}. \tag{4.7}$$

Observe that $c_1+c_2-2\lambda_r\neq 0$. In fact, if $c_1+c_2=2\lambda_r$, then $\lambda_r\neq 0$ and $|\lambda_1+\lambda_2|=|c_1+c_2|=|2\lambda_r|=2$, so $\lambda_1\neq 0$ and $\lambda_2\neq 0$ and then $|\lambda_1|=|\lambda_2|=1$. This implies $\lambda_1=\lambda_2$, which is impossible. So from (4.7) we get $B_{rr}=\beta_r I$ for some $\beta_r\in\mathbb{C}$. From $c_1A_{rr}+c_2B_{rr}=\lambda_r I$, it follows that $A_{rr}=\alpha_r I$ for some $\alpha_r\in\mathbb{C}$. If we denote

$$M_0:=\begin{bmatrix}\lambda_1 I & O\\ O & \lambda_2 I\end{bmatrix},\quad M_1:=\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix},\quad M_2:=\begin{bmatrix}B_{11} & B_{12}\\ B_{21} & B_{22}\end{bmatrix}$$

and

$$J_0:=\lambda_3 I\oplus\cdots\oplus\lambda_m I,\qquad J_1:=\alpha_3 I\oplus\cdots\oplus\alpha_m I,\qquad J_2:=\beta_3 I\oplus\cdots\oplus\beta_m I,$$

then

$$D=M_0\oplus J_0,\qquad Q_1=M_1\oplus J_1,\qquad Q_2=M_2\oplus J_2. \tag{4.8}$$

Again we recall that the blocks J_0, J_1, and J_2 may be absent. If not, $J_1J_2=J_2J_1$. Define $B:=I\oplus O$ partitioned as in D, Q_1, and Q_2 in (4.8). Note that B is a nonzero projector, $BQ_1=Q_1B$, and $BQ_2=Q_2B$. It is clear that if $\Pi:=SBS^{-1}$, then Π is a projector and $\Pi P_i=P_i\Pi$ for $i=1,2$. Note that M_1,M_2,J_1, and J_2 are also projectors since Q_1 and Q_2 are projectors.

Focusing on blocks M_i $(i=0,1,2)$, the equality $c_1M_1+c_2M_2=M_0$ holds. Since $M_0^2-(\lambda_1+\lambda_2)M_0+\lambda_1\lambda_2 I=O$, in order to apply Lemma 4.3, we have to prove that M_1 and M_2 do not commute. In fact, this is clear from (4.8) since $Q_1Q_2\neq Q_2Q_1$ and $J_1J_2=J_2J_1$. By Lemma 4.3, it follows that $(M_1-M_2)^2=\alpha I$, where $\alpha:=(\lambda_1\lambda_2)/(c_1c_2)$. So, $B(Q_1-Q_2)^2=\alpha B$, which implies $\Pi(P_1-P_2)^2=\alpha\Pi$.

Now, we focus on commuting blocks J_i $(i=0,1,2)$. Since $c_1J_1+c_2J_2$ is a $\{k\}$-group periodic matrix and J_1, J_2 are projectors, Theorem 3.1 gives the following cases:

0. $c_1\in\Omega_k$, $c_2\in\Omega_k$, and $c_1+c_2\in\Omega_k$. This implies that k is a multiple of 6, which is contrary to the assumption.
1. If blocks J_0,J_1 and J_2 are absent, then $\Delta^2=\alpha I$.
2. $c_1\in\Omega_k$, $c_1+c_2\in\Omega_k$, $c_2\notin\Omega_k$, and $J_1J_2=J_2$. It is easy to prove that J_1-J_2 is a projector, so $(I-B)(Q_1-Q_2)^2=(I-B)(Q_1-Q_2)$, which implies $(I-\Pi)(\Delta^2-\Delta)=O$.
3. $c_1\in\Omega_k$, $c_2\in\Omega_k$, $c_1+c_2\notin\Omega_k$, and $J_1J_2=O$. Because of $J_1J_2=J_2J_1=O$, we can obtain that $(I-B)Q_1Q_2=(I-B)Q_2Q_1=O$, which implies $(I-\Pi)P_1P_2=(I-\Pi)P_2P_1=O$.
4. $c_2\in\Omega_k$, $c_1+c_2\in\Omega_k$, $c_1\notin\Omega_k$, and $J_1J_2=J_1$. It is similar to case 2.

Note that we cannot apply Theorem 3.1 if $J_1=O$, $J_2=O$, or $J_1=J_2$. If $J_1=J_2=O$, then we obtain case 1. So, we must consider the following cases:

5. $J_1=O$ and $J_2\neq O$. Since $(c_1J_1+c_2J_2)^{k+1}=c_1J_1+c_2J_2$, we get $(c_2J_2)^{k+1}=c_2J_2$; hence $c_2\in\Omega_k$. Moreover, since $J_1=O$, we get $(I-\Pi)P_1=O$.
6. $J_2=O$ and $J_1\neq O$. This is analogous to the previous case.
7. $J_1=J_2\neq O$. We obtain $(c_1+c_2)^{k+1}=c_1+c_2$, and by assumption we get $c_1+c_2\in\Omega_k$. Since $J_1-J_2=O$, we get $(I-\Pi)\Delta=O$.

So, the necessity has been proved.

Let us prove the sufficiency for case 1: By Lemma 4.3 the matrix $Q := c_1P_1 + c_2P_2$ satisfies $Q^2 = (\lambda_1 + \lambda_2)Q - \lambda_1\lambda_2 I$. By recurrence, there exists $(a_m, b_m)_{m=0}^{\infty}$ such that $Q^m = a_m Q + b_m I$ for $m \in \{0\} \cup \mathbb{N}$. It is easy to prove

$$\begin{bmatrix} a_{m+1} \\ b_{m+1} \end{bmatrix} = \begin{bmatrix} \lambda_1 + \lambda_2 & 1 \\ -\lambda_1\lambda_2 & 0 \end{bmatrix} \begin{bmatrix} a_m \\ b_m \end{bmatrix}; \qquad \text{i.e., } v_{m+1} = Av_m.$$

Since A is diagonalizable and $\sigma(A) = \{\lambda_1, \lambda_2\} \subseteq \{0\} \cup \Omega_k$, by Theorem 2.1, we get $A^{k+1} = A$; hence $v_{k+1} = A^{k+1}v_0 = Av_0 = A[0\ 1]^{\mathrm{T}} = [1\ 0]^{\mathrm{T}}$. So $Q^{k+1} = a_{k+1}Q + b_{k+1}I = Q$. Thus the sufficiency for case 1 is proved.

For cases 2, 3, and 4 it is useful to observe that

$$\begin{aligned}(c_1P_1 + c_2P_2)^2 &= (c_1\Pi P_1 + c_2\Pi P_2 + c_1(I-\Pi)P_1 + c_2(I-\Pi)P_2)^2 \\ &= (c_1\Pi P_1 + c_2\Pi P_2)^2 + (c_1(I-\Pi)P_1 + c_2(I-\Pi)P_2)^2\end{aligned}$$

since $\Pi P_i(I-\Pi)P_j = (I-\Pi)P_i\Pi P_j = O$ for $i, j \in \{1, 2\}$. So, by recurrence

$$(c_1P_1 + c_2P_2)^{k+1} = [c_1\Pi P_1 + c_2\Pi P_2]^{k+1} + [c_1(I-\Pi)P_1 + c_2(I-\Pi)P_2]^{k+1}.$$

Hence, in order to prove that $(c_1, c_2) \in \mathcal{S}(P_1, P_2, k)$, it is enough to see that

$$(c_1, c_2) \in \mathcal{S}(\Pi P_1, \Pi P_2, k) \cap \mathcal{S}((I-\Pi)P_1, (I-\Pi)P_2, k).$$

2. Let $T_1 := \Pi P_1$ and $T_2 := \Pi P_2$. It is clear that T_1 and T_2 are projectors. Now we define $Q := c_1T_1 + c_2T_2$. It is easy to verify that under the assumptions of this case, $\Pi Q = Q$ and $Q^2 = (\lambda_1 + \lambda_2)Q - (\lambda_1\lambda_2)\Pi$. In the same manner as in case 1 we obtain $Q^{k+1} = Q$ and hence $(c_1, c_2) \in \mathcal{S}(T_1, T_2, k)$.
   Let $R_1 := (I-\Pi)P_1$ and $R_2 := (I-\Pi)P_2$. It is clear that R_1 and R_2 are projectors. Since $(I-\Pi)\Delta^2 = (I-\Pi)\Delta$ we get $2R_2 = R_1R_2 + R_2R_1$. Premultiplying first by R_1 we obtain $R_1R_2 = R_1R_2R_1$ and postmultiplying later by R_1 we obtain $R_2R_1 = R_1R_2R_1$. Hence R_1 and R_2 commute and now it is easy to deduce that $R_1R_2 = R_2$. By the sufficiency of Theorem 3.1 we get that $(c_1, c_2) \in \mathcal{S}(R_1, R_2, k)$.
3. In a similar manner as in the previous case we obtain $(c_1, c_2) \in \mathcal{S}(\Pi P_1, \Pi P_2, k)$. Let $R_1 := (I-\Pi)P_1$ and $R_2 := (I-\Pi)P_2$. It is clear that R_1 and R_2 are projectors satisfying $R_2R_1 = R_1R_2 = O$ and that sufficiency of Theorem 3.1 yields $(c_1, c_2) \in \mathcal{S}(R_1, R_2, k)$.
4. It is similar to case 2.

For cases 5, 6, and 7 we proceed as in case 2 and obtain $(c_1, c_2) \in \mathcal{S}(\Pi P_1, \Pi P_2, k)$. In these cases it is easy to check that $(c_1, c_2) \in \mathcal{S}((I-\Pi)P_1, (I-\Pi)P_2, k)$. □

Note: In cases 2, 4, and 7 it can be proved that $\alpha = 0$. In fact, $c_1 + c_2 = \lambda_1 + \lambda_2$ and $\lambda_1, \lambda_2, c_1 + c_2 \in \{0\} \cup \Omega_k$ implies $\lambda_1 = 0$ or $\lambda_2 = 0$. In the third case, α must be nonzero because if $\alpha = 0$, then $\lambda_1 = 0$ or $\lambda_2 = 0$, so $c_1 + c_2 = \lambda_1 + \lambda_2 \in \Omega_k$, which is a contradiction.

Systematic procedures can be designed in order to decide when each case of Theorem 4.4 may occur and furthermore when the involved numbers c_1 and c_2 may be found. We will present such a procedure to decide when case 3 of Theorem 4.4 may occur for given matrices P_1 and P_2. In addition, this algorithm calculates the projector Π and $\alpha \in \mathbb{C}$ appearing in Theorem 4.4. Similar algorithms for remaining cases may be designed.

ALGORITHM (value of α and matrix Π for case 3).

Step 1. Compute the eigenvalues $\alpha_1, \ldots, \alpha_m$ of Δ^2 and set $i := 1$.

Step 2. Set $\alpha := \alpha_i$.

Step 3. Construct a basis for null space of $\Delta^2 - \alpha I$ and denote by B_1 the matrix whose columns are the vectors of this basis.

Step 4. Construct a basis for the intersection of the null spaces of P_1P_2 and P_2P_1, respectively, and denote by B_2 the matrix whose columns are the vectors of this basis.

Step 5. Let $B := [B_1|B_2]$. If B is singular, let $i := i + 1$ and go to Step 2. If B is nonsingular, then go to Step 6.

Step 6. Let $\Pi := [B_1|B_2](I \oplus O)[B_1|B_2]^{-1}$. (It is obvious that Π is a projector.)

Step 7. If $\Pi P_1 = P_1\Pi$ and $\Pi P_2 = P_2\Pi$, then case 3 is satisfied and the projector Π and $\alpha \in \mathbb{C}$ have been found (it can be proved that $\Pi\Delta^2 = \alpha\Pi$ and $(I-\Pi)P_1P_2 = (I-\Pi)P_2P_1 = O$), and the algorithm is finished.
Otherwise,

7.1. If $i = m$, the algorithm is finished and we decide that case 3 of Theorem 4.4 is not satisfied.

7.2. If $i \neq m$ let $i := i + 1$ and go to Step 2.

Note that we have characterized all the possible structures of combinations of two idempotent matrices that are $\{k\}$-periodic. Moreover, using Theorem 4.4 and the algorithms we can compute the values c_1 and c_2 such that $c_1P_1+c_2P_2$ is a $\{k\}$-periodic matrix.

The main result when k is a multiple of 6 is given in the following theorem.

THEOREM 4.5. [illegible] k [illegible] 6 $P_1P_2 \neq P_2P_1$ $c_1 + c_2 \neq 0$ [illegible] $\Delta := P_1 - P_2$ [illegible] $(c_1, c_2) \in \mathcal{S}(P_1, P_2, k)$ [illegible]

1. [illegible] 4.4 [illegible]
2. [illegible] $\lambda_1, \lambda_2 \in \Omega_k$ [illegible] $\lambda_1 \neq \lambda_2$ [illegible] $\omega \in \Omega_k$ [illegible] $c_1 + c_2 = \lambda_1 + \lambda_2$ $\{c_1, c_2\} = \{\omega\omega_6, \omega\omega_6^{-1}\}$ [illegible] Π [illegible] P_1 [illegible] P_2 [illegible] $\Pi\Delta^2 = \alpha\Pi$ [illegible] $\alpha := \lambda_1\lambda_2/\omega^2$ [illegible] $(I-\Pi)P_1P_2 = (I-\Pi)P_2P_1$
3. [illegible] Π [illegible] P_1 [illegible] P_2 [illegible] $\omega \in \Omega_k$ [illegible] $\alpha \in \mathbb{C}$ [illegible] $c_1 + c_2 = \omega$ $\alpha c_1c_2 = \omega^2$ $\Pi\Delta^2 = O$ [illegible] $(I-\Pi)(\Delta^2 - \alpha I) = O$
4. [illegible] $\{\Pi_1, \Pi_2, \Pi_3\}$ [illegible] P_1 [illegible] P_2 [illegible] $I = \sum_{i=1}^{3} \Pi_i$ [illegible] $\Pi_1\Delta^2 = O$ $\Pi_2(\Delta^2 - \alpha I) = O$ [illegible] $\omega \in \Omega_k$ $\alpha \in \mathbb{C}$ [illegible] $c_1 + c_2 = \omega$ [illegible] $\alpha c_1c_2 = \omega^2$ [illegible]
   (a) $c_2 \in \Omega_k$ [illegible] $\Pi_3P_1 = O$
   (b) $c_1 \in \Omega_k$ [illegible] $\Pi_3P_2 = O$
   (c) $\Pi_3\Delta = O$
   (d) $\{c_1, c_2\} = \{\omega\omega_6, \omega\omega_6^{-1}\}$ $\alpha = 1$ [illegible] $\Pi_3P_1P_2 = \Pi_3P_2P_1$
   (e) $c_1 \in \Omega_k$ $c_2 \notin \Omega_k$ [illegible] $\Pi_3(\Delta^2 - \Delta) = O$
   (f) $c_2 \in \Omega_k$ $c_1 \notin \Omega_k$ [illegible] $\Pi_3(\Delta^2 + \Delta) = O$

[illegible]. As in the proof of Theorem 4.4, following the notation of Lemma 4.1, there exist $i, j \in \{1, \ldots, m\}$ such that $B_{ji} \neq O$ with $i \neq j$, and also $\lambda_i + \lambda_j = c_1 + c_2 \neq 0$ holds. We shall consider two possibilities:

$$\text{(4.9)} \qquad \text{(a) } \lambda_i + \lambda_j \notin \Omega_k \qquad \text{or} \qquad \text{(b) } \lambda_i + \lambda_j \in \Omega_k.$$

In case (a) of (4.9) it is evident that $\lambda_i \neq 0$ and $\lambda_j \neq 0$. If there exist $\lambda_r, \lambda_s \in \{0\} \cup \Omega_k$

such that $\lambda_r + \lambda_s = \lambda_i + \lambda_j$, it is also evident that $\lambda_r \neq 0$ and $\lambda_s \neq 0$. From Lemma 4.2 we get $\lambda_i = \lambda_r$ or $\lambda_i = \lambda_s$. In addition to the studied cases, another case appears:

2. $c_1 \in \Omega_k$, $c_2 \in \Omega_k$, and $c_1 + c_2 \in \Omega_k$. Let $\omega := c_1 + c_2$. In order to deduce $\{c_1, c_2\} = \{\omega\omega_6, \omega\omega_6^{-1}\}$ recall that $\omega_6 = \exp(i\pi/3)$. The same construction for J_1 and J_2 as in the previous theorem is valid for the present case. As before, $J_1J_2 = J_2J_1$ and hence $(I - \Pi)P_1P_2 = (I - \Pi)P_2P_1$. Moreover, $c_1c_2 = \omega^2$ and $\alpha = (\lambda_1\lambda_2)/(c_1c_2)$.

Now we consider case (b) of (4.9). Lemma 4.2 allows us to suppose that $\lambda_1 = 0$, $\lambda_2 = \omega$, $\lambda_3 = \omega_6^{-1}\omega$, and $\lambda_4 = \omega_6\omega$ for some $\omega \in \Omega_k$. In fact, this is possible by rearranging the eigenvalues of D and the blocks of Q_2, that is,

$$Q_2 = \left[\begin{array}{cc|cc|ccc} B_{11} & B_{12} & O & O & O & \cdots & O \\ B_{21} & B_{22} & O & O & O & \cdots & O \\ \hline O & O & B_{33} & B_{34} & O & \cdots & O \\ O & O & B_{43} & B_{44} & O & \cdots & O \\ \hline O & O & O & O & B_{55} & \cdots & O \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ O & O & O & O & O & \cdots & B_{mm} \end{array}\right] =: \left[\begin{array}{ccc} M_2 & O & O \\ O & N_2 & O \\ O & O & J_2 \end{array}\right].$$

The matrices Q_1 and D will be partitioned in the same way. Observe that some block may be absent. The equalities $B_{ii} = \beta_i I$ and $A_{ii} = \alpha_i I$ for some $\alpha_i, \beta_i \in \mathbb{C}$ can be obtained as before. Then $J_1J_2 = J_2J_1$ and this rearranging yields

$$M_0 = \left[\begin{array}{cc} O & O \\ O & \omega I \end{array}\right] \qquad \text{and} \qquad N_0 = \left[\begin{array}{cc} \omega_6^{-1}\omega I & O \\ O & \omega_6\omega I \end{array}\right].$$

If some of blocks M and N are absent, it reduces to some previously studied case.

Since $Q_1Q_2 \neq Q_2Q_1$, $M_1M_2 \neq M_2M_1$ or $N_1N_2 \neq N_2N_1$. If $M_1M_2 \neq M_2M_1$ and $N_1N_2 = N_2N_1$, then $N_1 \oplus J_1$ commutes with $N_2 \oplus J_2$ and now instead of $N_i \oplus J_i$ we shall write J_i and it becomes as in some previous case. The situation $M_1M_2 = M_2M_1$, $N_1N_2 \neq N_2N_1$ can be treated similarly. Now we focus on the situation $M_1M_2 \neq M_2M_1$ and $N_1N_2 \neq N_2N_1$. Applying Lemma 4.3 twice we get $(M_1 - M_2)^2 = O$ and $(N_1 - N_2)^2 = \alpha I$, where $\alpha := \omega^2/(c_1c_2)$ and $c_1 + c_2 = \omega$.

Now we split the proof into two cases:

3. If blocks J are absent, it is easy to see that $\Delta^2 = S(O \oplus \alpha I)S^{-1}$ and then there exists a projector Π such that $\Pi\Delta^2 = O$ and $(I - \Pi)\Delta^2 = \alpha(I - \Pi)$.
4. If blocks J are not absent, we denote $B_1 := I \oplus O \oplus O$, $B_2 := O \oplus I \oplus O$, and $B_3 := O \oplus O \oplus I$ with the same partition as in $Q_2 = M_2 \oplus N_2 \oplus J_2$. Now we define $\Pi_i := SB_iS^{-1}$ for $i = 1, 2, 3$. Thus we get $\Pi_1\Delta^2 = O$ and $\Pi_2(\Delta^2 - \alpha I) = O$. In this case $c_1J_1 + c_2J_2$ is a $\{k\}$-group periodic matrix. The following cases may occur:
   (a) $J_1 = O$, $J_2 \neq O$. This case yields $c_2 \in \Omega_k$ and $\Pi_3P_1 = O$.
   (b) $J_2 = O$, $J_1 \neq O$. The proof is similar to the previous one.
   (c) $J_1 = J_2 \neq O$. Since $J_1 - J_2 = O$, $\Pi_3\Delta = O$.

   We apply Theorem 3.1 for the remaining cases since $J_1 \neq J_2$.
   (d) $c_1 \in \Omega_k$, $c_2 \in \Omega_k$, and $c_1 + c_2 \in \Omega_k$. Since $c_1 + c_2 = \omega$, $\{c_1, c_2\} = \{\omega\omega_6, \omega\omega_6^{-1}\}$.
   (e) $c_1 \in \Omega_k$, $c_1 + c_2 \in \Omega_k$, and $c_2 \notin \Omega_k$. Theorem 3.1 yields $J_1J_2 = J_2J_1 = J_2$. It is easy to see that $(J_1 - J_2)^2 = J_1 - J_2$ and thus $\Pi_3(\Delta^2 - \Delta) = O$.
   (f) $c_2 \in \Omega_k$, $c_1 + c_2 \in \Omega_k$, and $c_1 \notin \Omega_k$. The proof is similar to the previous one.

Let us proof the sufficiency. By a similar argument used in the sufficiency of Theorem 4.4 we obtain the following useful fact: "If $\{\Pi_i\}_{i=1}^m$ is a commuting and disjoint family of projectors which commutes with P_1 and P_2 such that $I = \sum_{i=1}^m \Pi_i$, then $\cap_{i=1}^m \mathcal{S}(\Pi_i P_1, \Pi_i P_2, k) \subseteq \mathcal{S}(P_1, P_2, k)$."

2. Since $\Pi\Delta^2 = \alpha\Pi$ with $\alpha = (\lambda_1\lambda_2)/(c_1c_2)$, $(c_1, c_2) \in \mathcal{S}(\Pi P_1, \Pi P_2, k)$. This fact can be shown in a way similar to that of the proof of the sufficiency of case 2 (first part) in Theorem 4.4. As $(I-\Pi)P_1P_2 = (I-\Pi)P_2P_1$, $(I-\Pi)P_1$ and $(I-\Pi)P_2$ are two commuting projectors and since $c_1 \in \Omega_k$, $c_2 \in \Omega_k$, and $c_1 + c_2 \in \Omega_k$, by the sufficiency of Theorem 3.1 we get $(c_1, c_2) \in \mathcal{S}((I-\Pi)P_1, (I-\Pi)P_2, k)$.

The remaining cases are similar, and the proof is completed. □

. In case 2 of the above theorem, if $\alpha = 0$, then $c_1 + c_2 \in \Omega_k$. In fact, if $\alpha = 0$, then $\lambda_1\lambda_2 = 0$, which implies $c_1 + c_2 = \lambda_1 + \lambda_2 \in \Omega_k$.

Algorithms similar to the one depicted after Theorem 4.4 may be given for finding the projectors appearing in Theorem 4.5.

The necessary and sufficient conditions obtained in the previous theorem may be difficult to check. Here we present necessary (but sometimes not sufficient) conditions that are easier to check.

Corollary 4.6. 4.4

1. $\Delta^2 = \alpha I$
2. $\Delta^3 = \Delta^2$
3. $\alpha = 1$ $\Delta^3 = \Delta$ $\alpha \neq 1$ $\Delta^5 - \Delta^3 = \alpha(\Delta^3 - \Delta)$
4. $\Delta^3 = -\Delta^2$
5. $\alpha = 0$ $\Delta^3 + \Delta^2 = O$ $\alpha = 1$ $\Delta^3 = \Delta$ $\alpha \notin \{0, 1\}$ $\Delta^4 + \Delta^3 = \alpha(\Delta^2 + \Delta)$
6. $\alpha = 0$ $\Delta^3 = \Delta^2$ $\alpha = 1$ $\Delta^3 = \Delta$ $\alpha \notin \{0, 1\}$ $\Delta^4 - \Delta^3 = \alpha(\Delta^2 - \Delta)$
7. $\alpha = 0$ $\Delta^2 = O$ $\alpha \neq 0$ $\Delta^3 = \alpha\Delta$

. This corollary follows easily from the next claim: $C_1, \ldots, C_m$ $I = \sum_{i=1}^m C_i$ $p_1, \ldots, p_m$ $C_i p_i(\Delta) = O$ $i = 1, \ldots, m$ $p(\Delta) = O$ $p := \operatorname{lcm}(p_1, \ldots, p_m)$ " In fact, since $p_i | p$, there exist polynomials $q_1, \ldots, q_m$ such that $p = p_i q_i$ and thus

$$p(\Delta) = \sum_{i=1}^m C_i p(\Delta) = \sum_{i=1}^m C_i p_i(\Delta) q_i(\Delta) = O.$$

Now it is enough to apply this claim with $C_1 := \Pi$ and $C_2 := I - \Pi$. □

The next corollary gives some necessary conditions for cases 3, 5, and 6.

Corollary 4.7. 4.4

3. $(1-\alpha)P_1P_2 = (P_1P_2)^2$ $(1-\alpha)P_2P_1 = (P_2P_1)^2$
5. $(1-\alpha)P_1 = P_1P_2P_1$
6. $(1-\alpha)P_2 = P_2P_1P_2$

. We will use the following simple fact: $A, B \in \mathbb{C}^{n\times n}$ $f_1, \ldots, f_m$ $\{f_i(A, B)\}_{i=1}^m$ $\{\Pi_i\}_{i=1}^m$ $\Pi_i f_i(A, B) = O$ $\sum_{i=1}^m \Pi_i = I$ $f_1(A, B) \cdots f_n(A, B) = O$ " Setting $f_1(P_1, P_2) := (P_1 - P_2)^2 - \alpha I$ and $f_2(P_1, P_2) := P_1P_2$, we get the first equality of case 3. The remaining cases may be analogously obtained. □

COROLLARY 4.8. [illegible] 4.5

3. $\Delta^4 = \alpha\Delta^2$

4. (a) [illegible] $\alpha = 1$ [illegible] $\Delta^4 = \Delta^2$ [illegible] $\alpha \neq 1$ [illegible] $\Delta^5 + \Delta^4 = \alpha(\Delta^3 + \Delta^2)$
   (b) [illegible] $\alpha = 1$ [illegible] $\Delta^4 = \Delta^2$ [illegible] $\alpha \neq 1$ [illegible] $\Delta^5 - \Delta^4 = \alpha(\Delta^3 - \Delta^2)$
   (c) $\Delta^4 = \alpha\Delta^2$
   (e) [illegible] $\alpha = 1$ [illegible] $\Delta^4 = \Delta^2$ [illegible] $\alpha \neq 1$ [illegible] $\Delta^5 - \Delta^4 = \alpha(\Delta^3 - \Delta^2)$
   (f) [illegible] $\alpha = 1$ [illegible] $\Delta^4 = \Delta^2$ [illegible] $\alpha \neq 1$ [illegible] $\Delta^5 + \Delta^4 = \alpha(\Delta^3 + \Delta^2)$

[illegible]. It is enough to apply the claim of the proof of Corollary 4.7 by using $C_1 := \Pi$ and $C_2 := I - \Pi$ for case 3 and $C_i := \Pi_i$ for $i = 1, 2, 3$ for case 4. □

COROLLARY 4.9. [illegible] 4.5

2. $(1-\alpha)(P_1P_2 - P_2P_1) = (P_1P_2)^2 - (P_2P_1)^2$

4. (a) $P_1(I - P_2P_1)^2 = \alpha P_1(I - P_2P_1)$
   (b) $P_2(I - P_1P_2)^2 = \alpha P_2(I - P_1P_2)$
   (d) $(1-\alpha)(P_1P_2 - P_2P_1) + \alpha[(P_1P_2)^2 - (P_2P_1)^2] = (P_1P_2)^3 - (P_2P_1)^3$

[illegible]. Using the fact mentioned in Corollary 4.7, these conditions are easy to obtain. □

COROLLARY 4.10. [illegible] $P_1P_2 - P_2P_1$ [illegible] 1 [illegible] 4.4 [illegible] 3 [illegible] 4.5 [illegible].

[illegible]. Following the notation of the proof of Theorems 4.4 and 4.5 and using that $P_1P_2 - P_2P_1$ is nonsingular, we obtain that $Q_1Q_2 - Q_2Q_1$ is also nonsingular. Since $Q_1Q_2 - Q_2Q_1 = (M_1M_2 - M_2M_1) \oplus (N_1N_2 - N_2N_1) \oplus (J_1J_2 - J_2J_1) = (M_1M_2 - M_2M_1) \oplus (N_1N_2 - N_2N_1) \oplus O$, blocks J_i must be absent and hence the conclusion follows. □

THEOREM 4.11. [illegible] $P_1P_2 \neq P_2P_1$ $(c_1, c_2) \in \mathcal{S}(P_1, P_2, k)$ [illegible] $c_1 + c_2 = 0$ [illegible] k [illegible] $c_1^k(P_1 - P_2)^{k+1} = P_1 - P_2$

[illegible]. The matrix equality is obvious. By Theorem 4.1 there are $\alpha, \beta \in \{0\} \cup \Omega_k$ such that $\alpha \neq \beta$ and $\alpha + \beta = 0$. It is clear that $\alpha \neq 0$ and $\beta \neq 0$. Thus, $1 = \alpha^k = (-\beta)^k = (-1)^k\beta^k = (-1)^k$. Now it is evident that k must be even. □

[illegible] Let $(y, z) \in \mathbb{C}^2$ and $\alpha \in \mathbb{C} \setminus \{1/2\}$ such that $\alpha^2 + yz = \alpha$ and $yz \neq 0$. The matrices

$$P_1 := \begin{bmatrix} \alpha & y \\ z & 1-\alpha \end{bmatrix} \qquad \text{and} \qquad P_2 := \begin{bmatrix} 1-\alpha & y \\ z & \alpha \end{bmatrix}$$

are projectors which do not commute. The nonzero solutions of $(c_1P_1 - c_1P_2)^{k+1} = c_1P_1 - c_1P_2$ must satisfy $c_1^k(2\alpha - 1)^k = 1$. Since $\alpha \neq 1/2$ is arbitrary, this shows that c_1 is also arbitrary. So, the conclusion on c_1 of Theorem 4.11 cannot be improved.

[illegible] 1. This theorem asserts that if there are nontrivial elements of $\mathcal{S}(P_1, P_2, k) \cap \{(c_1, c_2) \in \mathbb{C}^2 : c_1 + c_2 = 0\}$, then there are exactly k different nontrivial solutions and if $(c_1, -c_1)$ is one of them, then

$$\mathcal{S}(P_1, P_2, k) \cap \{(c_1, c_2) \in \mathbb{C}^2 : c_1 + c_2 = 0\} = \{(c_1, -c_1), \omega_k(c_1, -c_1), \ldots, \omega_k^{k-1}(c_1, -c_1)\}.$$

[illegible] 2. From Theorem 4.11, we deduce that $(P_1 - P_2)^{k+1}$ is a scalar multiple of $P_1 - P_2$ if and only if there are nontrivial solutions in $\mathcal{S}(P_1, P_2, k) \cap \{(c_1, c_2) \in \mathbb{C}^2 : c_1 + c_2 = 0\}$. Moreover, we can suppose that $k = 2m$ for certain $m \in \mathbb{N}$.

The calculation of $(P_1 - P_2)^{2m+1}$ can be done more efficiently by using the next result.

Proposition 4.12. [illegible] $A, B \in \mathbb{C}^{n\times n}$ [illegible] $m \in \mathbb{N}$ [illegible]

$$(A-B)^{2m+1} = A(I-BA)^m - (I-BA)^m B.$$

[illegible]. First, we shall prove the following claim:

$$B(I-BA)^m B = (I-BA)^m B \qquad \text{and} \qquad A(I-BA)^m A = A(I-BA)^m.$$

The proof of the first equality follows by induction. The case $m = 0$ is evident. If it is true for m, then

$$B(I-BA)^{m+1}B = B(I-BA)(I-BA)^m B = B(I-BA)^m B - BA(I-BA)^m B.$$

By the induction hypothesis, $B(I-BA)^m B = (I-BA)^m B$. Now, by using the obvious fact, [illegible] $p(t)$ [illegible] $M \in \mathbb{C}^{n\times n}$ [illegible] $p(M)$ [illegible] M" we get $BA(I-BA)^m B = (I-BA)^m BAB$. So

$$\begin{aligned} B(I-BA)^{m+1}B &= (I-BA)^m B - (I-BA)^m BAB \\ &= (I-BA)^m (B - BAB) = (I-BA)^{m+1}B. \end{aligned}$$

The other equality of the claim can be proved in a similar way. Now, the theorem will also be proved by induction. The case $m = 0$ is evident. If the theorem is true for m, then

$$\begin{aligned} (A-B)^{2(m+1)+1} &= (A-B)(A-B)^{2m+1}(A-B) \\ &= (A-B)\left[A(I-BA)^m - (I-BA)^m B\right](A-B) \\ &= A(I-BA)^m A - A(I-BA)^m BA - BA(I-BA)^m A \\ &\quad + B(I-BA)^m BA - A(I-BA)^m B + A(I-BA)^m B \\ &\quad + BA(I-BA)^m B - B(I-BA)^m B. \end{aligned}$$

Now, by using the claim we obtain

$$A(I-BA)^m A - A(I-BA)^m BA = A(I-BA)^m - ABA(I-BA)^m = A(I-BA)^{m+1}$$

and

$$BA(I-BA)^m B - B(I-BA)^m B = (I-BA)^m BAB - (I-BA)^m B = -(I-BA)^{m+1}B.$$

Since $B(I-BA)^m BA = BA(I-BA)^m = (I-BA)^m BA = BA(I-BA)^m A$, we get

$$(A-B)^{2(m+1)+1} = A(I-BA)^{m+1} - (I-BA)^{m+1}B.$$

This completes the proof. □

REFERENCES

[1] J. K. Baksalary, *Algebraic characterizations and statistical implications of the commutativity of orthogonal projectors*, in Proceedings of the Second International Tampere Conferences in Statistics (Tampere, Finland, 1987), T. Pukkila and S. Puntanen, eds., University of Tampere, Tampere, Finland, 1987, pp. 113–142.

[2] J. K. Baksalary and O. M. Baksalary, *Idempotency of linear combinations of two idempotent matrices*, Linear Algebra Appl., 321 (2000), pp. 3–7.

[3] J. K. Baksalary, O. M. Baksalary, and G. P. H. Styan, *Idempotency of linear combinations of an idempotent matrix and a tripotent matrix*, Linear Algebra Appl., 354 (2002), pp. 21–34.

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

[5] R. Bru and N. Thome, *Group inverse and group involutory matrices*, Linear and Multilinear Algebra, 45 (1998), pp. 207–218.

[6] S. L. Campbell and C. D. Meyer, Jr., *Generalized Inverses of Linear Tranformations,* Dover, New York, 1979.

[7] N. Castro González, J. J. Koliha, and Y. Wei, *Perturbation of the Drazin inverse for matrices with equal eigenprojections*, Linear Algebra Appl., 312 (2000), pp. 181–189.

[8] N. Castro González, J. J. Koliha, and Y. Wei, *Error bounds for perturbation of the Drazin inverse of closed operators with equal spectral projections*, Appl. Anal., 81 (2002), pp. 915–928.

[9] C. Coll and N. Thome, *Oblique projectors and group involutory matrices*, Appl. Math. Comput., 140 (2003), pp. 517–522.

[10] J. Gross and G. Trenkler, *Generalized and hypergeneralized projectors*, Linear Algebra Appl., 264 (1997), pp. 463–474.

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

[12] I. Marek and D. B. Szyld, *Comparison of convergence of general stationary iterative methods for singular matrices*, SIAM J. Matrix Anal. Appl., 24 (2002), pp. 68–77.

[13] C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and Its Applications*, John Wiley and Sons, New York, 1971.

[14] J. L. Stuart, *Reducible pattern k-potent ray pattern matrices*, Linear Algebra Appl., 362 (2003), pp. 87–99.

[15] J. L. Stuart, L. Beasley, and B. Shader, *Irreducible, pattern k-potent ray pattern matrices*, Linear Algebra Appl., 346 (2002), pp. 261–271.

[16] Y. Wei and G. Wang, *The perturbation theory for the Drazin inverse and its applications*, Linear Algebra Appl., 258 (1997), pp. 179–186.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 1, pp. 26–39

CONVERGENCE ANALYSIS OF STRUCTURE-PRESERVING DOUBLING ALGORITHMS FOR RICCATI-TYPE MATRIX EQUATIONS*

WEN-WEI LIN† AND SHU-FANG XU‡

Abstract. In this paper, we introduce the doubling transformation, a structure-preserving transformation for symplectic pencils, and present its basic properties. Based on these properties, a unified convergence theory for the structure-preserving doubling algorithms for a class of Riccati-type matrix equations is established, using only elementary matrix theory.

Key words. matrix equation, structure-preserving doubling algorithm, convergence rate

AMS subject classifications. 15A24, 65H10, 93B50, 93D15

DOI. 10.1137/040617650

1. Introduction. In this paper, we investigate the convergence of the structure-preserving doubling algorithms (SDAs) for the symmetric positive (semi)definite solutions to the following Riccati-type matrix equations:

- Continuous-time algebraic Riccati equation (CARE) [22, 27]:

$$-XGX + A^T X + XA + H = 0, \tag{1.1}$$

  where $A, H, G \in \mathbb{R}^{n\times n}$ with G and H being symmetric positive semidefinite.
- Discrete-time algebraic Riccati equation (DARE) [22, 27]:

$$X = A^T X(I + GX)^{-1}A + H, \tag{1.2}$$

  where $A, H, G \in \mathbb{R}^{n\times n}$ with G and H being symmetric positive semidefinite.
- Nonlinear matrix equation with the plus sign (NME-P) [3]:

$$X + A^T X^{-1} A = Q, \tag{1.3}$$

  where $A, Q \in \mathbb{R}^{n\times n}$ with Q being symmetric positive definite.
- Nonlinear matrix equation with the minus sign (NME-M) [12]:

$$X - A^T X^{-1} A = Q, \tag{1.4}$$

  where $A, Q \in \mathbb{R}^{n\times n}$ with Q being symmetric positive definite.

The Riccati-type matrix equations occur in many important applications (see [3, 12, 22, 27] and references therein). The nonlinear matrix equations CARE and DARE have been studied extensively (see [1, 2, 4, 5, 6, 7, 19, 8, 14, 15, 18, 20, 21, 22, 23, 24, 25, 26, 27, 29, 30, 31, 34]). Recently, the nonlinear matrix equations NME-P and NME-M have been studied in [3, 10, 11, 12, 16, 17, 28, 32, 35].

*Received by the editors October 26, 2004; accepted for publication (in revised form) by M. Chu August 30, 2005; published electronically March 17, 2006.
http://www.siam.org/journals/simax/28-1/61765.html

†Department of Mathematics, National Tsing Hua University, Hsinchu 300, Taiwan (wwlin@math.nthu.edu.tw).

‡LMAM, School of Mathematical Sciences, Peking University, Beijing 100871, China (xsf@pku.edu.cn). The research of this author was supported in part by the National Center for Theoretical Sciences in Taiwan.

A class of methods, referred to as doubling algorithms, attracted much interest in the 1970s and '80s (see [2] and references therein). These methods originate from the fixed-point iteration derived from the DARE:

$$X_{k+1} = A^T X_k (I + G X_k)^{-1} A + H.$$

Instead of generating the sequence $\{X_k\}$, doubling algorithms generate $\{X_{2^k}\}$. Doubling algorithms were largely forgotten in the past decade. Recently, doubling algorithms have been revived for DAREs and CAREs because of their nice numerical behavior—a quadratic convergence rate, low computational cost, and high numerical reliability, despite the lack of a rigorous error analysis (see [19, 9, 8]). For the NME-Ps and NME-Ms, Meini [28] and Guo [17] proposed iterative methods with a numerical behavior similar to that of the SDAs for DAREs and CAREs.

In this paper, by employing techniques similar to those in [8], we derive two SDAs for solving NME-Ps and NME-Ms, similar to those proposed by Meini in [28]. In general, we discover that our SDAs can be viewed as repeated applications of some special structure-preserving transformations to the associated symplectic pencils. We first introduce these doubling transformations, then develop a unified convergence theory for the SDAs, based on the nice properties of the doubling transformations using only elementary matrix theory.

Throughout this paper, the symbols $\|\cdot\|_2$ denote the matrix spectral norm. For a given $n \times n$ matrix A we use $\rho(A)$ to denote the spectral radius of A. For real symmetric matrices X and Y we write $X > Y$ ($X \geq Y$) if $X - Y$ is symmetric positive definite (semidefinite).

The paper is organized as follows. In section 2, we introduce a structure-preserving transformation for symplectic pencils and show its basic properties. In section 3, we analyze the convergence of the SDAs for the DARE and the CARE. In section 4, we derive the SDAs for solving the NME-P and the NME-M by using the doubling transformations, and establish the convergence theory of SDAs. Concluding remarks are given in section 5.

2. Doubling transformation. In this section, we introduce a structure-preserving transformation for symplectic pencils and investigate its basic properties. Based on the swapping and collapsing techniques in [4, 7, 6, 5], we begin with the definition of the transformation.

For $M, L \in \mathbb{R}^{2n\times 2n}$, let $M - \lambda L$ be a symplectic pencil, i.e.,

$$MJM^T = LJL^T, \qquad J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}. \tag{2.1}$$

Define

$$\mathcal{N}(M,L) = \left\{ [M_*, L_*] \ : \ M_*, L_* \in \mathbb{R}^{2n\times 2n},\ \operatorname{rank}[M_*, L_*] = 2n,\ [M_*, L_*] \begin{bmatrix} L \\ -M \end{bmatrix} = 0 \right\}. \tag{2.2}$$

Since $\operatorname{rank}\left[\begin{smallmatrix} L \\ -M \end{smallmatrix}\right] \leq 2n$, it follows that $\mathcal{N}(M,L) \neq \emptyset$. For any given $[M_*, L_*] \in \mathcal{N}(M,L)$, define

$$\widehat{M} = M_* M, \qquad \widehat{L} = L_* L. \tag{2.3}$$

The transformation

$$M - \lambda L \longrightarrow \widehat{M} - \lambda \widehat{L}$$

is called a [illegible].

An important feature of this kind of transformation is that it is structure-preserving, eigenspace-preserving, and eigenvalue-squaring, which has been shown in [4, 5, 33]. We quote the basic properties in the following theorem.

THEOREM 2.1. $\widehat{M}-\lambda\widehat{L}$

$M-\lambda L$

(a) $\widehat{M}-\lambda\widehat{L}$

(b) $M\begin{bmatrix}U\\V\end{bmatrix}=L\begin{bmatrix}U\\V\end{bmatrix}S$ $U,V\in\mathbb{R}^{n\times m}$ $S\in\mathbb{R}^{m\times m}$ $\widehat{M}\begin{bmatrix}U\\V\end{bmatrix}=\widehat{L}\begin{bmatrix}U\\V\end{bmatrix}S^2$

(c) $M-\lambda L$

$$WMZ=\begin{bmatrix}J_r & 0\\ 0 & I_{2n-r}\end{bmatrix},\qquad WLZ=\begin{bmatrix}I_r & 0\\ 0 & N_{2n-r}\end{bmatrix}, \tag{2.4}$$

W,Z J_r N_{2n-r} $\widehat{W}$

$$\widehat{W}\widehat{M}Z=\begin{bmatrix}J_r^2 & 0\\ 0 & I_{2n-r}\end{bmatrix},\qquad \widehat{W}\widehat{L}Z=\begin{bmatrix}I_r & 0\\ 0 & N_{2n-r}^2\end{bmatrix}. \tag{2.5}$$

2.1. (i) A subspace $\mathcal{W}$ of $\mathbb{R}^{2n}$ is called a of a pencil $M-\lambda L$ if $\mathcal{W}$ is spanned by the columns of $W=\begin{bmatrix}U\\V\end{bmatrix}$, where $U,V\in\mathbb{R}^{n\times m}$, and W has full column rank and satisfies $MW=LWS$ with $S\in\mathbb{R}^{m\times m}$. Therefore, part (b) of Theorem 2.1 tells us that if $\mathcal{W}$ is a generalized eigenspace of a symplectic pencil $M-\lambda L$, then it is still a generalized eigenspace after a doubling transformation. This is a cornerstone for the convergence theory of the SDAs for the Riccati-type matrix equations in the next two sections.

(ii) A pencil $M-\lambda L$ is called if $\det(M-\lambda L)$ does not vanish identically. It is well known that a pencil is regular if and only if it has a Kronecker canonical form as in (2.4). Thus, part (c) of Theorem 2.1 says that doubling transformations preserve regularity and that λ is a eigenvalue of $M-\lambda L$ if and only if λ^2 is an eigenvalue of $\widehat{M}-\lambda\widehat{L}$.

A symplectic pencil $M-\lambda L$ is said to be in (SSF-1) if it has the form

$$M=\begin{bmatrix}A & 0\\ -H & I\end{bmatrix},\qquad L=\begin{bmatrix}I & G\\ 0 & A^T\end{bmatrix}, \tag{2.6}$$

with $H,G\geq 0$; it is said to be in (SSF-2) if

$$M=\begin{bmatrix}A & 0\\ Q & -I\end{bmatrix},\qquad L=\begin{bmatrix}-P & I\\ A^T & 0\end{bmatrix}, \tag{2.7}$$

with $P,Q\geq 0$.

Note that one standard symplective form cannot be transformed to another by left nonsingular and right symplectic equivalence transformations unless G in (2.6) or P in (2.7) is positive definite. The following theorem shows that the two standard symplectic forms are preserved by an appropriate choice of doubling transformations.

THEOREM 2.2. (a) $M-\lambda L$ 1 $[M_*,L_*]\in\mathcal{N}(M,L)$ $\widehat{M}-\lambda\widehat{L}$ 1

(b) [illegible] $M - \lambda L$ [illegible] 2 [illegible] $Q - P > 0$ [illegible] $Q - A^T(Q-P)^{-1}A \geq 0$ [illegible] $[M_*, L_*] \in \mathcal{N}(M, L)$ [illegible] $\widehat{M} - \lambda\widehat{L}$ [illegible] 2

[illegible] (a) Applying block Gaussian elimination and row permutation to $\left[\begin{smallmatrix} L \\ -M \end{smallmatrix}\right]$, we get

$$M_* = \begin{bmatrix} A(I+GH)^{-1} & 0 \\ -A^T(I+HG)^{-1}H & I \end{bmatrix}, \qquad L_* = \begin{bmatrix} I & AG(I+HG)^{-1} \\ 0 & A^T(I+HG)^{-1} \end{bmatrix} \tag{2.8}$$

such that

$$M_*L = L_*M, \tag{2.9}$$

i.e., $[M_*, L_*] \in \mathcal{N}(M, L)$. Here the Sherman–Morrison–Woodbury formula (see, e.g., [13, p. 50]) is used. For more details, see [8]. We then compute L_*L and M_*M to produce

$$\widehat{M} = M_*M = \begin{bmatrix} \widehat{A} & 0 \\ -\widehat{H} & I \end{bmatrix}, \qquad \widehat{L} = L_*L = \begin{bmatrix} I & \widehat{G} \\ 0 & \widehat{A}^T \end{bmatrix}, \tag{2.10}$$

where

$$\widehat{A} = A(I+GH)^{-1}A, \tag{2.11}$$

$$\widehat{G} = G + AG(I+HG)^{-1}A^T, \tag{2.12}$$

$$\widehat{H} = H + A^T(I+HG)^{-1}HA. \tag{2.13}$$

It is clear that the resulting pencil is still in SSF-1.

(b) Similarly, under the condition $Q-P > 0$, we can compute $[M_*, L_*] \in \mathcal{N}(M, L)$ with

$$M_* = \begin{bmatrix} A(Q-P)^{-1} & 0 \\ -A^T(Q-P)^{-1} & I \end{bmatrix}, \qquad L_* = \begin{bmatrix} I & -A(Q-P)^{-1} \\ 0 & A^T(Q-P)^{-1} \end{bmatrix}. \tag{2.14}$$

Direct calculation gives rise to

$$\widehat{M} = M_*M = \begin{bmatrix} \widehat{A} & 0 \\ \widehat{Q} & -I \end{bmatrix}, \qquad \widehat{L} = L_*L = \begin{bmatrix} -\widehat{P} & I \\ \widehat{A}^T & 0 \end{bmatrix}, \tag{2.15}$$

where

$$\widehat{A} = A(Q-P)^{-1}A, \tag{2.16}$$

$$\widehat{Q} = Q - A^T(Q-P)^{-1}A, \tag{2.17}$$

$$\widehat{P} = P + A(Q-P)^{-1}A^T. \tag{2.18}$$

The assumption $Q - A^T(Q-P)^{-1}A \geq 0$ implies that the resulting pencil is still in SSF-2. □

[illegible] 2.2. The proof of Theorem 2.2 provided us with the well-defined computation formulae for the special structure-preserving doubling transformations, which is the basis for the SDAs for solving the Riccati-type matrix equations.

3. SDAs for preserving SSF-1. In this section, we first state the SDAs proposed in [8] and [9], respectively, for solving DAREs and CAREs. Then we use the technique established in the last section to develop the convergence theory of the SDAs.

3.1. SDA for solving DAREs. It is well known [27] that the DARE (1.2) has a symmetric positive semidefinite solution X (i.e., $X \geq 0$) if and only if X satisfies that

$$M \begin{bmatrix} I \\ X \end{bmatrix} = L \begin{bmatrix} I \\ X \end{bmatrix} S \tag{3.1}$$

for some stable matrix $S \in \mathbb{R}^{n\times n}$, where

$$M = \begin{bmatrix} A & 0 \\ -H & I \end{bmatrix}, \qquad L = \begin{bmatrix} I & G \\ 0 & A^T \end{bmatrix}. \tag{3.2}$$

Notice that the pencil $M - \lambda L$ is in SSF-1. Therefore, repeated applications of the special doubling transformation defined in (2.11)–(2.13) gives rise to the following structure-preserving doubling algorithm.

Algorithm SDA-1.

$$\begin{aligned} &A_0 = A, \quad G_0 = G, \quad H_0 = H, \\ &A_{k+1} = A_k(I + G_kH_k)^{-1}A_k, \\ &G_{k+1} = G_k + A_kG_k(I + H_kG_k)^{-1}A_k^T, \\ &H_{k+1} = H_k + A_k^T(I + H_kG_k)^{-1}H_kA_k. \end{aligned}$$

This is the SDA described in [8], in which extensive numerical experiments show that this algorithm is efficient and competitive.

3.2. SDA for solving CAREs. Assume that $X \geq 0$ solves the CARE (1.1). It is well known that the CARE (1.1) can be rewritten as

$$\mathcal{H} \begin{bmatrix} I \\ X \end{bmatrix} = \begin{bmatrix} I \\ X \end{bmatrix} R, \tag{3.3}$$

where

$$\mathcal{H} = \begin{bmatrix} A & -G \\ -H & -A^T \end{bmatrix}, \qquad R = A - GX.$$

The matrix $\mathcal{H}$ is a Hamiltonian matrix, i.e., $(\mathcal{H}J)^T = \mathcal{H}J$. Using a Cayley transformation with some appropriate $\gamma > 0$, we can transform (3.3) into the form

$$M \begin{bmatrix} I \\ X \end{bmatrix} = L \begin{bmatrix} I \\ X \end{bmatrix} S, \tag{3.4}$$

where

$$M = \mathcal{H} + \gamma I, \qquad L = \mathcal{H} - \gamma I, \qquad S = (R - \gamma I)^{-1}(R + \gamma I).$$

Now assume that we have chosen a $\gamma > 0$ such that the matrices

$$A_\gamma = A - \gamma I \quad \text{and} \quad W_\gamma = A_\gamma^T + HA_\gamma^{-1}G \tag{3.5}$$

are nonsingular. Chu, Fan, and Lin [9] proposed a method for computing γ such that both A_γ and W_γ are well conditioned. Let

$$T_1 = \begin{bmatrix} A_\gamma^{-1} & 0 \\ HA_\gamma^{-1} & I \end{bmatrix}, \qquad T_2 = \begin{bmatrix} I & -A_\gamma^{-1}GW_\gamma^{-1} \\ 0 & -W_\gamma^{-1} \end{bmatrix}, \tag{3.6}$$

which are obtained by alternately applying block Gaussian elimination to the matrices L and M (see [9] for more details). Then, direct calculations give rise to

$$\widehat{M} = T_2T_1M = \begin{bmatrix} \widehat{A} & 0 \\ -\widehat{H} & I \end{bmatrix}, \qquad \widehat{L} = T_2T_1L = \begin{bmatrix} I & \widehat{G} \\ 0 & \widehat{A}^T \end{bmatrix},$$

where

$$\widehat{A} = I + 2\gamma W_\gamma^{-T}, \qquad \widehat{G} = 2\gamma A_\gamma^{-1}GW_\gamma^{-1}, \qquad \widehat{H} = 2\gamma W_\gamma^{-1}HA_\gamma^{-1}.$$

Here the Sherman–Morrison–Woodbury formula is used. Since $\gamma > 0$ and $H, G \geq 0$ implies that $\widehat{G}, \widehat{H} \geq 0$, it follows that the resulting pencil $\widehat{M} - \lambda\widehat{L}$ is in SSF-1. In addition, it follows from (3.4) that

$$\widehat{M}\begin{bmatrix} I \\ X \end{bmatrix} = \widehat{L}\begin{bmatrix} I \\ X \end{bmatrix} S. \tag{3.7}$$

Thus, beginning with (3.7), following the same lines as SDA-1 for solving the DARE, we can construct a matrix sequence to approximate the unique symmetric positive semidefinite solution X to the CARE (1.1). For more details, see [9].

3.3. Convergence analysis of SDA-1. Now we establish the convergence theory of SDA-1 using Theorem 2.1. The main results are listed in the following theorem.

THEOREM 3.1. [illegible] $X, Y \geq 0$ [illegible]

$$X = A^TX(I + GX)^{-1}A + H, \tag{3.8}$$

$$Y = AY(I + HY)^{-1}A^T + G, \tag{3.9}$$

[illegible] $G, H \geq 0$ [illegible]

$$S = (I + GX)^{-1}A, \qquad T = (I + HY)^{-1}A^T. \tag{3.10}$$

[illegible] $\{A_k\}$ $\{G_k\}$ [illegible] $\{H_k\}$ [illegible] 1 [illegible]

(a) $A_k = (I + G_kX)S^{2^k}$.

(b) $H \leq H_k \leq H_{k+1} \leq X$ [illegible]

$$X - H_k = (S^T)^{2^k}(X + XG_kX)S^{2^k} \leq (S^T)^{2^k}(X + XYX)S^{2^k}; \tag{3.11}$$

(c) $G \leq G_k \leq G_{k+1} \leq Y$ [illegible]

$$Y - G_k = (T^T)^{2^k}(Y + YH_kY)T^{2^k} \leq (T^T)^{2^k}(Y + YXY)T^{2^k}. \tag{3.12}$$

[illegible] Notice that $U, V \geq 0$ implies that $I + UV$ is nonsingular and $V(I + UV)^{-1}, (I + UV)^{-1}U \geq 0$. It follows that SDA-1 is well defined and

$$H = H_0 \leq H_k \leq H_{k+1} \quad \text{and} \quad G = G_0 \leq G_k \leq G_{k+1}. \tag{3.13}$$

Define

$$M_k = \begin{bmatrix} A_k & 0 \\ -H_k & I \end{bmatrix}, \qquad L_k = \begin{bmatrix} I & G_k \\ 0 & A_k^T \end{bmatrix}.$$

Then the pencil $M_{k+1} - \lambda L_{k+1}$ is the result of doubling-transforming the pencil $M_k - \lambda L_k$. Since (3.8) implies

$$M_0 \begin{bmatrix} I \\ X \end{bmatrix} = L_0 \begin{bmatrix} I \\ X \end{bmatrix} S, \tag{3.14}$$

where S is defined by (3.10), repeated applications of part (b) of Theorem 2.1 produce

$$M_k \begin{bmatrix} I \\ X \end{bmatrix} = L_k \begin{bmatrix} I \\ X \end{bmatrix} S^{2^k}. \tag{3.15}$$

Equating the blocks of (3.15) then yields

$$A_k = (I + G_k X) S^{2^k}, \tag{3.16}$$

$$X - H_k = A_k^T X S^{2^k}. \tag{3.17}$$

Combining (3.16) with (3.17) gives rise to

$$X - H_k = (S^T)^{2^k} (X + X G_k X) S^{2^k}. \tag{3.18}$$

This, together with $(I + XG_k)X \geq 0$, implies that $X - H_k \geq 0$, i.e., $X \geq H_k$.

Similarly, (3.9) can be rewritten as

$$M_0 \begin{bmatrix} -Y \\ I \end{bmatrix} T = L_0 \begin{bmatrix} -Y \\ I \end{bmatrix}, \tag{3.19}$$

where T is defined by (3.10), and from (3.19) we can derive that

$$Y - G_k = (T^T)^{2^k} (Y + Y H_k Y) T^{2^k},$$

implying that $Y \geq G_k$. Thus, the theorem is proved. □

Let

$$W = \left[L \begin{bmatrix} I \\ X \end{bmatrix}, M \begin{bmatrix} -Y \\ I \end{bmatrix} \right], \qquad Z = \begin{bmatrix} I & -Y \\ X & I \end{bmatrix}.$$

Noting that $M_0 = M$, $L_0 = L$, and $X, Y \geq 0$, it follows from (3.14) and (3.19) that W and Z are nonsingular and satisfy

$$W^{-1} M Z = \begin{bmatrix} S & 0 \\ 0 & I \end{bmatrix}, \qquad W^{-1} L Z = \begin{bmatrix} I & 0 \\ 0 & T \end{bmatrix}.$$

Thus, by the spectral properties of symplectic pencils, it follows that if $\rho(S) < 1$, then we must have $\rho(T) = \rho(S) < 1$. In addition, it is well known that $0 \leq U \leq V$ implies that $\|U\|_2 \leq \|V\|_2$. Consequently from Theorem 3.1, we immediately get the following convergence result for SDA-1.

COROLLARY 3.2. [illegible] 3.1 [illegible] $\rho(S) < 1$ [illegible]

(a) $\|A_k\|_2 \le (1 + \|X\|_2\|Y\|_2)\|S^{2^k}\|_2 \longrightarrow 0$, as $k \to \infty$.
(b) $\|X - H_k\|_2 \le \|X + XYX\|_2\|S^{2^k}\|_2^2 \longrightarrow 0$, as $k \to \infty$.
(c) $\|Y - G_k\|_2 \le \|Y + YXY\|_2\|T^{2^k}\|_2^2 \longrightarrow 0$, as $k \to \infty$

REMARK 3.1. (i) Convergence results similar to those in Corollary 3.2 were obtained in [8]. In contrast to the work of [8], however, our analysis is simpler and our convergence results are stronger. In Theorem 3.1, we show explicit expressions of A_k, $X - H_k$, and $Y - G_k$, respectively. Furthermore, Corollary 3.2 contains simple upper bounds of $\|A_k\|_2$, $\|X - H_k\|_2$, and $\|Y - G_k\|_2$ in terms of $\|S\|_2$, X, and Y.

(ii) Again from parts (b) and (c) of Theorem 3.1, the matrix sequences $\{H_k\}$ and $\{G_k\}$ are monotonically increasing and bounded above, and hence there exist symmetric positive semidefinite matrices $\bar{H}$ and $\bar{G}$ such that

$$\lim_{k\to\infty} H_k = \bar{H}, \qquad \lim_{k\to\infty} G_k = \bar{G}.$$

Corollary 3.2 tells us that if $\rho(S) < 1$, then $X = \bar{H}$ and $Y = \bar{G}$.

REMARK 3.2. Let $G = BR^{-1}B^T \ge 0$, with $R > 0$, let $H = C^TC \ge 0$ in the DARE (3.8), and assume that (A, B) is stabilizable and (A, C) is detectable. Then it is well known that the DARE (3.8) and its dual (3.9), respectively, have symmetric positive semidefinite solutions X and Y, and that $\rho(S) < 1$ (see, e.g., [25, 29] for details). Thus the conditions in Corollary 3.2 are satisfied. In fact, it is easy to verify that if the DARE (3.8) and its dual (3.9), respectively, have symmetric positive semidefinite solutions X and Y, with $S = (I + GX)^{-1}A$ and $\rho(S) < 1$, then (A, B) is stabilizable and (A, C) is detectable. A similar argument also holds for the CARE (1.1) (see [9] for details).

4. SDAs for preserving SSF-2. In this section, we shall use the doubling transformations defined in the last section to derive two SDAs for solving the NME-Ps and NME-Ms. Then, we use the technique established in the last section to develop the convergence theory of these SDAs.

4.1. SDA for solving NME-Ps. It is easy to verify that the NME-P (1.3) has a symmetric positive definite X (i.e., $X > 0$) if and only if X satisfies

$$M \begin{bmatrix} I \\ X \end{bmatrix} = L \begin{bmatrix} I \\ X \end{bmatrix} S \tag{4.1}$$

for some matrix $S \in \mathbb{R}^{n\times n}$, where

$$M = \begin{bmatrix} A & 0 \\ Q & -I \end{bmatrix}, \qquad L = \begin{bmatrix} 0 & I \\ A^T & 0 \end{bmatrix}. \tag{4.2}$$

Notice that the pencil $M - \lambda L$ is in SSF-2. Therefore, applying the special doubling transformation defined in (2.16)–(2.18) repeatedly gives rise to the following SDA.

ALGORITHM SDA-2.

$$\begin{aligned} &A_0 = A, \quad Q_0 = Q, \quad P_0 = 0, \\ &A_{k+1} = A_k(Q_k - P_k)^{-1}A_k, \\ &Q_{k+1} = Q_k - A_k^T(Q_k - P_k)^{-1}A_k, \\ &P_{k+1} = P_k + A_k(Q_k - P_k)^{-1}A_k^T. \end{aligned}$$

REMARK 4.1. To ensure that the iterations in SDA-2 are well defined, the matrix $Q_k - P_k$ must be symmetric positive definite for all k. This can be guaranteed if the

NME-P (1.3) has a symmetric positive solution (see Theorem 4.1), as we shall prove below.

Remark 4.2. It is interesting to note that SDA-2 is essentially the same as Algorithm 3.1 proposed in [28] with $Q_k - P_k$ and Q_k in SDA-2 replaced by Q_k and X_k, respectively. In other words, Algorithm 3.1 in [28] is an SDA. It was pointed out that this algorithm has very nice numerical behavior, with quadratical convergence rate, low computational cost, and good numerical stability. For more details, see [28, 17].

4.2. SDA for solving NEM-Ms. It is proved in [12] that there always exists a unique positive definite solution X to the NME-M

$$X - A^T X^{-1} A = Q \tag{4.3}$$

and, moreover, that the spectral radius of $X^{-1}A$ is strictly less than 1. The solution X is closely related to the generalized eigenspace of the pencil

$$M - \lambda L = \begin{bmatrix} A & 0 \\ -Q & I \end{bmatrix} - \lambda \begin{bmatrix} 0 & I \\ A^T & 0 \end{bmatrix}. \tag{4.4}$$

In fact, it is easy to verify that a symmetric positive definite matrix X is a solution to the NME-M (4.3) if and only if X satisfies that

$$M \begin{bmatrix} I \\ X \end{bmatrix} = L \begin{bmatrix} I \\ X \end{bmatrix} S \tag{4.5}$$

for some matrix $S \in \mathbb{R}^{n\times n}$.

Although the pencil (4.4) is not symplectic, we can use the same technique as described in section 2 to transform it into a symplectic pencil. Take

$$M_* = \begin{bmatrix} AQ^{-1} & 0 \\ A^T Q^{-1} & -I \end{bmatrix}, \qquad L_* = \begin{bmatrix} I & AQ^{-1} \\ 0 & A^T Q^{-1} \end{bmatrix};$$

then we have

$$M_* L = L_* M. \tag{4.6}$$

Direct calculations lead to

$$\widehat{M}_0 = M_* M = \begin{bmatrix} \widehat{A} & 0 \\ \widehat{Q} & -I \end{bmatrix}, \qquad \widehat{L}_0 = L_* L = \begin{bmatrix} \widehat{P} & I \\ \widehat{A}^T & 0 \end{bmatrix},$$

where

$$\widehat{A} = AQ^{-1}A, \quad \widehat{Q} = Q + A^T Q^{-1} A, \quad \widehat{P} = AQ^{-1}A^T. \tag{4.7}$$

The pencil $\widehat{M}_0 - \lambda \widehat{L}_0$ is symplectic but neither an SSF-1 nor an SSF-2.

Assume that $X > 0$ is the unique symmetric positive solution to the NME-M (4.3). Then it satisfies (4.5) with $S = X^{-1}A$. From part (b) of Theorem 2.1, we have

$$\widehat{M}_0 \begin{bmatrix} I \\ X \end{bmatrix} = \widehat{L}_0 \begin{bmatrix} I \\ X \end{bmatrix} S^2. \tag{4.8}$$

Now let

$$\begin{bmatrix} I \\ \widehat{X} \end{bmatrix} = \begin{bmatrix} I & 0 \\ \widehat{P} & I \end{bmatrix} \begin{bmatrix} I \\ X \end{bmatrix}, \qquad \widehat{M} = \begin{bmatrix} \widehat{A} & 0 \\ \widehat{Q} + \widehat{P} & -I \end{bmatrix}, \qquad \widehat{L} = \begin{bmatrix} 0 & I \\ \widehat{A}^T & 0 \end{bmatrix}.$$

Then it follows from (4.8) that

$$\widehat{M} \begin{bmatrix} I \\ \widehat{X} \end{bmatrix} = \widehat{L} \begin{bmatrix} I \\ \widehat{X} \end{bmatrix} S^2. \tag{4.9}$$

Clearly, the pencil $\widehat{M} - \lambda \widehat{L}$ is in SSF-2. Thus, beginning with (4.9), following the same lines as SDA-2 for solving the NME-P (1.3), we can construct an approximating matrix sequence with limit $\widehat{X}$. Then the unique symmetric positive definite solution X to the NME-M (4.3) can be obtained by computing $X = \widehat{X} - \widehat{P}$.

4.3. Convergence analysis of SDA-2. Now we establish the convergence theory of SDA-2 based on Theorem 2.1. The main results are listed in the following theorem.

THEOREM 4.1. [illegible] $X > 0$ [illegible]

$$X + A^T X^{-1} A = Q, \tag{4.10}$$

[illegible] $Q > 0$ [illegible] $S = X^{-1}A$ [illegible] $\{A_k\}$ $\{Q_k\}$ [illegible] $\{P_k\}$ [illegible] 2 [illegible]

(a) $A_k = (X - P_k)S^{2^k}$.

(b) $0 \le P_k \le P_{k+1} < X$ [illegible]

$$Q_k - P_k = (X - P_k) + A_k^T (X - P_k)^{-1} A_k > 0; \tag{4.11}$$

(c) $X \le Q_{k+1} \le Q_k \le Q$ [illegible]

$$Q_k - X = (S^T)^{2^k} (X - P_k) S^{2^k} \le (S^T)^{2^k} X S^{2^k}. \tag{4.12}$$

[illegible] Using mathematical induction, denote

$$M_k = \begin{bmatrix} A_k & 0 \\ Q_k & -I \end{bmatrix}, \qquad L_k = \begin{bmatrix} -P_k & I \\ A_k^T & 0 \end{bmatrix},$$

where $P_0 = 0$.

For $k = 1$, since $Q_0 - P_0 = Q > 0$, it follows that A_1, Q_1, P_1 are all well defined. Using (4.10), we have

$$\begin{bmatrix} X & A \\ A^T & Q \end{bmatrix} = \begin{bmatrix} I & 0 \\ A^T X^{-1} & I \end{bmatrix} \begin{bmatrix} X & 0 \\ 0 & X \end{bmatrix} \begin{bmatrix} I & X^{-1}A \\ 0 & I \end{bmatrix} > 0. \tag{4.13}$$

Further computations yield

$$\begin{bmatrix} I & -AQ^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} X & A \\ A^T & Q \end{bmatrix} \begin{bmatrix} I & 0 \\ -Q^{-1}A^T & I \end{bmatrix} = \begin{bmatrix} X - AQ^{-1}A^T & 0 \\ 0 & Q \end{bmatrix}. \tag{4.14}$$

Combining (4.14) and (4.13), we obtain

$$X - P_1 = X - AQ^{-1}A^T > 0. \tag{4.15}$$

From (4.10), it is easy to verify that X satisfies

$$M_0 \begin{bmatrix} I \\ X \end{bmatrix} = L_0 \begin{bmatrix} I \\ X \end{bmatrix} S$$

with $S = X^{-1}A$. Since $M_1 - \lambda L_1$ is the result of doubling-transforming $M_0 - \lambda L_0$, part (b) of Theorem 2.1 leads to

$$M_1 \begin{bmatrix} I \\ X \end{bmatrix} = L_1 \begin{bmatrix} I \\ X \end{bmatrix} S^2. \tag{4.16}$$

Equating the blocks of (4.16) gives rise to

$$A_1 = (X - P_1)S^2, \qquad Q_1 - X = A_1^T S^2.$$

This, together with (4.15), implies that

$$Q_1 - P_1 = (X - P_1) + A_1^T (X - P_1)^{-1} A_1 > 0, \tag{4.17}$$

$$Q_1 - X = (S^T)^2 (X - P_1) S^2 \geq 0. \tag{4.18}$$

Obviously, the inequalities $Q = Q_0 \geq Q_1$ and $0 = P_0 \leq P_1$ hold. Thus, we have proved that the theorem is true for $k = 1$.

Next, considering the $k+1$ case, we assume that the theorem is true for all positive integers less than or equal to k. Since $Q_k - P_k > 0$, it follows that A_{k+1}, Q_{k+1}, P_{k+1} are all well defined. Similar to the proof of (4.15), (4.11) implies

$$X - P_{k+1} = (X - P_k) - A_k (Q_k - P_k)^{-1} A_k^T > 0.$$

Recall that $M_{j+1} - \lambda L_{j+1}$ is the result of doubling-transforming $M_j - \lambda L_j$ for $j = 0, 1, \ldots, k$. Applying part (b) of Theorem 2.1 $k+1$ times, we get

$$M_{k+1} \begin{bmatrix} I \\ X \end{bmatrix} = L_{k+1} \begin{bmatrix} I \\ X \end{bmatrix} S^{2^{k+1}}. \tag{4.19}$$

From (4.19), following the same lines as the proof of (4.17) and (4.18), it can be proved that

$$Q_{k+1} - P_{k+1} = (X - P_{k+1}) + A_{k+1}^T (X - P_{k+1})^{-1} A_{k+1} > 0,$$
$$Q_{k+1} - X = (S^T)^{2^{k+1}} (X - P_{k+1}) S^{2^{k+1}} \geq 0.$$

Clearly, $P_k \leq P_{k+1}$ and $Q_k \geq Q_{k+1}$. This shows that the theorem is also true for integers $k+1$. By induction principle, the theorem is true for all positive integers k. □

REMARK 4.3. Similar results were obtained in [28] by using properties of cyclic reduction and spectral properties of block Toeplitz matrices with nonnegative definite matrix-valued generating functions. In contrast, our analysis is simpler and the results are stronger. In Theorem 4.1, we show the explicit expressions of A_k and $Q_k - X$, as well as the monotonicity properties of $\{P_k\}$ and $\{Q_k\}$. Furthermore, in part (b) we prove that $Q_k - P_k$ is symmetric positive definite for all k, which guarantees that SDA-2 is well defined.

It was proved in [11] that if the NME-P (1.3) has a symmetric positive definite solution, then all symmetric solutions are positive definite with the maximal and

minimal solutions X_+ and X_-. Since Theorem 4.1 is true for any symmetric positive definite solution X, the following result follows immediately.

COROLLARY 4.2. [illegible] 4.1 [illegible]

$$Q_k - P_k > X_+ - X_- \geq 0$$

[illegible] k [illegible] X_+ [illegible] X_- [illegible] (1.3) [illegible]

In addition, from Theorem 4.1, we obtain the following corollary.

COROLLARY 4.3. [illegible] 4.1 [illegible] $\rho(S) < 1$ [illegible]

(a) $\|A_k\|_2 \leq \|X\|_2 \|S^{2^k}\|_2 \longrightarrow 0$ [illegible] $k \to \infty$.

(b) $\|X - Q_k\|_2 \leq \|X\|_2 \|S^{2^k}\|_2^2 \longrightarrow 0$ [illegible] $k \to \infty$

[illegible] 4.4. (i) Here we see that the upper bounds of $\|A_k\|_2$ and $\|X - Q_k\|_2$ are in terms of [illegible] X and $S \equiv X^{-1}A$.

(ii) By Theorem 4.1, the matrix sequence $\{Q_k\}$ is monotonically decreasing and bounded below by $X > 0$. Hence, there exists $\bar{Q} > 0$ such that $\lim_{k\to\infty} Q_k = \bar{Q}$. Corollary 4.3 tells us that if $\rho(S) < 1$, then $X = \bar{Q}$. In fact, X will then be the maximal solution of (1.3). Moreover, it has been proved that X is the maximal solution of (1.3) if and only if $\rho(S) \leq 1$ (see [17]). Now assuming that $X = X_+$ is the maximal solution of (1.3), it is natural to ask whether $\bar{Q} = X_+$ if $\rho(S) = 1$. In [17], Guo proved that if $\rho(S) = 1$ and all eigenvalues of S on the unit circle are semisimple, then $\bar{Q} = X_+$ is still true. In this case, the convergence is at least linear with rate $1/2$. When S has nonsemisimple unimodular eigenvalues, it is unclear whether $\bar{Q} = X_+$.

[illegible] 4.5. It is proved that the NME-P (1.3) has a symmetric positive definite solution X if and only if the nonlinear matrix equation

$$Y + AY^{-1}A^T = Q \tag{4.20}$$

has a symmetric positive solution Y (see, for e.g., [28]). Assume that the maximal solution of (4.20) is Y_+. Then it follows from (4.20) that

$$\begin{bmatrix} A & 0 \\ Q & -I \end{bmatrix} \begin{bmatrix} I \\ Q - Y_+ \end{bmatrix} T = \begin{bmatrix} 0 & I \\ A^T & 0 \end{bmatrix} \begin{bmatrix} I \\ Q - Y_+ \end{bmatrix}, \tag{4.21}$$

where $T = Y_+^{-1}A^T$. Similar to the proof of (4.12), we can show from (4.21) that

$$0 \leq Q - Y_+ - P_k = (T^T)^{2^k}(Q_k - Q + Y_+)T^{2^k} \leq (T^T)^{2^k} Y_+ T^{2^k},$$

where P_k and Q_k are generated by SDA-2. Since $\rho(T) = \rho(Y_+^{-1}A^T) = \rho(X_+^{-1}A)$ (see, e.g., [28]), where X_+ is the maximal solution of the NME-P (1.3), it follows that $\lim_{k\to\infty} P_k = Q - Y_+$ under the conditions of Corollary 4.3. If A is nonsingular, then $X_- = Q - Y_+$ (see [28]), where X_- is the minimal solution of the NME-P (1.3), and thus in this case we have $\lim_{k\to\infty} P_k = X_-$.

[illegible] 4.6. Since $\lim_{k\to\infty}(Q_k - P_k) = X_+ - X_-$ if A is nonsingular and $\rho(S) < 1$, the lower bound $X_+ - X_-$ in Corollary 4.2 is the best one. However, $X_+ - X_-$ may be singular and, indeed, it can be the zero matrix. For example, the NME-P with $Q = I$ and $A = \frac{1}{2}I$ has $X_+ = X_- = \frac{1}{2}I$.

5. Conclusions. In this paper, we have introduced a structure-preserving transformation for symplectic pencils, referred to as the doubling transformation, and investigated its basic properties. Based on these nice properties, a unified convergence theory for the SDAs for solving a class of Riccati-type matrix equations has been established, using only elementary matrix theory.

REFERENCES

[1] G. Ammar and V. Mehrmann, *On Hamiltonian and symplectic Hessenberg forms*, Linear Algebra Appl., 149 (1991), pp. 55–72.

[2] B. Anderson, *Second-order convergent algorithms for the steady-state Riccati equation*, Internat. J. Control, 28 (1978), pp. 295–306.

[3] W. N. Anderson, T. D. Morley, and G. E. Trapp, *Positive solutions to* $X = A - BX^{-1}B^*$, Linear Algebra Appl., 134 (1990), pp. 53–62.

[4] Z. Bai, J. Demmel, and M. Gu, *An inverse free parallel spectral divide and conquer algorithm for nonsymmetric eigenproblems*, Numer. Math., 76 (1997), pp. 279–308.

[5] P. Benner, *Contributions to the Numerical Solution of Algebra Riccati Equations and Related Eigenvalue Problems*, Logos-Verlag, Berlin, 1997.

[6] P. Benner and R. Byers, *Disk functions and their relationship to the matrix sign function*, in Proceedings of the European Control Conference ECC97, BELWARE Information Technology, Waterloo, Belgium, 1997, Paper 936 (CD-ROM).

[7] P. Benner and R. Byers, *Evaluating products of matrix pencils and collapsing matrix products*, Numer. Linear Algebra Appl., 8 (2001), pp. 357–380.

[8] E. K.-W. Chu, H.-Y. Fan, W.-W. Lin, and C.-S. Wang, *A structure-preserving doubling algorithm for periodic discrete-time algebraic Riccati equations*, Internat. J. Control, 77 (2004), pp. 767–788.

[9] 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.

[10] 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.

[11] 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.

[12] A. Ferrante and B. C. Levy, *Hermitian solutions of the equation* $X = Q + NX^{-1}N^*$, Linear Algebra Appl., 247 (1996), pp. 359–373.

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

[14] T. Gudmundsson, C. Kenney, and A. J. Laub, *Scaling of the discrete-time algebraic Riccati equation to enhance stability of the Schur solution method*, IEEE Trans. Automat. Control, 37 (1992), pp. 513–518.

[15] 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.

[16] C.-H. Guo and P. Lancaster, *Iterative solution of two matrix equations*, Math. Comp., 68 (1999), pp. 1589–1603.

[17] C.-H. Guo, *Comvergence rate of an iterative method for a nonlinear matrix equation*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 295–302.

[18] J. J. Hench and A. J. Laub, *Numerical solution of the discrete-time periodic Riccati equation*, IEEE Trans. Automat. Control, 39 (1994), pp. 1197–1210.

[19] 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.

[20] M. Kimura, *Convergence of the doubling algorithm for the discrete-time algebraic Riccati equation*, Internat. J. Systems Sci., 19 (1988), pp. 701–711.

[21] D. Lainiotis, N. Assimakis, and S. Katsikas, *New doubling algorithm for the discrete periodic Riccati equation*, Appl. Math. Comput., 60 (1994), pp. 265–283.

[22] P. Lancaster and L. Rodman, *Algebraic Riccati Equations*, Oxford University Press, Oxford, UK, 1995.

[23] A. J. Laub, *A Schur method for solving algebraic Riccati equations*, IEEE Trans. Automat. Control, 24 (1979), pp. 913–921.

[24] A. J. Laub, *Invariant subspace methods for numerical solution of Riccati equations*, in The Riccati Equation, S. Bittanti, A. J. Laub, and J. C. Willems, eds., Springer-Verlag, Berlin, 1991, pp. 163–196.

[25] L.-Z. Lu and W.-W. Lin, *An iterative algorithm for the solution of the discrete time algebraic Riccati equations*, Linear Algebra Appl., 189 (1993), pp. 465–488.

[26] L.-Z. Lu, W.-W. Lin, and C. E. M. Pearce, *An efficient algorithm for the discrete-time algebraic Riccati equation*, IEEE Trans. Automat. Control, 44 (1999), pp. 1216–1220.

[27] V. Mehrmann, *The Autonomous Linear Quadratic Control Problem*, Lecture Notes in Control and Inform. Sci. 163, Springer-Verlag, Berlin, 1991.

[28] 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 (2001), pp. 1189–1204.

[29] C. Paige and C. Van Loan, *A Schur decomposition for Hamiltonian matrices*, Linear Algebra Appl., 41 (1981), pp. 11–32.

[30] T. Pappas, A. J. Laub, and N. R. Sandell, *On the numerical solution of the discrete-time algebraic Riccati equation*, IEEE Trans. Automat. Control, 25 (1980), pp. 631–641.

[31] J.-G. Sun, *Sensitivity analysis of the discrete-time algebraic Riccati equation*, Linear Algebra Appl., 275/276 (1998), pp. 595–615.

[32] J.-G. Sun and S. F. Xu, *Perturbation analysis of the maximal solution of the matrix equation* $X + A^TX^{-1}A = P$. II, Linear Algebra Appl., 362 (2003), pp. 211–228.

[33] X. Sun and E. S. Quintana-Ortí, *Spectral division methods for block generalized Schur decompositions*, Math. Comp., 73 (2004), pp. 1827–1847.

[34] P. Van Dooren, *A generalized eigenvalue approach for solving Riccati equations*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 121–135.

[35] S. F. Xu, *On the maximal solution of the matrix equation* $X + A^TX^{-1}A = I$, Acta Sci. Natur. Univ. Pekinensis, 36 (2000), pp. 29–38.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 1, pp. 40–67

AN ERROR ANALYSIS OF A UNITARY HESSENBERG *QR* ALGORITHM*

MICHAEL STEWART†

Abstract. This paper proves the stability of a variant of Gragg's unitary Hessenberg QR algorithm (UHQR). It is shown that a single UHQR iteration with numerically unimodular shift applied to the Schur parameters and complementary parameters of a unitary Hessenberg matrix H gives computed parameters for $Q^{\mathrm{H}}HQ$ that are close to those obtained from a perturbed set of parameters using a perturbed shift with no numerical error. The perturbations of the parameters and the shift are bounded.

Key words. unitary eigenvalue problem, unitary Hessenberg matrix

AMS subject classifications. 65F15, 65G50

DOI. 10.1137/04061948X

1. Background. If an unreduced unitary Hessenberg matrix is scaled by a diagonal similarity so as to have real positive subdiagonal, then it has the form

$$H = \begin{bmatrix} -\overline{a}_0 a_1 & -\overline{a}_0 b_1 a_2 & -\overline{a}_0 b_1 b_2 a_3 & \cdots & -\overline{a}_0 b_1 \cdots b_{n-1} a_n \\ b_1 & -\overline{a}_1 a_2 & -\overline{a}_1 b_2 a_3 & \cdots & -\overline{a}_1 b_2 \cdots b_{n-1} a_n \\ & b_2 & -\overline{a}_2 a_3 & & \vdots \\ & & \ddots & \ddots & \vdots \\ & & & b_{n-1} & -\overline{a}_{n-1} a_n \end{bmatrix}, \tag{1}$$

where b_k is real, $b_k > 0$, $|a_k|^2 + b_k^2 = 1$ for $1 \le k < n$, and $|a_n| = 1$. We also assume that $a_0 = 1$. The a_k are the [illegible] and the b_k are the [illegible] of H.

A unitary Hessenberg matrix can also be written as a product of modified elementary rotations

$$H = G_1(a_1) G_2(a_2) \cdots G_n(a_n), \tag{2}$$

where

$$G_k(a_k) = I_{k-1} \oplus \begin{bmatrix} -a_k & b_k \\ b_k & \overline{a}_k \end{bmatrix} \oplus I_{n-k-1}$$

for $1 \le k < n$ and

$$G_n(a_n) = I_{n-1} \oplus (-a_n).$$

The parameters in the rotations are the same as the parameters in (1).

Unitary Hessenberg structure is similar to symmetric tridiagonal structure in that it can be exploited by the QR algorithm. If we do not accumulate eigenvectors,

*Received by the editors November 23, 2004; accepted for publication (in revised form) by Q. Ye September 15, 2005; published electronically March 17, 2006.
http://www.siam.org/journals/simax/28-1/61948.html
†Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30303 (mastewart@gsu.edu).

then the unitary Hessenberg QR (UHQR) algorithm requires $O(n^2)$ floating point operations [3]. As originally presented the algorithm was unstable. A stable algorithm based on the application of a parameterized QZ algorithm to a structured matrix pencil with the same eigenvalues as H was analyzed in [1].

In [4] the main source of the instability in a rational variant of the UHQR algorithm was identified as cancellation in a single formula. A stabilizing alternate formula was proposed in the same paper and numerical results suggested that the modified algorithm was numerically stable. The main features of an error analysis of a rational UHQR algorithm were summarized in [6] where it was shown that an additional modificaton to enforce the normalization condition $|a_k|^2 + b_k^2 = 1$ results in a provably stable algorithm. In this paper we prove the numerical stability of a simpler modification for which the normalization errors on a_k and b_k can be bounded over repeated QR iterations without explicit renormalization. The algorithm considered here admits a somewhat simpler proof of stability than that of [6].

2. The algorithm. For a particular choice of shift z one iteration of the QR algorithm computes

$$H - zI = QR$$

and

$$\hat{H} = RQ + zI = Q^{\mathrm{H}} H Q,$$

where

$$Q = G_1(c_1)G_2(c_2)\cdots G_n(c_n)$$

is a product of plane rotations computed to zero the subdiagonal elements of $H - zI$ and give real, positive values on the diagonal of R. Clearly $\hat{H}$ is a unitary Hessenberg matrix. If we let the Schur parameters of $\hat{H}$ be $\hat{a}_k$ and $\hat{b}_k$, then the UHQR algorithm efficiently computes the parameters $\hat{a}_k$ and $\hat{b}_k$ from a_k and b_k. The algorithm also computes the cosines c_k and the corresponding real sines s_k. While $G_n(c_n)$ is not needed to triangularize $Q^{\mathrm{H}}(H - zI) = R$, it is required to make the last diagonal element of R real and nonnegative. We assume that H is unreduced (i.e., that $b_k > 0$ for $k = 1, 2, \ldots, n-1$). If this is not the case, then the problem deflates in the usual way. Algorithm 1 is the UHQR algorithm in the form originally given in [3].

Algorithm 1: UHQR

$c_0 = d_0 = 1, \qquad s_0 = 1$
for $k = 1 : n-1$
  $p_k = zc_{k-1} + a_k d_{k-1}$
  $q_k = \overline{a}_k z c_{k-1} + d_{k-1}$
  $r_k = \sqrt{|p_k|^2 + b_k^2}$
  $c_k = p_k / r_k, \qquad s_k = b_k / r_k, \qquad d_k = q_k / r_k$
  $\hat{b}_{k-1} = r_k s_{k-1}$
  $\hat{a}_k = c_k \overline{d}_k - \overline{z} s_k^2 a_{k+1}$
end
$p_n = zc_{n-1} + a_n d_{n-1}, \qquad r_n = |p_n|$
$\hat{b}_{n-1} = r_n s_{n-1}, \qquad \hat{a}_n = a_n$
if $r_n > 0$, **then** $c_n = \mathrm{sign}(p_n)$
  else $c_n = 1$

In its original formulation the algorithm is numerically unstable. Three modifications lead to a provably stable algorithm. The first, from [4], gives an algorithm, specialized to the case $|z| = 1$, from which square roots are easily eliminated if the eigenvectors are not needed. The resulting algorithm is still unstable. A further modification of a single formula, also proposed in [4], stabilizes the algorithm in practice. Unfortunately, it is not possible to prove stability. The difficulty is that there is no guarantee that the normalization errors on the Schur parameters remain small. If

$$|a_k|^2 + b_k^2 = 1 + \delta_k,$$

where δ_k represents the normalization error on a_k and b_k, then the analysis given in this paper suggests that δ_k could grow exponentially with repeated UHQR iterations.

From this point on we assume that $|z| = 1$. Even with this restriction the shift can be chosen to ensure global quadratic convergence. In [8] it is shown that the QR algorithm applied to a unitary Hessenberg matrix converges globally and quadratically if z is chosen to be the eigenvalue of the 2×2 unitary matrix

$$\begin{bmatrix} -\overline{a}_{n-2}a_{n-1}/|a_{n-2}| & -\overline{a}_{n-2}b_{n-1}a_n/|a_{n-2}| \\ b_{n-1} & -\overline{a}_{n-1}a_n \end{bmatrix} \tag{3}$$

which is closer to $-\overline{a}_{n-1}a_n$. If $a_{n-2} = 0$, then the factor $a_{n-2}/|a_{n-2}|$ is replaced by a_n. This is the [illegible]

Instead of computing both c_k and d_k we work with c_k and the ratio

$$f_k = \frac{c_k}{d_k} = \frac{p_k}{q_k}.$$

If we define the Szegö polynomial $\chi_k(z) = \det(zI - H_k)$, where H_k is the $k \times k$ leading principal submatrix of H, then it can be shown that

$$p_k = \frac{\chi_k(z)}{r_1 r_2 \cdots r_{k-1}}, \qquad \text{and} \qquad q_k = \frac{z^k \overline{\chi}_k(1/z)}{r_1 r_2 \cdots r_{k-1}}.$$

If H is unreduced, then it follows from well-known properties of Szegö polynomials [7] that for $k = 1, 2, \ldots, n-1$ the polynomial $\chi_k(z)$ has all zeros strictly within the unit circle and $z^k\overline{\chi}_k(1/z)$ has all zeros strictly outside the unit circle; thus for $|z| = 1$, $q_k \neq 0$. It also follows that $z^k\overline{\chi}_k(1/z) = z^k\overline{\chi_k(z)}$ so that $|p_k| = |q_k|$ and $|f_k| = 1$. Further details on the connection of Szegö polynomials and the UHQR algorithm can be found in [3].

If we define

$$w_k = z f_{k-1} \qquad \text{and} \qquad g_k = w_k + a_k,$$

then we can compute f_k from

$$f_k = \frac{z c_{k-1} + a_k d_{k-1}}{\overline{a}_k z c_{k-1} + d_{k-1}} = \frac{z f_{k-1} + a_k}{\overline{a}_k z f_{k-1} + 1} = \frac{g_k}{\overline{g}_k w_k} = \frac{\overline{w}_k g_k}{\overline{g}_k}.$$

We can compute p_k from g_k, f_{k-1}, and c_{k-1} using

$$p_k = z c_{k-1} + a_k d_{k-1} = (z + a_k \overline{f}_{k-1}) c_{k-1} = g_k \overline{f}_{k-1} c_{k-1},$$

where we have used the fact that $|f_{k-1}| = 1$. Since $c_k = f_k d_k$ so that $|d_k| = |c_k|$ we have

$$\hat{a}_k = |c_k|^2 f_k - \overline{z} s_k^2 a_{k+1}.$$

The stabilizing formula of [4] is an alternate formula for g_k. We define $t_k = \overline{a}_k z f_{k-1}$ so that

$$g_k = z f_{k-1} + a_k = z f_{k-1}(1 + \overline{t}_k).$$

If $\mathrm{Re}(t_k) \geq 0$, then there is no cancellation and it can be shown that g_k is computed to high relative accuracy from z, f_{k-1}, and a_k. However, if $\mathrm{Re}(t_k) < 0$, then the formula can be inaccurate. In the latter case it was proposed in [4] to use

$$\begin{aligned} g_k &= z f_{k-1} + a_k \\ &= \frac{(z f_{k-1} + a_k)\overline{(z f_{k-1} - a_k)}}{\overline{(z f_{k-1} - a_k)}} \\ &= \frac{1 - |a_k|^2 - 2i\,\mathrm{Imag}(\overline{a}_k z f_{k-1})}{\overline{(z f_{k-1} - a_k)}} \\ &= \frac{b_k^2 - 2i\,\mathrm{Imag}(t_k)}{\overline{(z f_{k-1} - a_k)}}, \end{aligned} \tag{4}$$

where $x = \mathrm{Re}(x) + i\,\mathrm{Imag}(x)$. The use of (4) to compute g_k when $\mathrm{Re}(t_k) < 0$ dramatically improves the numerical properties of the algorithm. However, if we always apply (4) when $\mathrm{Re}(t_k) < 0$, then we cannot prove that the normalization errors on the Schur parameters remain small.

There are two possible ways to prevent the growth of normalization errors: the parameters a_k and b_k can be explicitly renormalized with every iteration or we can further restrict the use of the alternate formula (4). We take the latter approach and use (4) only when

$$\mathrm{Re}(t_k) < 0 \qquad \text{and} \qquad |a_k| > \frac{\sqrt{2}}{2}.$$

We show in section 4 that this additional restriction stabilizes the propagation of normalization errors. This results in the following stabilized algorithm.

Algorithm 2: Stabilized UHQR

$z = z/|z|$
$f_0 = c_0 = 1, \qquad s_0 = 0$
for $k = 1 : n-1$
  $w_k = z f_{k-1}, \qquad t_k = \overline{a}_k w_k$
  if $\mathrm{Re}(t_k) \geq 0$ **or** $|a_k| \leq \frac{\sqrt{2}}{2}$, **then** $g_k = w_k + a_k$
    else $g_k = (b_k^2 - 2i\mathrm{Imag}(t_k))/(\overline{w}_k - \overline{a}_k)$
  $p_k = g_k \overline{f}_{k-1} c_{k-1}$
  $r_k = \mathrm{sqrt}(|p_k|^2 + b_k^2), \qquad \widehat{b}_{k-1} = r_k s_{k-1}$
  $c_k = p_k / r_k, \qquad s_k = b_k / r_k$
  $f_k = \overline{w}_k g_k / \overline{g}_k$
  $\hat{a}_k = |c_k|^2 f_k - \overline{z} s_k^2 a_{k+1}$
end
$w_n = z f_{n-1} \qquad t_n = \overline{a}_n w_n$
if $\mathrm{Re}(t_n) \geq 0$, **then** $g_n = w_n + a_n$
  else $g_n = (-2i\mathrm{Imag}(t_n))/(\overline{w}_n - \overline{a}_n)$
$p_n = g_n \overline{f}_{n-1} c_{n-1}$
$r_n = |p_n|, \qquad \hat{b}_{n-1} = r_n s_{n-1}$

$\hat{a}_n = a_n$
if $r_n > 0$, **then** $c_n = \text{sign}(p_n)$
else $c_n = 1$

The algorithm uses several divisions that might be of some concern. The matrix H is unreduced so that in the absence of underflow in the computation of b_k^2 we have $r_k \neq 0$. Underflow of b_k^2 should justify deflation. Thus the computation of $c_k = p_k/r_k$ and $s_k = b_k/r_k$ will not result in division by zero when H is numerically unreduced. In the absence of numerical error $w_k = 1$ so that $g_k = w_k + a_k = 0$ implies $|a_k| = 1$ and $b_k = 0$. Thus $b_k \neq 0$ implies $g_k \neq 0$. To state conditions under which $g_k \neq 0$ in finite precision arithmetic requires a bound on the numerical errors in the normalization relation $|w_k| = 1$. This will be fully dealt with in section 4.

If the eigenvectors are not desired, then Algorithm 2 can be converted to a rational algorithm that avoids the use of square roots. In particular we can replace the computation of p_k, r_k, $\hat{b}_{k-1}$, c_k, and s_k with

$$|p_k|^2 = |g_k|^2|c_{k-1}|^2, \qquad r_k^2 = |p_k|^2 + b_k^2, \qquad \hat{b}_{k-1}^2 = r_k^2 s_{k-1}^2,$$

$$|c_k|^2 = |p_k|^2/r_k^2, \qquad \text{and} \qquad s_k^2 = b_k^2/r_k^2.$$

This is the form of the algorithm emphasized in [4]. The analysis of this paper can be modified to apply to the rational form of the algorithm with only minor modifications.

3. General issues. The analysis involves both forward and backward errors. The backward errors are placed in a structured way on the Schur parameters a_k, the complementary parameters b_k, and the shift z instead of in an unstructured way on the matrix H. Similarly the forward errors are placed on the quantities computed by the algorithm and not in an unstructured way on $\hat{H}$ or Q.

It is not reasonable to assume that we start with perfectly normalized a_k and b_k and exactly unimodular z. Instead we assume that

$$|a_k|^2 + b_k^2 = 1 + \delta_k, \tag{5}$$

where $b_n = 0$ so that $|a_n|^2 = 1 + \delta_n$ and

$$z = \tilde{z}(1 + \delta_z), \tag{6}$$

where δ_k and $\delta_z = |z| - 1$ are assumed to be of the order of the unit roundoff. The shift $\tilde{z}$ is not the exact projected Wilkinson shift. Instead, since

$$|\tilde{z}| = \frac{|z|}{|1 + \delta_z|} = \frac{|z|}{|z|} = 1,$$

it is the projection of an arbitrary computed shift on the unit circle. If the shift is computed in such a way as to guarantee numerical unimodularity, then $|\delta_z|$ can be bounded as a small multiple of the unit roundoff. This can be guaranteed by the explicit normalization $z = z/|z|$.

We use δ as an upper bound on the normalization errors on the Schur parameters so that

$$\delta = \max_{1 \leq k \leq n} |\delta_k|.$$

With ϵ defined to be the unit roundoff, bounds on the forward and backward errors are written in terms of the three basic sources of error ϵ, δ, and δ_z. All three errors must be small for the resulting bounds to imply stability. The shift can be explicitly normalized to keep δ_z small. In the case of δ, we must prove that repeated UHQR iterations do not cause dramatic growth in the normalization errors on the Schur parameters.

Structured bounds with errors on the Schur parameters imply normwise error bounds. For the moment we assume that $\tilde{a}_k$, $\tilde{b}_k$, and $\tilde{z}$ are such that the computed $\hat{a}_k$, $\hat{b}_k$, s_k, and c_k are close to the quantities $\tilde{\hat{a}}_k$, $\tilde{\hat{b}}_k$, $\tilde{s}_k$, and $\tilde{c}_k$ computed by stabilized UHQR applied to $\tilde{a}_k$, $\tilde{b}_k$, and $\tilde{z}$ without error. We have already chosen $\tilde{z}$ so that $|\tilde{z}| = 1$. The perturbed Schur parameters are constructed to satisfy $|\tilde{a}_k|^2 + \tilde{b}_k^2 = 1$ and $|\tilde{a}_n| = 1$. Throughout the analysis a tilde indicates a quantity computed from $\tilde{a}_k$, $\tilde{b}_k$, and $\tilde{z}$ without error.

It follows from the fact that the tilde quantities are computed without error that

$$\tilde{\hat{H}} = \tilde{Q}^{\mathrm{H}} \tilde{H} \tilde{Q} \qquad \text{and} \qquad \tilde{Q}^{\mathrm{H}} \tilde{Q} = I, \tag{7}$$

where

$$\tilde{H} = G_1(\tilde{a}_1) G_2(\tilde{a}_2) \cdots G_n(\tilde{a}_n),$$

$$\tilde{\hat{H}} = G_1(\tilde{\hat{a}}_1) G_2(\tilde{\hat{a}}_2) \cdots G_n(\tilde{\hat{a}}_n),$$

and

$$\tilde{Q} = G_1(\tilde{c}_1) G_2(\tilde{c}_2) \cdots G_n(\tilde{c}_n).$$

In proving the stability of the algorithm we give bounds for $K_a(\epsilon, \delta, \delta_z, k)$ and $K_b(\epsilon, \delta, \delta_z, k)$ in the relative backward error bounds

$$|\tilde{a}_k - a_k| \le |\tilde{a}_k| K_a(\epsilon, \delta, \delta_z, k) \qquad \text{and} \qquad \left|\tilde{b}_k - b_k\right| \le |\tilde{b}_k| K_b(\epsilon, \delta, \delta_z, k),$$

for $K_{\hat{b}}(\epsilon, \delta, \delta_z, k)$, $K_c(\epsilon, \delta, \delta_z, k)$, and $K_s(\epsilon, \delta, \delta_z, k)$ in the relative forward error bounds

$$\left|\tilde{\hat{b}}_k - \hat{b}_k\right| \le |\tilde{\hat{b}}_k| K_{\hat{b}}(\epsilon, \delta, \delta_z, k), \qquad |\tilde{c}_k - c_k| \le |\tilde{c}_k| K_c(\epsilon, \delta, \delta_z, k),$$

and

$$|\tilde{s}_k - s_k| \le |\tilde{s}_k| K_s(\epsilon, \delta, \delta_z, k),$$

and for $K_{\hat{a}}(\epsilon, \delta, \delta_z, k)$ in the absolute forward error bound

$$|\tilde{\hat{a}}_k - \hat{a}_k| \le K_{\hat{a}}(\epsilon, \delta, \delta_z, k).$$

The use of an absolute error bound for a_k is a result of the details of the analysis. In contrast to the other quantities, it does not appear that there is a satisfactory relative error bound for a_k.

Define

$$\hat{H} = G_1(\hat{a}_1, \hat{b}_1) G_2(\hat{a}_2, \hat{b}_2) \cdots G_{n-1}(\hat{a}_{n-1}, \hat{b}_{n-1}) G_n(\hat{a}_n)$$

and

$$Q = G_1(c_1, s_1)G_2(c_2, s_2)\cdots G_{n-1}(c_{n-1}, s_{n-1})G_n(c_n)$$

with

$$G_k(a_k, b_k) = I_{k-1} \oplus \begin{bmatrix} -a_k & b_k \\ b_k & \overline{a}_k \end{bmatrix} \oplus I_{n-k-1} \qquad \text{and} \qquad G_n(a_n) = I_{n-1} \oplus (-a_n).$$

We have shown explicitly the dependence of $G_k(\hat{a}_k, \hat{b}_k)$ on $\hat{b}_k$ since, when dealing with computed quantities that are not exactly normalized, $\hat{b}_k$ is not determined by $\hat{a}_k$.

The above error bounds for the Schur parameters imply that

$$\tilde{H} = H + E \qquad \text{and} \qquad \tilde{\hat{H}} = \hat{H} + F,$$

where

$$\|E\|_2 \leq \sum_{k=1}^{n} 2\max(K_a(\epsilon, \delta, \delta_z, k), K_b(\epsilon, \delta, \delta_z, k)) + O(\epsilon^2)$$

and

$$\|F\|_2 \leq \sum_{k=1}^{n} 2\max(K_{\hat{a}}(\epsilon, \delta, \delta_z, k), K_{\hat{b}}(\epsilon, \delta, \delta_z, k)) + O(\epsilon^2).$$

Similarly

$$\|Q - \tilde{Q}\|_2 \leq \sum_{k=1}^{n} 2\max(K_c(\epsilon, \delta, \delta_z, k), K_s(\epsilon, \delta, \delta_z, k)) + O(\epsilon^2). \tag{8}$$

Writing (7) in terms of H and $\hat{H}$ gives

$$\hat{H} = \tilde{Q}^{\mathrm{H}} \left(H + E - \tilde{Q} F \tilde{Q}^{\mathrm{H}} \right) \tilde{Q}. \tag{9}$$

Thus the computed $\hat{H}$ is similar to a matrix that is close to H.

If we now consider a sequence of QR iterations starting with $H = H_0$ and satisfying the error relation

$$H_{j+1} = \tilde{Q}_j^{\mathrm{H}} \left[H_j + E_j \right] \tilde{Q}_j,$$

then

$$\begin{aligned} H_{j+1} = (\tilde{Q}_j^{\mathrm{H}} \tilde{Q}_{j-1}^{\mathrm{H}} \cdots \tilde{Q}_0^{\mathrm{H}})[H_0 + E_0 + \tilde{Q}_0 E_1 \tilde{Q}_0^{\mathrm{H}} + \tilde{Q}_0 \tilde{Q}_1 E_2 \tilde{Q}_1^{\mathrm{H}} \tilde{Q}_0^{\mathrm{H}} \\ + \cdots (\tilde{Q}_0 \cdots \tilde{Q}_{j-1}) E_j (\tilde{Q}_{j-1}^{\mathrm{H}} \cdots \tilde{Q}_0)] \tilde{Q}_0 \tilde{Q}_1 \cdots \tilde{Q}_j. \end{aligned}$$

Thus H_{j+1} is similar to a matrix that is close to H_0. If the iteration converges so that H_{j+1} is diagonal or numerically diagonal, then the diagonal matrix is similar to a matrix that is close to H_0. The similarity Q can be accurately computed by accumulating the rotations $G_k(c_k, s_k)$ from each iteration. We use these observations to give normwise error bounds in section 7.

We begin the analysis in section 4 with an analysis of the normalization errors and continue in section 5 with a backward error analysis. In both sections, the computation of f_k is of primary importance: The normalization condition $|f_k| = 1$ is crucial throughout the analysis and the backward errors on a_k and b_k are chosen to show that the computation of f_k is stable. The relevant fragment of Algorithm 2 is

the following.

Algorithm 3: Computation of f_k

for $k = 1 : n$

$w_k = z f_{k-1}, \qquad t_k = \overline{a}_k w_k$

if $\mathrm{Re}(t_k) \geq 0$ **or** $|a_k| \leq \frac{\sqrt{2}}{2}$, **then** $g_k = w_k + a_k$

else $g_k = (b_k^2 - 2i\mathrm{Imag}(t_k))/(\overline{w}_k - \overline{a}_k)$

$f_k = \overline{w}_k g_k / \overline{g}_k$

end

To include g_n we have extended the loop to run from $k = 1$ to $k = n$. This computes the unnecessary quantity f_n and uses $b_n = 0$. With $b_n = 0$ the g_n computed by this loop is the same as that computed by Algorithm 2.

In the backward part of the analysis we construct $\tilde{a}_k$ and $\tilde{b}_k$ close to a_k and b_k such that $|\tilde{a}_k|^2 + \tilde{b}_k^2 = 1$ holds exactly. The computed f_k satisfies

$$f_k = (1 + \eta_k)\tilde{f}_k \qquad \text{and} \qquad \eta_k = \frac{f_k - \tilde{f}_k}{\tilde{f}_k} \tag{10}$$

for some relative error $\eta_k = O(\epsilon)$.

In addition to the normalization errors δ_k and δ_z we are also concerned with the normalization of f_k. Since $|\tilde{f}_k| = 1$, the size of $\mathrm{Re}(\eta_k)$ can be viewed as a measure of how far f_k departs from unimodularity. In particular

$$|f_k| = \sqrt{f_k \overline{f}_k} = \sqrt{(1 + \eta_k)(1 + \overline{\eta}_k)} = (1 + \mathrm{Re}(\eta_k)) + O(|\eta_k|^2) \tag{11}$$

so that $\mathrm{Re}(\eta_k) = |f_k| - 1 + O(|\eta_k|^2)$. In section 4 we prove that f_k is numerically unimodular for $k = 1, 2, \ldots, n-1$. This immediately gives a first order bound on $\mathrm{Re}(\eta_k)$ that does not depend on the choice of the backward errors in $\tilde{a}_k$ and $\tilde{b}_k$.

To keep the equations relatively short for as long as possible, we will only substitute bounds on normalization errors into the forward and backward error bounds at the end. Prior to this many of the bounds will be expressed in terms of

$$\delta_f(k) = |\mathrm{Re}(\eta_k)|,$$

$$\delta_f^{(1)}(k) = \sum_{l=1}^{k} \delta_f(l),$$

and

$$\delta_f^{(2)}(k) = \sum_{l=1}^{k} \delta_f^{(1)}(l).$$

The following is an outline of the analysis.

1. The normalization errors: In section 4 we bound $\mathrm{Re}(\eta_k)$ and thus $\delta_f(k)$, $\delta_f^{(1)}(k)$, and $\delta_f^{(2)}(k)$. The bounds are used to show that $g_k \neq 0$ under reasonable assumptions on ϵ and δ_z. We show that the normalization errors δ_k do not increase over the course of repeated UHQR iterations.
2. The backward errors: In section 5 we define $\tilde{a}_k$ and $\tilde{b}_k$. We then bound the backward errors on a_k and b_k and the imaginary part, $\mathrm{Imag}(\eta_k)$, of the forward relative error on f_k. The normalization errors $\mathrm{Re}(\eta_k)$, δ_k, and δ_z are present in these bounds.

3. The forward errors: In section 6 we analyze the error propagation in the computation of c_k, concluding that c_k is close to $\tilde{c}_k$ computed without error from $\tilde{a}_k$ and $\tilde{b}_k$. We also give forward error bounds for all other computed quantities.

Most of the analysis neglects second order terms. One exception is the construction of backward errors for which $|\tilde{a}_k|^2 + \tilde{b}_k^2 = 1$ holds exactly. Another exception is the condition that ensures $g_k \neq 0$. When we wish to keep track of second order terms we use the following result from [5].

LEMMA 1. [illegible] $|\epsilon_k| \leq \epsilon$ $\rho_k = \pm 1$ [illegible] $j\epsilon < 1$ [illegible]

$$\prod_{k=1}^{j} (1+\epsilon_k)^{\rho_k} = 1 + \theta_j,$$

[illegible] $|\theta_j| \leq \gamma_j$ [illegible]

$$\gamma_j = \frac{j\epsilon}{1-j\epsilon}.$$

The error γ_j is used when it is inappropriate to use the first order expansion

$$\prod_{k=1}^{j} (1+\epsilon_k)^{\rho_k} = 1 + \sum_{k=1}^{j} \rho_k \epsilon_j + O(\epsilon^2). \tag{12}$$

Along the same lines, the following is useful for manipulating expressions involving the multiple sources of error ϵ, δ_z, $\mathrm{Re}(\eta_k)$, and δ.

LEMMA 2. [illegible] $\delta_1 \geq 0$ $\delta_2 \geq 0$ [illegible] $\delta_1 + \delta_2 < 1$ [illegible]

$$|\theta_1| \leq \frac{\delta_1}{1-\delta_1} \qquad \text{[illegible]} \qquad |\theta_2| \leq \frac{\delta_2}{1-\delta_2},$$

[illegible]

$$(1+\theta_1)(1+\theta_2) = 1 + \theta$$

[illegible]

$$|\theta| \leq \frac{\delta_1 + \delta_2}{1-(\delta_1+\delta_2)}.$$

[illegible] Clearly

$$|\theta| = |\theta_1 + \theta_2 + \theta_1\theta_2| \leq \frac{\delta_1}{1-\delta_1} + \frac{\delta_2}{1-\delta_2} + \frac{\delta_2}{1-\delta_2}\frac{\delta_1}{1-\delta_1} = \frac{\delta_1+\delta_2-\delta_1\delta_2}{1-\delta_1-\delta_2+\delta_1\delta_2}$$
$$\leq \frac{\delta_1+\delta_2}{1-\delta_1-\delta_2}. \qquad \square$$

The obvious extension of Lemma 2 to the expression $(1+\theta_1)/(1+\theta_2)$ is not true.

We make regular use of the expansions

$$\sqrt{1+\epsilon} = 1 + \frac{\epsilon}{2} + O(\epsilon^2) \qquad \text{and} \qquad \frac{1}{1+\epsilon} = 1 - \epsilon + O(\epsilon^2)$$

without comment.

Throughout the analysis we use ϵ as an effective unit roundoff for complex arithmetic. Thus our model for complex floating point arithmetic is

$$\mathrm{fl}(x \mathrm{\,op\,} y) = (x \mathrm{\,op\,} y)(1+\epsilon_{\mathrm{op}}), \qquad \text{and} \qquad \mathrm{fl}(\sqrt{x}) = \sqrt{x}(1+\epsilon_{\mathrm{op}}),$$

where $|\epsilon_{\mathrm{op}}| \le \epsilon$. If ϵ is replaced with $\sqrt{2}\gamma_4$ where $\gamma_4 = 4\epsilon/(1-4\epsilon)$, then the model holds for the true unit roundoff [5].

We signal the presence of neglected second order terms by $O(\epsilon^2)$. Equations or inequalities in which the $O(\epsilon^2)$ term is absent hold strictly.

We use both subscripts and superscripts on errors. When we write $\epsilon_k^{(j)}$ where $|\epsilon_k^{(j)}| \le \epsilon$ the subscript refers to the index k in the loop of Algorithm 2. The superscript j is present to distinguish errors resulting from different computations within the loop. The superscripts are used consistently throughout the paper; they increase with each error encountered. Similar comments apply to constants $x_k^{(j)}$ for which we require $|x_k^{(j)}| \le 1$.

4. Normalization errors. The normalization error on f_k is the easiest to deal with. Since the shift is by assumption numerically unimodular and $\mathrm{fl}(g_k/\overline{g}_k)$ is numerically unimodular, the formula

$$f_k = \mathrm{fl}\left(\overline{zf_{k-1}}\frac{g_k}{\overline{g}_k}\right)$$

implies that the normalization of f_k is not much worse than that of f_{k-1}. The following result makes this observation precise.

THEOREM 1. [illegible] f_k [illegible] 2 [illegible] $g_j \ne 0$, $j = 1, 2, \ldots, k$ [illegible] γ_3 [illegible] δ_z [illegible] $k\gamma_3 + k|\delta_z| < 1$ [illegible] ϕ_k [illegible] $1 + \phi_k \ne 0$

$$\hat{f}_k = \frac{f_k}{1+\phi_k}, \qquad |\hat{f}_k| = 1, \qquad |\phi_k| \le \frac{k\gamma_3 + k|\delta_z|}{1-(k\gamma_3+k|\delta_z|)} \le k|\delta_z| + 3k\epsilon + O(\epsilon^2).$$

[illegible] γ_j [illegible] 1 [illegible] η_k [illegible] $f_k = (1+\eta_k)\tilde{f}_k$ [illegible] $|\tilde{f}_k| = 1$ [illegible]

$$\phi_k = \mathrm{Re}(\eta_k) + O(\epsilon^2)$$

[illegible]

$$|\mathrm{Re}(\eta_k)| \le k|\delta_z| + 3k\epsilon + O(\epsilon^2).$$

[illegible]. The proof is inductive. Since $f_0 = 1$ we may choose $\phi_0 = 0$. We assume that there is a real ϕ_{k-1} satisfying the upper bound stated in the theorem and such that $\hat{f}_{k-1} = 1$ and $1 + \phi_{k-1} \ne 0$. The numerical errors in the computation of f_k are

$$f_k = \mathrm{fl}\left(\frac{\overline{zf_{k-1}}g_k}{\overline{g}_k}\right) = \frac{\overline{zf_{k-1}}g_k}{\overline{g}_k}(1+3\epsilon_k^{(1)}) = \frac{\overline{\tilde{z}\hat{f}_{k-1}}g_k}{\overline{g}_k}(1+3\epsilon_k^{(1)})(1+\delta_z)(1+\phi_{k-1})$$

for $|\epsilon_k^{(1)}| \le \gamma_3/3$. We choose ϕ_k so that

$$1 + \phi_k = \left|1 + 3\epsilon_k^{(1)}\right|(1+\delta_z)(1+\phi_{k-1}).$$

Clearly the assumptions on γ_3 and δ_z imply that $1+\phi_k \neq 0$ so that

$$|\hat{f}_k| = \left|\frac{f_k}{1+\phi_k}\right| = \left|\frac{\overline{\tilde{z}\hat{f}_{k-1}}g_k}{(1+\phi_k)\overline{g}_k}(1+3\epsilon_k^{(1)})(1+\delta_z)(1+\phi_{k-1})\right| = 1.$$

Applying Lemma 2 we conclude that when $\gamma_3 + |\delta_z| < 1$ we have

$$\begin{aligned} 1+|\phi_k| &\leq \left(1+\frac{\gamma_3+|\delta_z|}{1-(\gamma_3+|\delta_z|)}\right)(1+|\phi_{k-1}|) \\ &\leq \left(1+\frac{\gamma_3+|\delta_z|}{1-(\gamma_3+|\delta_z|)}\right)\left(1+\frac{(k-1)\gamma_3+(k-1)|\delta_z|}{1-((k-1)\gamma_3+(k-1)|\delta_z|)}\right). \end{aligned}$$

A further application of Lemma 2 gives the upper bound for $|\phi_k|$. The first order upper bound is a first order expansion of the strict upper bound.

The claim that $\phi_k = \mathrm{Re}(\eta_k) + O(\epsilon^2)$ follows from

$$\left|(1+\phi_k)\hat{f}_k\right|^2 = |f_k|^2 = \left|(1+\eta_k)\tilde{f}_k\right|^2$$

which implies $|1+\phi_k|^2 = |1+\eta_k|^2$ or

$$1 + 2\mathrm{Re}(\phi_k) + |\phi_k|^2 = 1 + 2\mathrm{Re}(\eta_k) + |\eta_k|^2.$$

Since ϕ_k is real this implies $\phi_k = \mathrm{Re}(\eta_k) + O(\epsilon^2)$. □

The theorem shows that

$$\delta_f(k) \leq k|\delta_z| + 3k\epsilon + O(\epsilon^2), \tag{13}$$

$$\delta_f^{(1)}(k) \leq \sum_{l=1}^{k} l|\delta_z| + 3l\epsilon = \frac{3}{2}(k^2+k)\epsilon + \frac{1}{2}(k^2+k)|\delta_z| + O(\epsilon^2), \tag{14}$$

and

$$\delta_f^{(2)}(k) \leq \sum_{l=1}^{k} \frac{3}{2}(l^2+l)\epsilon + \frac{1}{2}(l^2+l)|\delta_z| = \frac{1}{2}(k^3+3k^2+2k)\epsilon + \frac{1}{6}(k^3+3k^2+2k)|\delta_z| + O(\epsilon^2). \tag{15}$$

We can now use the bounds on the normalization on f_k to give conditions under which $g_k \neq 0$; we assume that underflow does not occur.

THEOREM 2. [illegible] $1 \leq k \leq n-1$

- [illegible] $|a_k| \leq \sqrt{2}/2$ [illegible] $g_k =$ [illegible] $(w_k + a_k)$ [illegible]

$$\frac{2-\sqrt{2}/2}{1-\sqrt{2}/2}(\epsilon + (k-1)\gamma_3 + k|\delta_z|) < 1. \tag{16}$$

- [illegible] $\mathrm{Re}(t_k) \geq 0$ [illegible] $g_k =$ [illegible] $(w_k + a_k)$ [illegible]

$$4(k-1)\gamma_3 + 4k|\delta_z| + 2\gamma_2 + 2\epsilon < 1. \tag{17}$$

- [illegible] $\mathrm{Re}(t_k) < 0$ [illegible] $|a_k| > \sqrt{2}/2$ [illegible] $g_k =$ [illegible] $(b_k^2 - 2i\mathrm{Imag}(t_k))/(\overline{w}_k - \overline{a}_k)$ [illegible] (17) [illegible]

$(\overline{w}_k - \overline{a}_k) \neq 0$ $g_k \neq 0$ 1 $|\phi_k| < 1$ $|\hat{f}_k| = 1$ $\hat{f}_k = f_k/(1+\phi_k)$

We consider the cases in which $g_k = w_k + a_k$ first. The computation of g_k with errors is

$$g_k = \left(z f_{k-1}(1+\epsilon_k^{(2)}) + a_k \right)(1+\epsilon_k^{(3)}),$$

where $|\epsilon_k^{(2)}| < \epsilon$ and $|\epsilon_k^{(3)}| < \epsilon$. Either (16) or (17) guarantees that $|\epsilon_k^{(3)}| < 1$ so that g_k is zero if and only if

$$z f_{k-1}(1+\epsilon_k^{(2)}) + a_k = 0$$

or equivalently

$$\hat{z}\hat{f}_{k-1}(1+\epsilon_k^{(2)})(1+\phi_{k-1})(1+\delta_z) + a_k = 0. \tag{18}$$

Assuming that $|a_k| \leq \sqrt{2}/2$ and that (16) holds we clearly have $\epsilon+(k-1)\gamma_3+k|\delta_z| < 1$. Lemma 2 and Theorem 1 then imply that

$$\hat{z}\hat{f}_{k-1}(1+\theta) + a_k = 0,$$

where

$$|\theta| \leq \frac{\epsilon + (k-1)\gamma_3 + k|\delta_z|}{1 - \epsilon - (k-1)\gamma_3 - k|\delta_z|}.$$

Since $|a_k| \leq \sqrt{2}/2$ and $|\hat{z}\hat{f}_{k-1}| = 1$, $g_k = 0$ implies

$$|\theta| \geq 1 - \sqrt{2}/2.$$

However, (16) can be put in the form

$$\frac{\epsilon + (k-1)\gamma_3 + k|\delta_z|}{1 - (\epsilon + (k-1)\gamma_3 + k|\delta_z|)} < 1 - \frac{\sqrt{2}}{2}$$

so that $|\theta| < 1 - \sqrt{2}/2$ and $g_k \neq 0$.

Considering the case $\mathrm{Re}(t_k) \geq 0$ we assume that (17) holds and return to (18) which holds if and only if

$$(1+\epsilon_k^{(2)})(1+\phi_{k-1})(1+\delta_z) + \frac{a_k \overline{z}\,\overline{f}_{k-1}}{(1+\phi_{k-1})(1+\delta_z)} = 0.$$

The denominator of the second term is nonzero since (17) implies that $|\delta_z| < 1$ and that

$$\frac{2(k-1)\gamma_3 + 2k|\delta_z| + \gamma_2 + \epsilon}{1 - (2(k-1)\gamma_3 + 2k|\delta_z| + \gamma_2 + \epsilon)} < 1$$

which by Theorem 1 implies $|\phi_{k-1}| < 1$. The errors in the computation of t_k are

$$t_k = \overline{a}_k z f_{k-1}(1+2\epsilon_k^{(4)}),$$

where $|\epsilon_k^{(4)}| \leq \gamma_2/2$. Thus g_k is zero if and only if

$$(1+\epsilon_k^{(2)})(1+\phi_{k-1})(1+\delta_z) + \frac{\bar{t}_k}{(1+\phi_{k-1})(1+\delta_z)(1+2\bar{\epsilon}_k^{(4)})} = 0.$$

Since $\gamma_2 < 1$ the denominator is nonzero. It follows that $g_k = 0$ if and only if

$$(1+\epsilon_k^{(2)})(1+\phi_{k-1})^2(1+\delta_z)^2(1+2\bar{\epsilon}_k^{(4)}) + \bar{t}_k = 0.$$

By Theorem 1 and repeated application of Lemma 2 this is equivalent to

$$1 + \theta + \bar{t}_k = 0$$

for some θ satisfying

$$|\theta| < \frac{2(k-1)\gamma_3 + 2k|\delta_z| + \gamma_2 + \epsilon}{1 - (2(k-1)\gamma_3 + 2k|\delta_z| + \gamma_2 + \epsilon)}.$$

Since $\mathrm{Re}(t_k) \geq 0$ this can only happen if $|\mathrm{Re}(\theta)| > 1$. The inequality (17) implies that $|\theta| < 1$ so that $g_k \neq 0$.

Finally, we observe that the proof that $\overline{w}_k - \overline{a}_k \neq 0$ when (17) holds and $\mathrm{Re}(t_k) < 0$ parallels the proof that $w_k + a_k \neq 0$ when (17) holds and $\mathrm{Re}(t_k) \geq 0$. Since H is by assumption unreduced, $b_k \neq 0$ so that

$$\mathrm{Re}(\mathrm{fl}(b_k^2 - 2i\mathrm{Imag}(t_k))) = b_k^2(1+\epsilon_k^{(5)}) \neq 0$$

for some $|\epsilon_k^{(5)}| < \epsilon$. The computation of g_k is then

$$g_k = \frac{\mathrm{fl}(b_k^2 - 2i\mathrm{Imag}(t_k))}{\mathrm{fl}(\overline{w}_k - \overline{a}_k)}(1+\epsilon_k^{(6)}) \neq 0. \qquad \square$$

We have neglected the possibility of underflow in proving Theorem 2. However, underflow appears unlikely if H is numerically unreduced. If $|a_k| \leq \sqrt{2}/2$ or $\mathrm{Re}(t_k) \geq 0$ so that the formula $g_k = w_k + a_k$ is used, then $|g_k| \geq 1 - \sqrt{2}/2 + O(\epsilon)$. If $|a_k| > \sqrt{2}/2$ and $\mathrm{Re}(t_k) < 0$ so that the alternate formula for g_k is used, then $|\overline{w_k} - \overline{a_k}| < 2 + O(\epsilon)$ so that the numerator must underflow or very nearly underflow to cause an underflow in the computation of g_k. This occurs only if the computation of b_k^2 underflows.

The following result shows that for any given $k = 1, 2, \ldots, n-1$ the normalization error for $\hat{a}_k$ and $\hat{b}_k$ computed by Algorithm 2 is never much larger than δ_{k+1} provided that the formula $g_k = w_k + a_k$ is used.

THEOREM 3. [illegible] $g_l \neq 0$, $l = 1, 2, \ldots k+1$ [illegible] g_{k+1} [illegible] $g_{k+1} = w_{k+1} + a_{k+1}$ [illegible] $|a_k|^2 + b_k^2 = 1 + \delta_k$, $k = 1, 2, \ldots, n$ [illegible] $b_n = 0$ [illegible] $k = 1, 2, \ldots n-1$ [illegible] $\hat{a}_k$ [illegible] $\hat{b}_k$ [illegible]

$$|\hat{a}_k|^2 + \hat{b}_k^2 = 1 + \hat{\delta}_k$$

[illegible]

$$\hat{\delta}_k = s_k^2\delta_{k+1} + x_k\left[(66k+70)\epsilon + (22k+18)|\delta_z|\right] + O(\epsilon^2)$$

[illegible] $|x_k| \leq 1$

. We first consider the normalization errors in the relation $|c_k|^2 + s_k^2 = 1$. The errors involved in computing c_k and s_k are

$$r_k = \sqrt{|p_k|^2 + b_k^2}\left(1 + 4\epsilon_k^{(7)}\right) + O(\epsilon^2),$$

$$s_k = \frac{b_k}{r_k}\left(1 + \epsilon_k^{(8)}\right), \qquad \text{and} \qquad c_k = \frac{p_k}{r_k}\left(1 + \epsilon_k^{(9)}\right).$$

It follows that

$$\begin{aligned} |c_k|^2 + s_k^2 &= \frac{b_k^2 + |p_k|^2}{r_k^2} + 2\mathrm{Re}(\epsilon_k^{(8)})\frac{b_k^2}{r_k^2} + 2\mathrm{Re}(\epsilon_k^{(9)})\frac{|p_k|^2}{r_k^2} + O(\epsilon^2) \\ &= \frac{r_k^2(1 - 8\epsilon_k^{(7)})}{r_k^2} + 2\mathrm{Re}(\epsilon_k^{(8)})\frac{b_k^2}{r_k^2} + 2\mathrm{Re}(\epsilon_k^{(9)})\frac{|p_k|^2}{r_k^2} + O(\epsilon^2) \end{aligned}$$

so that

$$s_k^2 + |c_k|^2 = 1 + 10\epsilon_k^{(10)} + O(\epsilon^2).$$

Let ϕ_k be chosen as in Theorem 1 so that $\hat{f}_k(1 + \phi_k) = f_k$, where $|\hat{f}_k| = 1$. The errors in the computation of g_{k+1} are

$$\begin{aligned} g_{k+1} &= \mathrm{fl}(zf_k + a_{k+1}) \\ &= (zf_k + a_{k+1})\left(1 + 5\epsilon_k^{(11)}\right) + O(\epsilon^2) \\ &= (\tilde{z}\hat{f}_k + a_{k+1})\left(1 + 5\epsilon_k^{(11)} + 4x_k^{(1)}(|\delta_z| + |\phi_k|)\right) + O(\epsilon^2). \\ &= \tilde{z}\hat{f}_k\left(1 + a_{k+1}\overline{\tilde{z}\hat{f}_k}\right)\left(1 + 5\epsilon_k^{(11)} + 4x_k^{(1)}(|\delta_z| + |\phi_k|)\right) + O(\epsilon^2), \end{aligned}$$

where we have used the fact that $|g_{k+1}| \geq 1/4 + O(\epsilon)$ to cast the errors associated with the multiplication zf_k as relative errors on g_{k+1}.

We now consider the computation

$$\hat{b}_k = \mathrm{fl}(r_{k+1}s_k) = \sqrt{|p_{k+1}|^2 + b_{k+1}^2} \cdot s_k\left(1 + 5\epsilon_k^{(12)}\right) + O(\epsilon^2)$$

and

$$\begin{aligned} p_{k+1} &= \mathrm{fl}(g_{k+1}\overline{f}_k c_k) \\ &= g_{k+1}\overline{f}_k c_k(1 + 2\epsilon_k^{(13)}) + O(\epsilon^2) \\ &= g_{k+1}\overline{\hat{f}}_k c_k(1 + 2\epsilon_k^{(13)} + \phi_k) + O(\epsilon^2). \end{aligned}$$

Combining the formula for $\hat{b}_k$ with those for p_{k+1} and g_{k+1} gives

$$\hat{b}_k = \sqrt{|1 + a_{k+1}\overline{\tilde{z}\hat{f}_k}|^2|c_k|^2 + b_{k+1}^2} \cdot s_k\left(1 + x_k^{(2)}\left(19\epsilon + 10|\phi_k| + 8|\delta_z|\right)\right) + O(\epsilon^2).$$

We have used the lack of cancellation in the equation to combine the errors on p_{k+1} and g_{k+1} into a relative error on $\hat{b}_k$. Note that the above expression for $\hat{b}_k$ applies for $k = n - 1$ under the assumption that $b_n = 0$ even though $\hat{b}_{n-1}$ is computed outside the loop of Algorithm 2.

The errors in the computation of $\hat{a}_k$ are

$$\begin{aligned}\hat{a}_k &= \mathrm{fl}(|c_k|^2 f_k - \overline{z} s_k^2 a_{k+1}) \\ &= |c_k|^2 f_k - \overline{z} s_k^2 a_{k+1} + 6\epsilon_k^{(14)} + O(\epsilon^2) \\ &= |c_k|^2 \hat{f}_k - \overline{\tilde{z}} s_k^2 a_{k+1} + x_k^{(3)} \left(6\epsilon + |\delta_z| + |\phi_k|\right) + O(\epsilon^2).\end{aligned}$$

Finally, combining the expressions for $\hat{a}_k$ and $\hat{b}_k$ gives

$$\begin{aligned}|\hat{a}_k|^2 + \hat{b}_k^2 &= |c_k|^4 + |a_{k+1}|^2 s_k^4 - 2\mathrm{Re}\left(|c_k|^2 s_k^2 \hat{f}_k \tilde{z} \overline{a}_{k+1}\right) \\ &\quad + \left(|c_k|^2 \left|1 + a_{k+1}\overline{\tilde{z}\hat{f}_k}\right|^2 + b_{k+1}^2\right) s_k^2 + x_k^{(4)} \left(50\epsilon + 18|\delta_z| + 22|\phi_k|\right) + O(\epsilon^2) \\ &= |c_k|^4 + |a_{k+1}|^2 s_k^4 - 2|c_k|^2 s_k^2 \mathrm{Re}\left(\hat{f}_k \tilde{z} \overline{a}_{k+1}\right) \\ &\quad + \left(|c_k|^2 \left(1 + 2\mathrm{Re}\left(\tilde{z}\hat{f}_k \overline{a}_{k+1}\right) + |a_{k+1}|^2\right) + b_{k+1}^2\right) s_k^2 \\ &\quad + x_k^{(4)} \left(50\epsilon + 18|\delta_z| + 22|\phi_k|\right) + O(\epsilon^2) \\ &= |c_k|^2 \left(|c_k|^2 + s_k^2\right) + |a_{k+1}|^2 s_k^2 \left(|c_k|^2 + s_k^2\right) + b_{k+1}^2 s_k^2 \\ &\quad + x_k^{(4)} \left(50\epsilon + 18|\delta_z| + 22|\phi_k|\right) + O(\epsilon^2) \\ &= |c_k|^2 + s_k^2 \left(|a_{k+1}|^2 + b_{k+1}^2\right) + \left(|c_k|^2 + |a_{k+1}|^2 s_k^2\right) 10\epsilon_k^{(10)} \\ &\quad + x_k^{(4)} \left(50\epsilon + 18|\delta_z| + 22|\phi_k|\right) + O(\epsilon^2) \\ &= 1 + s_k^2 \delta_{k+1} + 10\left(1 + |c_k|^2 + |a_{k+1}|^2 s_k^2\right) \epsilon_k^{(10)} \\ &\quad + x_k^{(5)} \left(50\epsilon + 18|\delta_z| + 22|\phi_k|\right) + O(\epsilon^2) \\ &= 1 + s_k^2 \delta_{k+1} + x_k^{(6)} \left(70\epsilon + 18|\delta_z| + 22|\phi_k|\right) + O(\epsilon^2).\end{aligned}$$

The lemma follows by applying the first order bound on $|\phi_k|$ from Theorem 1. □

The main result of Theorem 3 is of the form

$$\hat{\delta}_k = s_k^2 \delta_{k+1} + O(\epsilon),$$

where the $O(\epsilon)$ term hides only bounded errors. It is significant that this is an equality and not an upper bound; the theorem accurately describes the effect of δ_{k+1} on $\hat{\delta}_k$. Since $|s_k^2| \leq 1$ the propagation of normalization errors is stable when the formula $g_{k+1} = w_{k+1} + a_{k+1}$ is used.

To show that error propagation is stable in general, we must give an analogous result that applies when the algorithm uses the other formula for g_{k+1}.

THEOREM 4. [illegible] $g_l \neq 0$ [illegible] $l = 1, 2, \ldots, k+1$ [illegible] g_{k+1} [illegible] (4) [illegible] $|a_k|^2 + b_k^2 = 1 + \delta_k$ [illegible] $k = 1, 2, \ldots, n$ [illegible] $b_n = 0$ [illegible] $k = 1, 2, \ldots, n-1$ [illegible] $\hat{a}_k$ [illegible] $\hat{b}_k$ [illegible]

$$|\hat{a}_k|^2 + \hat{b}_k^2 = 1 + \hat{\delta}_k$$

[illegible]

$$\hat{\delta}_k = \left(s_k^2 + \frac{2b_{k+1}^2 |c_k|^2 s_k^2}{|z f_k - a_{k+1}|^2}\right) \delta_{k+1} + x_k \left[(78k + 118)\epsilon + (26k + 18)|\delta_z|\right] + O(\epsilon^2)$$

$|x_k| < 1$

. Let $\hat{f}_k(1+\phi_k) = f_k$, where $|\hat{f}_k| = 1$. As in the proof of Theorem 3 we have

$$|c_k|^2 + s_k^2 = 1 + 10\epsilon_k^{(10)} + O(\epsilon^2).$$

The assumption that (4) is used to compute g_{k+1} implies that $\mathrm{Re}(t_{k+1}) < 0$ which in turn implies that $|f_k z - a_{k+1}| > 1 + O(\epsilon)$. We also have $g_{k+1} \le 2 + O(\epsilon)$. Using these facts we can represent errors in (4) as small absolute errors so that g_{k+1} satisfies

$$\begin{aligned} g_{k+1} &= \frac{b_{k+1}^2 - 2i\,\mathrm{Imag}(\overline{a}_{k+1} f_k z)}{\overline{f_k z} - \overline{a}_{k+1}} + 15\epsilon_k^{(15)} + O(\epsilon^2) \\ &= \frac{b_{k+1}^2 - 2i\,\mathrm{Imag}(\overline{a}_{k+1} \hat{f}_k \tilde{z})}{\overline{\hat{f}_k \tilde{z}} - \overline{a}_{k+1}} + x_k^{(7)}\,(15\epsilon + 4|\phi_k| + 4|\delta_z|) + O(\epsilon^2). \end{aligned} \tag{19}$$

We also have

$$\begin{aligned} \tilde{z}\hat{f}_k + a_{k+1} &= \frac{(\tilde{z}\hat{f}_k + a_{k+1})\overline{(\tilde{z}\hat{f}_k - a_{k+1})}}{\overline{(\tilde{z}\hat{f}_k - a_{k+1})}} \\ &= \frac{1 - |a_{k+1}|^2 - 2i\,\mathrm{Imag}(\overline{a}_{k+1}\hat{f}_k\tilde{z})}{\overline{(\tilde{z}\hat{f}_k - a_{k+1})}} \end{aligned}$$

so that

$$\begin{aligned} \left|1 + a_{k+1}\overline{\tilde{z}\hat{f}_k}\right|^2 &= |\tilde{z}\hat{f}_k + a_{k+1}|^2 \\ &= \frac{\left(b_{k+1}^2 - \delta_{k+1}\right)^2 + 4\,\mathrm{Imag}(\overline{a}_{k+1}\hat{f}_k\tilde{z})^2}{|\tilde{z}\hat{f}_k - a_{k+1}|^2} \\ &= |g_{k+1}|^2 - \frac{2b_{k+1}^2}{|\tilde{z}\hat{f}_k - a_{k+1}|^2}\delta_{k+1} + x_k^{(8)}\,(60\epsilon + 16|\phi_k| + 16|\delta_z|) + O(\epsilon^2), \end{aligned}$$

where we have used the fact that $|g_{k+1}| \le 2 + O(\epsilon)$ when using (19) to evaluate $|g_{k+1}|^2$.

As in the proof of Theorem 3 we have

$$\hat{b}_k = \sqrt{|p_{k+1}|^2 + b_{k+1}^2} \cdot s_k \left(1 + 5\epsilon_k^{(12)}\right) + O(\epsilon^2)$$

with

$$p_{k+1} = g_{k+1}\overline{\hat{f}_k} c_k \left(1 + 2\epsilon_k^{(13)} + \phi_k\right).$$

Thus

$$\begin{aligned} \hat{b}_k^2 &= \left(|g_{k+1}|^2|c_k|^2\left(1 + 4\mathrm{Re}(\epsilon_k^{(13)}) + 2\phi_k\right) + b_{k+1}^2\right) s_k^2(1 + 10\epsilon_k^{(12)}) + O(\epsilon^2) \\ &= \left(|g_{k+1}|^2|c_k|^2 + b_{k+1}^2\right) s_k^2 + x_k^{(9)}\,(26\epsilon + 8|\phi_k|) + O(\epsilon^2) \\ &= \left(\left|1 + a_{k+1}\tilde{z}\hat{f}_k\right|^2 |c_k|^2 + b_{k+1}^2\right) s_k^2 + \frac{2|c_k|^2 s_k^2 b_{k+1}^2}{|\tilde{z}\hat{f}_k - a_{k+1}|^2}\delta_{k+1} \\ &\quad + x_k^{(10)}\,(86\epsilon + 24|\phi_k| + 16|\delta_z|) + O(\epsilon^2). \end{aligned}$$

As before the computed $\hat{a}_k$ is

$$\hat{a}_k = |c_k|^2 \hat{f}_k - \overline{\tilde{z}} s_k^2 a_{k+1} + x_k^{(3)} \left(6\epsilon + |\delta_z| + |\phi_k|\right) + O(\epsilon^2)$$

so that

$$\begin{aligned}|\hat{a}_k|^2 + \hat{b}_k^2 = |c_k|^4 + |a_{k+1}|^2 s_k^4 - 2\mathrm{Re}\left(|c_k|^2 s_k^2 \hat{f}_k \tilde{z}\overline{a}_{k+1}\right) \\ + \left(|c_k|^2 \left|1 + a_{k+1}\overline{\tilde{z}\hat{f}_k}\right|^2 + b_{k+1}^2\right) s_k^2 + x_k^{(11)} \left(98\epsilon + 18|\delta_z| + 26|\phi_k|\right) \\ + \frac{2b_{k+1}^2 |c_k|^2 s_k^2}{|\tilde{z}\hat{f}_k - a_{k+1}|^2} \delta_{k+1} + O(\epsilon^2).\end{aligned}$$

A repetition of the final part of the proof of Theorem 3 gives

$$\begin{aligned}|\hat{a}_k|^2 + \hat{b}_k^2 &= |c_k|^2 + s_k^2(|a_{k+1}|^2 + b_{k+1}^2) + 10(|c_k|^2 + |a_{k+1}|^2 s_k^2)\epsilon_k^{(10)} \\ &\quad + x_k^{(11)} \left(98\epsilon + 26|\phi_k| + 18|\delta_z|\right) + \frac{2b_{k+1}^2 |c_k|^2 s_k^2}{|\tilde{z}\hat{f}_k - a_{k+1}|^2} \delta_{k+1} + O(\epsilon^2) \\ &= 1 + \left(s_k^2 + \frac{2b_{k+1}^2 |c_k|^2 s_k^2}{|\tilde{z}\hat{f}_k - a_{k+1}|^2}\right) \delta_{k+1} + x_k^{(12)} \left(118\epsilon + 26|\phi_k| + 18|\delta_z|\right) + O(\epsilon^2).\end{aligned}$$

The lemma follows upon using Theorem 1 to bound $|\phi_k|$. ∎

The reason for requiring $|a_k| > \sqrt{2}/2$ when applying the alternate formula for g_k can now be made clear. If (4) is used to compute g_{k+1}, then $\mathrm{Re}(t_{k+1}) < 0$ so that $|zf_k - a_{k+1}| > 1 + O(\epsilon)$. Since $|a_{k+1}|^2 > 1/2$ we have $b_{k+1}^2 \leq 1/2 + O(\epsilon)$. Neglecting numerical errors, the multiplier of δ_{k+1} is

$$s_k^2 + \frac{2b_{k+1}^2 |c_k|^2 s_k^2}{|zf_k - a_{k+1}|^2} \leq s_k^2 + 2b_{k+1}^2 |c_k|^2 s_k^2 \leq s_k^2 + |c_k|^2 s_k^2 = s_k^2(2 - s_k^2) \leq 1.$$

The function $s_k^2(2 - s_k^2)$ reaches its maximum at $s_k^2 = 1$. With errors we have

$$s_k^2 + \frac{2b_{k+1}^2 |c_k|^2 s_k^2}{|zf_k - a_{k+1}|^2} \leq 1 + O(\epsilon).$$

The $O(\epsilon)$ term is second order when multiplied by δ_{k+1}. Thus when the alternate formula for g_k is used we have

$$|\hat{\delta}_k| \leq |\delta_{k+1}| + O(\epsilon),$$

where the $O(\epsilon)$ term hides only bounded errors.

Whichever formula for g_k is used we can conclude by taking the largest error coefficients from each equation for $\hat{\delta}_k$ that

$$|\hat{\delta}_k| \leq |\delta_{k+1}| + (78k + 118)\epsilon + (26k + 18)|\delta_z| + O(\epsilon^2).$$

If the initial errors are all less than δ, then after j UHQR iterations the normalization errors satisfy

$$|\delta_k| \leq \delta + j(78k + 118)\epsilon + j(26k + 18)|\delta_z| + O(\epsilon^2) \tag{20}$$

so that the errors grow at worst linearly in the number of UHQR iterations.

If we apply the alternate formula for g_k when $|a_k| \le \sqrt{2}/2$, then the most that we can conclude is that

$$\frac{s_k^2 + 2b_{k+1}^2|c_k|^2 s_k^2}{zf_k - a_{k+1}} \le s_k^2(1 + 2|c_k|^2) = s_k^2(3 - 2s_k^2) \le \frac{9}{8},$$

where the maximum occurs when $b_{k+1} = 1$ and $s_k = \sqrt{3}/2$. This upper bound is achieved by the 5×5 unitary Hessenberg matrix with parameters

$$a_0 = 1, \qquad (a_1, b_1) = \left(-\frac{1}{2}i, \frac{\sqrt{3}}{2}\right), \qquad (a_2, b_2) = (0, 1), \qquad (a_3, b_3) = \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right),$$

$$(a_4, b_4) = (0, 1), \qquad a_5 = 1.$$

The projected Wilkinson shift is $z = i$. Application of the UHQR algorithm gives $g_1 = i/2$, $p_1 = i/2$, $r_1 = 1$, $c_1 = i/2$, and $s_1 = \sqrt{3}/2$. Thus

$$\hat{\delta}_1 = \left(s_1^2 + \frac{2b_2^2|c_1|^2 s_1^2}{zf_1 - a_2}\right)\delta_2 + O(\epsilon) = \frac{9}{8}\delta_2 + O(\epsilon).$$

This suggests that if we do not require $|a_k| > \sqrt{2}/2$ when using the alternate formula for g_k, then the normalization errors could grow as $O((9/8)^j)$ over j UHQR iterations. The possibility of exponential growth in errors is inconvenient when attempting to prove the stability of the algorithm, but it seems to have no practical impact on the accuracy of the computation. Sustained exponential error growth has not been observed.

5. Backward errors. The backward errors on a_k and b_k have the form

$$\tilde{a}_k = (1 - \alpha_k + i\mu_k)a_k$$

and

$$\tilde{b}_k = (1 - \beta_k)b_k,$$

where α_k, β_k, and μ_k are real. The α_k and β_k are chosen to enforce the normalization $|\tilde{a}_k|^2 + \tilde{b}_k^2 = 1$. The imaginary part μ_k is chosen to ensure that the computed f_k is close to $\tilde{f}_k$. That is, the choice of μ_k guarantees that η_k in (10) is small. In defining μ_k we assume that the relative error η_{k-1} on f_{k-1} has already been determined. This leads to a recurrence for η_k in terms of η_{k-1}. Since $f_0 = \tilde{f}_0 = 1$ the process starts with $\eta_0 = 0$. The recurrence can be used to bound η_k for $k = 1, 2, \ldots, n-1$.

For the computed t_k define the error $|\epsilon_k^{(4)}| \le \gamma_2/2$ by

$$t_k = \mathrm{fl}(\overline{a}_k z f_{k-1}) = (1 + 2\epsilon_k^{(4)})\overline{a}_k z f_{k-1}. \tag{21}$$

Let

$$\mu_k = -\mathrm{Imag}(\eta_{k-1}) - 2\mathrm{Imag}(\epsilon_k^{(4)}).$$

In defining α_k and β_k we consider two cases.

1. If $|a_k| > \sqrt{2}/2$ we set $\beta_k = 0$ and

$$\alpha_k = 1 - \sqrt{1 - \frac{\delta_k}{|a_k|^2} - \mu_k^2} = \frac{\delta_k}{2|a_k|^2} + O(\epsilon^2). \tag{22}$$

In this case we have

$$(1-\alpha_k)^2 + \mu_k^2 = 1 - \frac{\delta_k}{|a_k|^2}$$

so that

$$|\tilde{a}_k|^2 + \tilde{b}_k^2 = |a_k|^2\left((1-\alpha_k)^2 + \mu_k^2\right) + b_k^2 = |a_k|^2\left(1 - \frac{\delta_k}{|a_k|^2}\right) + b_k^2 = 1.$$

2. If $|a_k| \leq \sqrt{2}/2$ we set $\alpha_k = 0$ and

$$\beta_k = 1 - \sqrt{1 - \frac{|a_k|^2}{b_k^2}\mu_k^2 - \frac{\delta_k}{b_k^2}} = \frac{\delta_k}{2b_k^2} + O(\epsilon^2).$$

In this case we have

$$(1-\beta_k)^2 = 1 - \frac{\delta_k}{b_k^2} - \frac{|a_k|^2}{b_k^2}\mu_k^2$$

so that

$$\begin{aligned} |\tilde{a}_k|^2 + \tilde{b}_k^2 &= |a_k|^2(1+\mu_k^2) + b_k^2|(1-\beta_k)^2 \\ &= |a_k|^2(1+\mu_k^2) + b_k^2\left(1 - \frac{\delta_k}{b_k^2} - \frac{|a_k|^2}{b_k^2}\mu_k^2\right) = 1. \end{aligned}$$

Note that for a_n the first case applies with $\tilde{b}_n = b_n = 0$.

Whichever case holds we have

$$|\tilde{a}_k|^2 + |\tilde{b}_k|^2 = 1 \qquad \text{and} \qquad \max(|\alpha_k|, |\beta_k|) \leq |\delta_k| + O(\epsilon^2). \tag{23}$$

We now perform a forward error analysis for each of the two equations for g_k. In each case we show that the choice of μ_k leads to a recurrence for η_k in terms of η_{k-1}.

Analysis for the first formula. We start by assuming that $\mathrm{Re}(t_k) \geq 0$ or $|a_k| \leq \sqrt{2}/2$ so that

$$g_k = \mathrm{fl}(w_k + a_k).$$

The inequalities that hold when this formula is used imply that there is no significant cancellation in computing g_k. Hence we may represent the effect of all errors on a_k and w_k as small forward relative errors on g_k.

To make this precise we give a lower bound on $|g_k|$. First assume that $|a_k| \leq \sqrt{2}/2$. Since $|w_k| = 1 + O(\epsilon)$ the formula $g_k = w_k + a_k$ gives

$$|g_k| \geq 1 - \frac{\sqrt{2}}{2} + O(\epsilon).$$

If we instead assume that $\mathrm{Re}(t_k) \geq 0$, then $g_k = w_k(1 + \bar{t}_k) + O(\epsilon)$ implies

$$|g_k| \geq 1 + O(\epsilon).$$

Thus whenever the formula $g_k = w_k + a_k$ is used

$$|g_k| \geq 1 - \frac{\sqrt{2}}{2} + O(\epsilon) \geq \frac{1}{4} + O(\epsilon) \geq \frac{1}{4}|a_k| + O(\epsilon). \tag{24}$$

We have used $1/4$ instead of $1 - \sqrt{2}/2 \approx .29$ in order to avoid unwieldy constants in already long error bounds.

Let the errors in computing g_k from z, f_{k-1}, and a_k be

$$g_k = \mathrm{fl}(a_k + w_k) = \left[a_k + z f_{k-1}(1 + \epsilon_k^{(16)})\right](1 + \epsilon_k^{(17)}),$$

where $|\epsilon_k^{(16)}| \leq \epsilon$ and $|\epsilon_k^{(17)}| \leq \epsilon$. Since $a_k = (1 + \alpha_k - i\mu_k)\tilde{a}_k + O(\epsilon^2)$, $z = (1 + \delta_z)\tilde{z}$, and $f_{k-1} = (1+\eta_{k-1})\tilde{f}_{k-1}$ and since we have chosen $\mu_k = -\mathrm{Imag}(\eta_{k-1}) - 2\mathrm{Imag}(\epsilon_k^{(4)})$ we have

$$\begin{aligned}
\mathrm{fl}(g_k) &= \left[\tilde{a}_k\,(1 + \alpha_k - i\mu_k) + \tilde{z}\tilde{f}_{k-1}\left(1 + \eta_{k-1} + \delta_z + \epsilon_k^{(16)}\right)\right](1 + \epsilon_k^{(17)}) + O(\epsilon^2) \\
&= \left[\tilde{a}_k\left(1 + \alpha_k + i\mathrm{Imag}(\eta_{k-1}) + 2i\mathrm{Imag}(\epsilon_k^{(4)})\right)\right. \\
&\quad \left. + \tilde{z}\tilde{f}_k\left(1 + \mathrm{Re}(\eta_{k-1}) + i\mathrm{Imag}(\eta_{k-1}) + \delta_z + \epsilon_k^{(16)}\right)\right](1 + \epsilon_k^{(17)}) + O(\epsilon^2).
\end{aligned}$$

Note that $i\mathrm{Imag}(\eta_{k-1})$ is common to both the $\tilde{a}_k$ term and the $\tilde{z}\tilde{f}_{k-1}$ term and to first order can be factored out of the expression as $(1 + i\mathrm{Imag}(\eta_{k-1}))$. Since (24) implies that relative errors on $\tilde{a}_k$ and $\tilde{z}\tilde{f}_{k-1}$ correspond to relative errors on g_k that are no more than four times larger we get

$$\mathrm{fl}(g_k) = \tilde{g}_k\left(1 + i\mathrm{Imag}(\eta_{k-1}) + 4x_k^{(13)}\big(|\alpha_k| + |\delta_z| + |\mathrm{Re}(\eta_{k-1})| + (13/4)\epsilon\big)\right) + O(\epsilon^2), \tag{25}$$

where $|x_k^{(13)}| \leq 1$. This expression gives the forward errors on g_k.

Analysis for the second formula. We assume that $\mathrm{Re}(t_k) < 0$ and $|a_k| > \sqrt{2}/2$. The reason for choosing μ_k as we have is that it implies that $\mathrm{Imag}(t_k)$ is close to $\mathrm{Imag}(\tilde{t}_k)$. The verification of this is

$$\begin{aligned}
\mathrm{fl}(\mathrm{Imag}(t_k)) &= \mathrm{Imag}\left(\overline{a}_k z f_{k-1}\left(1 + 2\epsilon_k^{(4)}\right)\right) \\
&= \mathrm{Imag}\left(\overline{\tilde{a}}_k \tilde{z}\tilde{f}_{k-1}\left(1 + \alpha_k + i\mu_k + \delta_z + \eta_{k-1} + 2\epsilon_k^{(4)}\right)\right) + O(\epsilon^2) \\
&= \mathrm{Imag}\left(\tilde{t}_k\right)\left(1 + \alpha_k + \delta_z + \mathrm{Re}(\eta_{k-1}) + 2\mathrm{Re}\left(\epsilon_k^{(4)}\right)\right) \\
&\quad + \mathrm{Re}\left(\tilde{t}_k\right)\left(\mu_k + \mathrm{Imag}(\eta_{k-1}) + 2\mathrm{Imag}\left(\epsilon_k^{(4)}\right)\right) + O(\epsilon^2) \\
&= \mathrm{Imag}\left(\tilde{t}_k\right)\left(1 + \alpha_k + \delta_z + \mathrm{Re}(\eta_{k-1}) + 2\mathrm{Re}\left(\epsilon_k^{(4)}\right)\right) + O(\epsilon^2), \qquad (26)
\end{aligned}$$

where $\tilde{t}_k = \overline{\tilde{a}}_k \tilde{z}\tilde{f}_{k-1}$ and $\epsilon_k^{(4)}$ is defined by (21). The last line follows upon substituting in the expression for μ_k. Thus if the Schur parameters are normalized so that $|\alpha_k| \leq |\delta_k| + O(\epsilon^2)$ is small, if δ_z is small, and if $\mathrm{Re}(\eta_{k-1})$ is small, then $\mathrm{Imag}(\tilde{t}_k)$ is close to $\mathrm{Imag}(t_k)$ in a relative sense.

The condition $\mathrm{Re}(t_k) < 0$ implies that $zf_{k-1} - a_k = w_k(1 - \overline{t}_k) + O(\epsilon)$ satisfies

$$|zf_{k-1} - a_k| \geq 1 + O(\epsilon) \geq |a_k| + O(\epsilon). \tag{27}$$

It follows that small relative errors on zf_{k-1} or a_k correspond to small relative errors on $zf_{k-1} - a_k$. Similarly

$$|b_k^2 - 2i\mathrm{Imag}(\overline{a}_k z f_{k-1})| \geq b_k^2 + O(\epsilon)$$

and

$$|b_k^2 - 2i\mathrm{Imag}(\overline{a}_k z f_{k-1})| \geq |2i\mathrm{Imag}(\overline{a}_k z f_{k-1})| + O(\epsilon)$$

so that relative errors due to squaring b_k and multiplying $\mathrm{Imag}(t_k)$ by 2 correspond to small relative errors on g_k. Thus the local errors in computing g_k are

$$\begin{aligned} g_k &= \mathrm{fl}\left(\frac{b_k^2 - 2i\mathrm{Imag}(\overline{a}_k z f_{k-1})}{\overline{zf_{k-1} - a_k}}\right) \\ &= \frac{b_k^2 - 2i\mathrm{fl}\left(\mathrm{Imag}(\overline{a}_k z f_{k-1})\right)}{\overline{zf_{k-1} - a_k}}\left(1 + 5\epsilon_k^{(18)}\right) + O(\epsilon^2), \end{aligned}$$

where $|\epsilon_k^{(18)}| \leq \epsilon$. We have treated the subtraction in the numerator as exact. If we substitute $\tilde{z}$, $\tilde{f}_{k-1}$, and $\tilde{a}_k$ in the denominator, the associated relative errors can also be put on g_k. Since $\tilde{b}_k = b_k$ whenever $|a_k| > \sqrt{2}/2$ we have

$$\begin{aligned} g_k &= \frac{\tilde{b}_k^2 - 2i\mathrm{Imag}(\overline{\tilde{a}}_k \tilde{z}\tilde{f}_{k-1})\left(1 + \alpha_k + \delta_z + \mathrm{Re}(\eta_{k-1}) + 2\mathrm{Re}(\epsilon_k^{(4)})\right)}{\overline{\tilde{z}\tilde{f}_{k-1}}(1 + \delta_z + \overline{\eta}_{k-1}) - \overline{\tilde{a}}_k(1 + \alpha_k + i\mu_k)}(1 + 5\epsilon_k^{(18)}) \\ &\quad + O(\epsilon^2) \\ &= \frac{\tilde{b}_k^2 - 2i\mathrm{Imag}(\overline{\tilde{a}}_k \tilde{z}\tilde{f}_{k-1})\left(1 + \alpha_k + \delta_z + \mathrm{Re}(\eta_{k-1}) + 2\mathrm{Re}(\epsilon_k^{(4)})\right)}{\overline{\tilde{z}\tilde{f}_{k-1}}(1 + \delta_z + \mathrm{Re}(\eta_{k-1})) - \overline{\tilde{a}}_k(1 + \alpha_k - 2i\mathrm{Imag}(\epsilon_k^{(4)}))} \\ &\quad \cdot \left(1 + i\mathrm{Imag}(\eta_{k-1}) + 5\epsilon_k^{(18)}\right) + O(\epsilon^2) \\ &= \tilde{g}_k\left(1 + i\mathrm{Imag}(\eta_{k-1}) + x_k^{(14)}\big(2|\delta_z| + 2|\mathrm{Re}(\eta_{k-1})| + 2|\alpha_k| + 9\epsilon\big)\right) + O(\epsilon^2), \end{aligned} \tag{28}$$

where $|x_k^{(14)}| \leq 1$. In the first line we have used (26). In the second line we have used $\mu_k = -\mathrm{Imag}(\eta_{k-1}) - 2\mathrm{Imag}(\epsilon_k^{(4)})$ to factor out $(1 - i\mathrm{Imag}(\eta_{k-1}))$ from both denominator terms. In the last line we have cast all other errors as small relative errors on $\tilde{g}_k$.

Comparing (28) and (25) and taking the coefficients of (25), which are larger, we conclude that whichever formula is used

$$g_k = \tilde{g}_k\left[1 + i\mathrm{Imag}(\eta_{k-1}) + \omega_k\right] + O(\epsilon^2),$$

where

$$|\omega_k| \leq 4|\delta_z| + 4|\alpha_k| + 4|\mathrm{Re}(\eta_{k-1})| + 13\epsilon.$$

We now turn our attention to the computation of f_k. The local numerical errors in computing f_k from z, f_{k-1}, and g_k are

$$f_k = \mathrm{fl}\left(\frac{\overline{zf_{k-1}}g_k}{\overline{g}_k}\right) = \frac{\overline{zf_{k-1}}g_k}{\overline{g}_k}(1 + 3\epsilon_k^{(19)}) + O(\epsilon^2)$$

for $|\epsilon_k^{(19)}| \le \epsilon$. In terms of $\tilde{g}_k$, $\tilde{z}$, and $\tilde{f}_{k-1}$ we get

$$\begin{aligned} f_k &= \frac{\overline{\tilde{z}\tilde{f}_{k-1}}\tilde{g}_k}{\overline{\tilde{g}}_k(1 - i\mathrm{Imag}(\eta_{k-1}) + \overline{\omega}_k)} \left(1 + 3\epsilon_k^{(19)} + \delta_z + \overline{\eta}_{k-1} + i\mathrm{Imag}(\eta_{k-1}) + \omega_k\right) \\ &\quad + O(\epsilon^2) \\ &= \tilde{f}_k \left(1 + \eta_{k-1} + 3\epsilon_k^{(19)} + \delta_z + 2i\mathrm{Imag}(\omega_k)\right) + O(\epsilon^2). \end{aligned}$$

Since $\eta_k = (f_k - \tilde{f}_k)/\tilde{f}_k$ we have

$$\eta_k = \eta_{k-1} + 3\epsilon_k^{(19)} + \delta_z + 2i\mathrm{Imag}(\omega_k) + O(\epsilon^2)$$

or

$$\eta_k = \eta_{k-1} + \nu_k,$$

where

$$|\nu_k| \le 9|\delta_z| + 8|\alpha_k| + 8|\mathrm{Re}(\eta_{k-1})| + 29\epsilon + O(\epsilon^2). \tag{29}$$

These results are summarized in the following theorem.

THEOREM 5.

$$|a_k|^2 + b_k^2 = 1 + \delta_k, \qquad z = (1 + \delta_z)\tilde{z}$$

$k = 1, 2, \ldots, n$ $b_n = 0$ f_k 3 $g_k \neq 0$ $k = 1, 2, \ldots, n-1$ $\tilde{a}_k$ $\tilde{b}_k$

$$\tilde{a}_k = (1 - \alpha_k + i\mu_k)a_k, \qquad \tilde{b}_k = (1 - \beta_k)b_k$$

α_k β_k μ_k $\tilde{f}_k$ $\tilde{g}_k$ $\tilde{a}_k$ $\tilde{b}_k$ 3

$$f_k = (1 + \eta_k)\tilde{f}_k$$

$$g_k = (1 + i\mathrm{Imag}(\eta_{k-1}) + \omega_k)\,\tilde{g}_k.$$

1. $|\tilde{a}_k|^2 + \tilde{b}_k^2 = 1$
2. $\max(|\alpha_k|, |\beta_k|) \le |\delta_k| + O(\epsilon^2)$
3. μ_k

$$|\mu_k| \le |\mathrm{Imag}(\eta_{k-1})| + 2\epsilon.$$

4. ω_k

$$|\omega_k| \le 4|\delta_z| + 4|\alpha_k| + 4|\mathrm{Re}(\eta_{k-1})| + 13\epsilon + O(\epsilon^2).$$

5. η_k

$$\eta_k = \eta_{k-1} + \nu_k$$

$$\eta_0 = 0 \qquad |\nu_k| \le 9|\delta_z| + 8|\alpha_k| + 8|\mathrm{Re}(\eta_{k-1})| + 29\epsilon + O(\epsilon^2).$$

Recall that we are treating $\mathrm{Re}(\eta_k)$ as a normalization error on f_k. The normalization errors were bounded in section 4. However, it is worth noting that if all normalization errors are small, then we have already proven that the computation of f_k is stable. In particular, there are perfectly normalized $\tilde{a}_k$ and $\tilde{b}_k$, defined as in the theorem, such that the relative error on f_k satisfies

$$\left|\frac{f_k - \tilde{f}_k}{\tilde{f}_k}\right| = |\eta_k| \leq \sum_{j=1}^{k} |\nu_k| \leq 9k|\delta_z| + 29k\epsilon + 8\sum_{j=1}^{k} |\delta_j| + \delta_f(j-1) + O(\epsilon^2)$$

or

$$|\eta_k| \leq 9k|\delta_z| + 29k\epsilon + 8k\delta + 8\delta_f^{(1)}(k-1) + O(\epsilon^2).$$

6. Forward errors. The computation of c_k is the only nontrivial portion of the algorithm left to analyze. We start by assuming that

$$c_k = (1+\zeta_k)\tilde{c}_k.$$

Since $c_0 = \tilde{c}_0 = 1$ is exact, we have $\zeta_0 = 0$. Recall that p_k is a scaled Szegö polynomial so that when $\tilde{H}$ is unreduced $\tilde{p}_k \neq 0$ for $k = 1, 2, \ldots, n-1$. It follows that $\tilde{c}_k = \tilde{p}_k/\tilde{r}_k \neq 0$ for $k = 1, 2, \ldots, n-1$. Since $\tilde{b}_k = (1-\beta_k)b_k$, if $|\beta_k| \leq |\delta_k| + O(\epsilon^2) < 1$ and H is unreduced, then $\tilde{H}$ is unreduced. Thus $\tilde{c}_k \neq 0$ and $\zeta_k = (c_k - \tilde{c}_k)/\tilde{c}_k$ is well defined under the assumption that H is unreduced and $|\delta_k| < 1$.

We show that the relative error ζ_k is small by constructing a recurrence for ζ_k in terms of ζ_{k-1}. The recurrence immediately leads to an upper bound for $|\zeta_k|$. The local errors in the computation of c_k are

$$c_k = \frac{g_k f_{k-1} c_{k-1}}{\sqrt{|g_k f_{k-1} c_{k-1}|^2 + b_k^2}} \left(1 + 9\epsilon_k^{(20)}\right) + O(\epsilon^2),$$

where $|\epsilon_k^{(20)}| \leq \epsilon$. We wish to derive an expression for c_k in terms of $\tilde{g}_k$, $\tilde{f}_{k-1}$, and $\tilde{c}_{k-1}$. We start by changing c_{k-1} to $(1+\zeta_{k-1})\tilde{c}_{k-1}$ to get

$$\begin{aligned}
c_k &= \frac{g_k f_{k-1} \tilde{c}_{k-1}}{\sqrt{|g_k f_{k-1} \tilde{c}_{k-1}|^2 (1 + 2\mathrm{Re}(\zeta_{k-1})) + b_k^2}} \left(1 + 9\epsilon_k^{(20)} + \zeta_{k-1}\right) + O(\epsilon^2) \\
&= \frac{g_k f_{k-1} \tilde{c}_{k-1}}{\sqrt{|g_k f_{k-1} \tilde{c}_{k-1}|^2 + b_k^2}} \cdot \frac{1}{\sqrt{1 + \frac{2|g_k f_{k-1} \tilde{c}_{k-1}|^2}{|g_k f_{k-1} \tilde{c}_{k-1}|^2 + b_k^2} \mathrm{Re}(\zeta_{k-1})}} \left(1 + 9\epsilon_k^{(20)} + \zeta_{k-1}\right) \\
&\quad + O(\epsilon^2) \\
&= \frac{g_k f_{k-1} \tilde{c}_{k-1}}{\sqrt{|g_k f_{k-1} \tilde{c}_{k-1}|^2 + b_k^2}} \left(1 + 9\epsilon_k^{(20)} + \zeta_{k-1} - |\tilde{c}_k|^2 \mathrm{Re}(\zeta_{k-1})\right) + O(\epsilon^2).
\end{aligned}$$

Similarly we can replace f_{k-1} with $(1+\eta_{k-1})\tilde{f}_{k-1}$, g_k with $(1 + i\mathrm{Imag}(\eta_{k-1}) + \omega_k)\tilde{g}_k$, and b_k with $(1+\beta_k)\tilde{b}_k$ to get

$$\begin{aligned}
c_k = \frac{\tilde{g}_k \tilde{f}_{k-1} \tilde{c}_{k-1}}{\sqrt{\left|\tilde{g}_k \tilde{f}_{k-1} \tilde{c}_{k-1}\right|^2 + \tilde{b}_k^2}} \Bigg(& 1 + 9\epsilon_k^{(20)} + \zeta_{k-1} - |\tilde{c}_k|^2 \mathrm{Re}(\zeta_{k-1}) + \eta_{k-1} \\
& - |\tilde{c}_k|^2 \mathrm{Re}(\eta_{k-1}) + i\mathrm{Imag}(\eta_{k-1}) + \omega_k - |\tilde{c}_k|^2 \mathrm{Re}(\omega_k) - \tilde{s}_k^2 \beta_k \Bigg) + O(\epsilon^2).
\end{aligned}$$

The relative errors on f_{k-1} and g_k have been put on c_k by the same method as was used for the relative errors on c_{k-1}. The relative error on b_k has been handled similarly using

$$\tilde{s}_k^2 = \frac{\tilde{b}_k^2}{|\tilde{g}_k \tilde{f}_{k-1} \tilde{c}_{k-1}|^2 + \tilde{b}_k^2}.$$

It follows that

$$|\zeta_k| \le |\zeta_{k-1}| + 9\epsilon + 2|\eta_{k-1}| + |\omega_k| + \tilde{s}_k^2 |\beta_k| + O(\epsilon^2).$$

Substituting in the bounds on η_{k-1}, ω_k, and β_k from Theorem 5 gives

$$\begin{aligned}|\zeta_k| \le |\zeta_{k-1}| &+ (58k-36)\epsilon + (18k-14)|\delta_z| + (16k-11)\delta + 16\delta_f^{(1)}(k-2) + 4\delta_f(k-1) \\ &+ O(\epsilon^2).\end{aligned}$$

Summing this gives

$$\begin{aligned}|\zeta_k| &\le \sum_{l=1}^{k} (58l-36)\epsilon + (18l-14)|\delta_z| + (16l-11)\delta + 16\delta_f^{(1)}(l-2) + 4\delta_f(l-1) \\ &\quad + O(\epsilon^2) \\ &= (29k^2-7k)\epsilon + (9k^2-5k)|\delta_z| + (8k^2-3k)\delta + 4\delta_f^{(1)}(k-1) \\ &\quad + 16\delta_f^{(2)}(k-2) + O(\epsilon^2). \qquad (30)\end{aligned}$$

Bounding the errors on p_k, r_k, s_k, $\hat{b}_k$, and $\hat{a}_k$ is straightforward but tedious.[1] We start with p_k for which

$$p_k = g_k \overline{f}_{k-1} c_{k-1} \left(1 + 2\epsilon_k^{(13)}\right) + O(\epsilon^2),$$

where $\epsilon_k^{(13)} \le \epsilon$. Substituting in $\tilde{g}_k$, $\tilde{f}_k$, and $\tilde{c}_k$ we get

$$p_k = \tilde{g}_k \overline{\tilde{f}_{k-1}} \tilde{c}_{k-1} \left(1 + 2\epsilon_k^{(13)} + i\mathrm{Imag}(\eta_{k-1}) + \omega_k + \overline{\eta}_{k-1} + \zeta_{k-1}\right) + O(\epsilon^2)$$

from which it follows that

$$\left|\frac{p_k - \tilde{p}_k}{\tilde{p}_k}\right| \le 2\epsilon + 2|\eta_{k-1}| + |\zeta_{k-1}| + |\omega_k| + O(\epsilon^2).$$

Substituting the bounds on $|\eta_{k-1}|$, $|\zeta_{k-1}|$, and $|\omega_k|$ we get

$$\begin{aligned}\left|\frac{p_k - \tilde{p}_k}{\tilde{p}_k}\right| &\le (29k^2-7k-7)\epsilon + (9k^2-5k)|\delta_z| + (8k^2-3k-1)\delta \\ &\quad + 4\delta_f(k-1) + 20\delta_f^{(1)}(k-2) + 16\delta_f^{(2)}(k-3) + O(\epsilon^2). \qquad (31)\end{aligned}$$

For r_k we have

$$r_k = \sqrt{|p_k|^2 + b_k^2}\left(1 + 4\epsilon_k^{(7)}\right) + O(\epsilon^2)$$

[1] The bound on $|\zeta_k|$ and the remaining bounds were derived using the computer algebra program Mupad.

for $|\epsilon_k^{(7)}| \leq \epsilon$. Thus

$$r_k = \sqrt{|\tilde{p}_k|^2 \left(1 + 2\mathrm{Re}\left(\frac{p_k - \tilde{p}_k}{\tilde{p}_k}\right)\right) + \tilde{b}_k^2(1+2\beta_k)}\left(1 + 4\epsilon_k^{(7)}\right) + O(\epsilon^2)$$
$$= \tilde{r}_k\left(1 + 4\epsilon_k^{(7)} + x_k^{(15)}\left(\left|\frac{p_k - \tilde{p}_k}{\tilde{p}_k}\right| + |\beta_k|\right)\right) + O(\epsilon^2).$$

Thus

$$\left|\frac{r_k - \tilde{r}_k}{\tilde{r}_k}\right| \leq 4\epsilon + \left|\frac{p_k - \tilde{p}_k}{\tilde{p}_k}\right| + |\beta_k|.$$

Substituting in the bound for $|\beta_k|$ and the bound for the error on p_k we get

$$\left|\frac{r_k - \tilde{r}_k}{\tilde{r}_k}\right| \leq (29k^2 - 7k - 3)\epsilon + (9k^2 - 5k)|\delta_z| + (8k^2 - 3k)\delta$$
$$+ 4\delta_f(k-1) + 20\delta_f^{(1)}(k-2) + 16\delta_f^{(2)}(k-3) + O(\epsilon^2). \tag{32}$$

In computing s_k the single division results in an error

$$s_k = \frac{b_k}{r_k}\left(1 + \epsilon_k^{(8)}\right),$$

where $|\epsilon_k^{(8)}| \leq \epsilon$ so that

$$s_k = \frac{\tilde{b}_k\left(1 + \epsilon_k^{(8)} + \beta_k\right)}{\tilde{r}_k\left(1 + \frac{r_k - \tilde{r}_k}{\tilde{r}_k}\right)} + O(\epsilon^2) = \tilde{s}_k\left(1 + \epsilon_k^{(8)} + \beta_k - \frac{r_k - \tilde{r}_k}{\tilde{r}_k}\right) + O(\epsilon^2).$$

Thus

$$\left|\frac{s_k - \tilde{s}_k}{\tilde{s}_k}\right| \leq (29k^2 - 7k - 2)\epsilon + (9k^2 - 5k)|\delta_z| + (8k^2 - 3k + 1)\delta$$
$$+ 4\delta_f(k-1) + 20\delta_f^{(1)}(k-2) + 16\delta_f^{(2)}(k-3) + O(\epsilon^2). \tag{33}$$

For the computation of $\hat{b}_{k-1}$ we get

$$\hat{b}_{k-1} = r_k s_{k-1}\left(1 + \epsilon_k^{(21)}\right) = \tilde{r}_k\tilde{s}_{k-1}\left(1 + \epsilon_k^{(21)} + \frac{r_k - \tilde{r}_k}{\tilde{r}_k} + \frac{s_{k-1} - \tilde{s}_{k-1}}{\tilde{s}_{k-1}}\right) + O(\epsilon^2)$$

so that

$$\left|\frac{\hat{b}_k - \tilde{\hat{b}}_k}{\tilde{\hat{b}}_k}\right| \leq (58k^2 + 44k + 18)\epsilon + (18k^2 + 8k + 4)|\delta_z| + (16k^2 + 10k + 6)\delta$$
$$+ 4\delta_f(k) + 4\delta_f(k-1) + 20\delta_f^{(1)}(k-1) + 36\delta_f^{(1)}(k-2)$$
$$+ 32\delta_f^{(2)}(k-3) + O(\epsilon^2). \tag{34}$$

Finally, we consider the computation of $\hat{a}_k$. The relative difference between $\hat{a}_k$ and the perturbed $\tilde{\hat{a}}_k$ is not necessarily small. Nevertheless all quantities involved in

the computation of $\hat{a}_k$ are less than one in magnitude so that

$$\begin{aligned}\hat{a}_k &= |c_k|^2 f_k - \overline{z} s_k^2 a_{k+1} + 6\epsilon_k^{(14)} + O(\epsilon^2)\\ &= |\tilde{c}_k|^2 \tilde{f}_k \left(1 + 2\mathrm{Re}(\zeta_k) + \eta_k\right) + \overline{\tilde{z}} \tilde{s}_k^2 \tilde{a}_{k+1} \left(1 + \delta_z + 2\frac{s_k - \tilde{s}_k}{\tilde{s}_k} + \alpha_{k+1} - i\mu_{k+1}\right)\\ &\quad + 6\epsilon_k^{(14)} + O(\epsilon^2).\end{aligned}$$

Since $|\mu_{k+1}| \leq |\mathrm{Imag}(\eta_k)| + 2\epsilon$ and $|\alpha_{k+1}| \leq \delta$ we have

$$\left|\hat{a}_k - \tilde{\hat{a}}_k\right| \leq 2|\zeta_k| + 2|\eta_k| + |\delta_z| + 2\left|\frac{s_k - \tilde{s}_k}{\tilde{s}_k}\right| + \delta + 8\epsilon + O(\epsilon^2)$$

so that

$$\begin{aligned}\left|\hat{a}_k - \tilde{\hat{a}}_k\right| \leq{}& (116k^2 + 30k + 4)\epsilon + (36k^2 - 2k + 1)|\delta_z| + (32k^2 + 4k + 3)\delta\\ &+ 8\delta_f(k-1) + 24\delta_f^{(1)}(k-1) + 72\delta_f^{(1)}(k-2)\\ &+ 64\delta_f^{(2)}(k-3) + O(\epsilon^2). \qquad (35)\end{aligned}$$

Combining the results of this section with Theorem 5 we have bounds on the relative backward errors on a_k and b_k and bounds on the relative forward errors on g_k, p_k, r_k, $\hat{b}_{k-1}$, c_k, s_k, and f_k. We have a bound on the absolute forward error on $\hat{a}_k$. All of these bounds are given in terms of ϵ, δ, $\delta_f^{(j)}(k)$, and δ_z. If the starting normalization errors δ_k and δ_z are small and if the normalization error $\delta_f(l)$ is small for each $l = 1, 2, \ldots, n$, then we have a proof of stability for a single iteration of Algorithm 2.

7. Final error bounds. We have now bounded all forward and backward errors and all normalization errors. The results are summarized in the following theorem.

THEOREM 6. $b_k \neq 0$ $k = 1, 2, \ldots, n-1$

$$\max_{1\leq k\leq n} \left|1 - |a_k|^2 - b_k^2\right| = \delta,$$

$b_n = 0$ z $|z| = 1 + \delta_z$ $|\delta_z| < 1$ $\delta = O(\epsilon)$ $\delta_z = O(\epsilon)$

$$\frac{2 - \sqrt{2}/2}{1 - \sqrt{2}/2}\left(\epsilon + (n-1)\gamma_3 + n|\delta_z|\right) + 2\gamma_2 < 1,$$

2 $g_k \neq 0$ $k = 1, 2, \ldots, n-1$ $\tilde{a}_k$ $\tilde{b}_k$ $|\tilde{a}_k|^2 + \tilde{b}_k^2 = 1$ $\tilde{z}$

$$|\tilde{a}_k - a_k| \leq \left[(12k^2 - 7k - 3)\epsilon + (8k - 7)\delta + (4k^2 - 3k - 1)|\delta_z|\right] |\tilde{a}_k| + O(\epsilon^2)$$

$$|\tilde{b}_k - b_k| \leq \delta|\tilde{b}_k| + O(\epsilon^2).$$

$\tilde{c}_k$ $\tilde{s}_k$ $\tilde{\hat{a}}_k$ $\tilde{\hat{b}}_{k-1}$ $\tilde{a}_k$ $\tilde{b}_k$ $\tilde{z}$ $\hat{a}_k$ $\hat{b}_{k-1}$ c_k s_k

$$\begin{aligned}\left|\tilde{\hat{a}}_k - \hat{a}_k\right| \leq{}& (32k^3 + 68k^2 + 46k + 4)\epsilon + \frac{1}{3}(32k^3 + 60k^2 + 10k + 3)|\delta_z|\\ &+ (32k^2 + 4k + 3)\delta + O(\epsilon^2),\end{aligned}$$

$$\left|\tilde{\hat{b}}_k - \hat{b}_k\right| \le \Big[(16k^3+46k^2+52k+18)\epsilon + \frac{1}{3}(16k^3+42k^2+32k+12)|\delta_z| + (16k^2+10k+6)\delta\Big]|\tilde{\hat{b}}_{k-1}| + O(\epsilon^2),$$

$$|\tilde{c}_k - c_k| \le \left[(8k^3+11k^2+3k)\epsilon + \frac{1}{3}(8k^3+9k^2-5k)|\delta_z| + (8k^2-3k)\delta\right]|\tilde{c}_k| + O(\epsilon^2),$$

$$|\tilde{s}_k - s_k| \le \Big[(8k^3+11k^2+3k-2)\epsilon + \frac{1}{3}(8k^3+9k^2-5k)|\delta_z| + (8k^2-3k+1)\delta\Big]|\tilde{s}_k| + O(\epsilon^2).$$

. The stated condition on the unit roundoff and $|\delta_z|$ implies both of the conditions assumed in Theorem 2 so that $g_k \neq 0$ for $k = 1, 2, \ldots, n-1$. All that is needed to prove the forward error bounds is to substitute (13), (14), and (15) into (35), (34), (30), and (33). Similarly the bounds on the backward errors are from Theorem 5 combined with the bounds on $\delta_f^{(j)}(k)$. □

In terms of normwise errors we have the following.

THEOREM 7. 6 H Q $\hat{H}$ $\tilde{H}$ $\tilde{Q}$ $\tilde{\hat{H}}$ 3

$$\hat{H} = \tilde{Q}^{\mathrm{H}}\,[H+E]\,\tilde{Q}$$

$$\|E\|_2 \le \frac{1}{3}(48n^4+256n^3+405n^2+203n)\epsilon + \frac{1}{3}(16n^4+80n^3+89n^2+25n)|\delta_z| + \frac{1}{3}(64n^3+132n^2+44n)\delta + O(\epsilon^2)$$

$$\|Q-\tilde{Q}\|_2 \le \frac{1}{3}(12n^4+46n^3+54n^2+20n)\epsilon + \frac{1}{3}(4n^4+14n^3+8n^2-2n)|\delta_z| + \frac{1}{3}(16n^3+15n^2+5n)\delta + O(\epsilon^2).$$

. In terms of the notation used in section 3, we see from Theorem 6 that

$$\max(K_a(\epsilon,\delta,\delta_z,k), K_b(\epsilon,\delta,\delta_z,k)) \le (12k^2-7k-3)\epsilon + (4k^2-3k-1)|\delta_z| + (8k-7)\delta + O(\epsilon^2),$$

$$\max(K_{\hat{a}}(\epsilon,\delta,\delta_z,k), K_{\hat{b}}(\epsilon,\delta,\delta_z,k)) \le (32k^3+68k^2+46k+4)\epsilon + \frac{1}{3}(32k^3+60k^2+10k+3)|\delta_z| + (32k^2+4k+3)\delta + O(\epsilon^2),$$

and

$$\max(K_c(\epsilon,\delta,\delta_z,k),K_s(\epsilon,\delta,\delta_z,k)) \le (8k^3+11k^2+3k)\epsilon + \frac{1}{3}(8k^3+9k^2-5k)|\delta_z| + (8k^2-3k+1)\delta + O(\epsilon^2).$$

The theorem then follows from (9) and (8). □

Although we have stated the bounds only for a single UHQR iteration, the observations of section 3 apply. The normalization errors on the Schur parameters grow at worst linearly in the number of iterations so that a sequence of $j+1$ UHQR iterations will compute H_{j+1} that is similar to a matrix $\tilde{H}$ that is close to $H_0 = H$. The similarity transformation is the product $\tilde{Q} = \tilde{Q}_0\tilde{Q}_1\cdots\tilde{Q}_j$. The product of computed transformations $Q = Q_0Q_1\cdots Q_j$ is close to $\tilde{Q}$. If H_{j+1} is diagonal, then the diagonal elements are eigenvalues of $\tilde{H}$. The matrix Q is close to the matrix of eigenvectors of $\tilde{H}$.

We have not taken into account errors due to neglecting a small b_k when performing a deflation. The deflation contributes an additional error to the bound on $\|H_0 - \tilde{H}_0\|$. The error is proportional to the size of the subdiagonal neglected. Neglecting b_k also affects the normalization error on a_k. Suppose that for some small b_k we have before deflation $|a_k|^2 + b_k^2 = 1 + \delta_k$. After setting b_k to zero we have a new normalization error associated with a_k only, $|a_k|^2 = 1 + (\delta_k - b_k^2)$. Thus in neglecting b_k we have $\delta_k \leftarrow \delta_k - b_k^2$. If $b_k = O(\epsilon)$ the effect on δ_k is $O(\epsilon^2)$.

Finally, we warn against relying on implicit normalization of the projected shift. Suppose that the algorithm is implemented using the projected Wilkinson shift computed directly as an eigenvalue of (3) without explicit normalization. The unimodularity of the eigenvalue depends on the normalization on a_{n-1} and b_{n-1}. If the Schur parameters are not properly normalized, the computed eigenvalue of (3) is not perfectly unimodular. This increases the normalization error in the computed parameters $\hat{a}_{n-1}$ and $\hat{b}_{n-1}$, leading to even poorer normalization of the shift in the next iteration. The process is highly unstable and can quickly lead to overflow in the Schur parameters. The explicit normalization $z \leftarrow z/|z|$ solves the problem ensuring that $|\delta_z|$ is less than a fixed multiple of ϵ regardless of the normalization on a_{n-1} and b_{n-1}.

Acknowledgment. The author would like to thank Bill Gragg for inspiring and encouraging this work.

REFERENCES

[1] A. Bunse-Gerstner and L. Elsner, *Schur parameter pencils for the solution of the unitary eigenproblem*, Linear Algebra Appl., 154/156 (1991), pp. 741–778.

[2] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., Johns Hopkins University Press, Baltimore, MD, 1989.

[3] W. B. Gragg, *The QR algorithm for unitary Hessenberg matrices*, J. Comput. Appl. Math., 16 (1986), pp. 1–8. Cited in [2].

[4] W. B. Gragg, *Stabilization of the UHQR algorithm*, in Advances in Computational Mathematics, Z. Chen, Y. Li, C. Micchelli, and Y. Xu, eds., Marcel Dekker, New York, 1999, pp. 139–154.

[5] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.

[6] M. Stewart, *Stability properties of several variants of the unitary Hessenberg QR algorithm*, Contemp. Math. 281, AMS, Providence, RI, 2001.

[7] G. Szegö, *Orthogonal Polynomials*, 3rd ed., AMS, Providence, RI, 1967.

[8] T.-L. Wang and W. B. Gragg, *Convergence of the unitary QR algorithm with a unimodular Wilkinson shift*, Math. Comp., 72 (2003), pp. 375–385.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 1, pp. 68–80

COMPRESSIONS OF TOTALLY POSITIVE MATRICES*

SHAUN M. FALLAT†, ALLEN HERMAN†, MICHAEL I. GEKHTMAN‡, AND CHARLES R. JOHNSON§

Abstract. A matrix is called totally positive if all its minors are positive. If a totally positive matrix A is partitioned as $A = (A_{ij})\ i, j = 1, 2, \ldots, k$, in which each block A_{ij} is $n \times n$, we show that the $k \times k$ compressed matrix given by $(\det A_{ij})$ is also totally positive and that the determinant of the compressed matrix exceeds $\det A$ when $k = 2, 3$. An extension that allows for overlapping blocks is also presented when $k = 2, 3$. For $k \geq 4$ we verify, by example, that the $k \times k$ compressed matrix of a totally positive matrix need not be totally positive.

Key words. totally positive matrices, generalized matrix functions, compressed matrix, bidiagonal factorizations

AMS subject classifications. Primary, 15A48; Secondary, 05C38, 05C50, 15A15

DOI. 10.1137/S0895479803437827

1. Introduction.

An $n \times n$ matrix A is called [illegible] (TP) (resp., [illegible] (TN)) if every minor of A is positive (nonnegative) (see [1, 6, 10]). Such matrices have a wide variety of applications in approximation theory, numerical mathematics, statistics, and combinatorics [7].

We consider $nk \times nk$ partitioned matrices $A = (A_{ij})_{i,j=1}^{k}$, in which each block A_{ij} is $n \times n$. It has long been known (see [18], for example) that if A is a Hermitian positive definite matrix, then the $k \times k$ [illegible] (or [illegible] A)

$$C_k(A) = (\det A_{ij})_{i,j=1}^{k}$$

is also positive definite, and

$$\det C_k(A) > \det A.$$

Analogous results have also been shown to hold for M-matrices (namely, matrices with nonpositive off-diagonal entries and entrywise positive inverses), with the extra condition that the comparison matrix of the compression of A is used in place of the usual compression of A (see [9]).

Naturally, comparing and identifying common properties between these positivity classes and totally positive matrices is both important and useful.

Thus we formulate the corresponding problem in the context of totally positive matrices and ask, If A is a totally positive $nk \times nk$ matrix, then is the compressed matrix $C_k(A)$ a totally positive matrix?

Due to the evident inductive nature of this problem, for general k it suffices to prove that $\det(C_k(A)) > 0$, because determinants of proper square submatrices of $C_k(A)$ can be assumed positive by induction.

*Received by the editors November 13, 2003; accepted for publication (in revised form) by H. J. Woerdeman September 16, 2005; published electronically March 17, 2006.

http://www.siam.org/journals/simax/28-1/43782.html

†Department of Mathematics and Statistics, University of Regina, Regina, Saskatchewan, S4S 0A2 (sfallat@math.uregina.ca, aherman@math.uregina.ca). Research supported in part by an NSERC research grant.

‡Department of Mathematics, University of Notre Dame, Notre Dame, IN 46556-5683 (Michael.Gekhtman.1@nd.edu). Research supported in part by NSF grant DMS 0400484.

§Department of Mathematics, College of William and Mary, Williamsburg, VA 23187-8795 (crjohnso@math.wm.edu).

On the other hand, we note that $C_k(A)$ is a principal submatrix of the kth compound of A (see [1]). Using this fact, it is easy to conclude that $C_k(A)$ is positive semidefinite whenever A is positive semidefinite. Furthermore, recall that the $(n-1)$st compound of an $n \times n$ totally nonnegative matrix is again totally nonnegative (see [1]). However, for $1 < k < n-1$, the kth compound of a totally nonnegative matrix need not be totally nonnegative. Further along these lines if we consider Sylvester's determinantal identity (see [6, p. 12 or p. 92]), it is known that if A is totally positive, then the matrix corresponding to Sylvester's determinantal identity (see also a related result in [15, Prop. 2]) is also totally positive.

In an effort to answer these questions we consider small values of k ($k = 2, 3$) initially and provide positive resolutions in these cases. However, it is shown by example that for $k \geq 4$, the compression of a totally positive matrix need not be totally positive. This is in stark constrast to the situations for both positive semidefinite matrices and for M-matrices.

The rest of the paper is organized as follows. In section 2 we consider the case $k = 2$ in complete detail. Section 3 begins by identifying the determinant of the compressed matrix as a generalized matrix function. We then make use of the related generalized Cauchy–Binet identity in concert with an associated bidiagonal factorization of TP matrices. These results yield an affirmative answer to our question for the case $k = 3$. Section 5 contains an example of an 8×8 totally nonnegative matrix A for which $C_4(A)$ is not totally nonnegative. Finally, in section 6, we derive some interesting consequences for the cases $k = 2, 3$.

2. The case $k = 2$. Let A be presented as follows:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

where A_{ij} are $n \times n$.

Clearly, in this case to verify that the compression of A is totally positive it is enough to show that $\det(C_2(A)) > 0$. In fact we will prove a little more by demonstrating that not only is $\det(C_2(A))$ positive, but it can be naturally expanded as a subtraction-free expression involving only nonnegative terms. To accomplish this we make use of the fact that any totally positive matrix can be written as the product LU (see [4]), where L (resp., U) is a lower (resp., upper) triangular matrix and is ΔTP (see definition below).

We now introduce some important notation. Let A be an $n \times n$ real matrix, and let α, β be nonempty subsets of $\{1, 2, \ldots, n\}$, arranged in increasing order. Then $A[\alpha|\beta]$ denotes the submatrix of A lying in rows indexed by α and columns indexed by β. If, in addition, $\alpha = \beta$, then we abbreviate the [illegible] $A[\alpha|\alpha]$ to $A[\alpha]$. An $m \times m$ lower (upper) triangular matrix A is called [illegible], denoted by ΔTP, if for each $l = 1, 2, \ldots, m$ and for each pair of index sets $\alpha = \{i_1, \ldots, i_l\}$ and $\beta = \{j_1, \ldots, j_l\}$ with the property that $i_s \geq j_s$ ($i_s \leq j_s$) for $s = 1, 2, \ldots, l$, the minor $\det A[\alpha|\beta]$ is positive. We are now in a position to prove the result.

LEMMA 2.1. [illegible] A [illegible] $2n \times 2n$ [illegible]

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

[illegible] A_{ij} [illegible] $n \times n$ [illegible] $C_2(A)$ [illegible]

$$\det C_2(A) > \det A.$$

. Since A is totally positive A can be written as $A = LU$, where L and U are ΔTP matrices. We can partition L and U into $n \times n$ blocks and rewrite $A = LU$ as

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix} \begin{pmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{pmatrix}.$$

Observe that

$$\det A_{11} \det A_{22} = \det(L_{11}U_{11}) \det \left([L_{21}L_{22}] \begin{bmatrix} U_{12} \\ U_{22} \end{bmatrix} \right).$$

Applying the classical Cauchy–Binet identity to the far right term above yields

$$\det \left([L_{21}L_{22}] \begin{bmatrix} U_{12} \\ U_{22} \end{bmatrix} \right) = \sum_{\gamma} \det L[\{n+1, \ldots, 2n\}|\gamma] \det U[\gamma|\{n+1, \ldots, 2n\}], \tag{2.1}$$

where the sum is taken over all ordered subsets γ of $\{1, 2, \ldots, 2n\}$ with cardinality n. If we separate the terms with $\gamma = \{1, 2, \ldots n\}$ and $\gamma = \{n+1, n+2, \ldots 2n\}$, then the sum on the right in (2.1) reduces to

$$\sum_{\gamma} \det L[\{n+1, \ldots, 2n\}|\gamma] \det U[\gamma|\{n+1, \ldots, 2n\}]$$
$$= \det L_{21} \det U_{12} + \det L_{22} \det U_{22} + (\text{positive terms}).$$

Since L and U are ΔTP, all summands are positive.

Hence

$$\det A_{11} \det A_{22} = \det(L_{11}U_{11})[\det L_{21} \det U_{12} + \det L_{22} \det U_{22}] + (\text{positive terms}),$$

which is equivalent to

$$\det A_{11} \det A_{22} = \det A_{12} \det A_{21} + \det A + (\text{positive terms}).$$

Thus we have

$$\det C_2(A) = \det A_{11} \det A_{22} - \det A_{12} \det A_{21}$$
$$= \det A + (\text{positive terms}),$$

and so $\det C_2(A) > \det A > 0$, which completes the proof. □

The following is a consequence of Lemma 2.1 and the classical fact (see [10], for example) that the TP matrices are dense in the TN matrices.

Corollary 2.2. A $2n \times 2n$

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

A_{ij} $n \times n$ $C_2(A)$

$$\det C_2(A) \geq \det A.$$

Next we establish a connection between $\det(C_k(A))$ and a certain generalized matrix function, which is included for completeness.

3. $\det(C_k(A))$ as a generalized matrix function. For brevity, we let $N = nk$. Let S_N be the group of permutations of $\{1, \ldots, N\}$, which we view as acting on the row and column indices of A. For each $g \in S_N$, the [illegible] of A determined by g is

$$A_g = a_{1,g(1)} a_{2,g(2)} \cdots a_{N,g(N)}.$$

The collection of all generalized diagonals of A determines a specific element of the group algebra $\mathbb{R}S_N$ given by

$$[A] = \sum_{g \in S_N} A_g g.$$

If χ is a character of a subgroup H of S_N, then χ can be viewed as a linear map from $\mathbb{R}S_N$ to $\mathbb{R}$ by extending it as 0 on $S_N \setminus H$. Applying this to $[A]$ we have

$$\chi[A] = \sum_{g \in H} A_g \chi(g).$$

A [illegible] is a function $d_\chi : M_N(\mathbb{R}) \to \mathbb{R}$ such that $d_\chi(A) = \chi[A]$, for a fixed character χ of some subgroup of S_N. For a general reference on generalized matrix functions see [14].

The aim of this section is to show that there exists a real-valued character χ of a subgroup H of S_N having degree 1 such that $\det(C_k(A)) = d_\chi(A)$ for all $N \times N$ matrices A.

We begin by describing the subgroup H. For each $s \in \{1, \ldots, k\}$, let $S_n(s)$ be the subgroup of S_N consisting of all permutations on the set of n consecutive indices $\Delta_s = \{(s-1)n + 1, (s-1)n + 2, \ldots, sn\}$. The subgroups $S_n(s)$, $s = 1, \ldots, k$, generate an internal direct product $S_n(1) \times S_n(2) \times \cdots \times S_n(k)$ as a subgroup of S_N. The k distinct blocks $\Delta_1, \ldots, \Delta_k$ of n consecutive indices are also permuted by a subgroup $\tilde{S}_k$ of S_N. This subgroup is isomorphic to S_k. The element of $\tilde{S}_k$ sending Δ_s to Δ_t is represented in S_N as the product of the nonoverlapping 2-cycles $((s-1)n+u, (t-1)n+u)$, $u = 1, \ldots, n$. The subgroup $\tilde{S}_k$ normalizes $S_n(1) \times \cdots \times S_n(k)$, so the subgroup H generated by $\tilde{S}_k$ and $S_n(1) \times \cdots \times S_n(k)$ in S_N is a semidirect product of the form $(S_n(1) \times \cdots \times S_n(k)) \rtimes \tilde{S}_k$. This is the subgroup H we require. Note that this subgroup H is one of several isomorphic copies of the wreath product group $S_n \wr S_k$ occurring as subgroups of S_N.

Let $\tau \to \tilde{\tau}$ be the natural isomorphism from S_k to $\tilde{S}_k$. Then the sign character ε_k corresponds to a character $\tilde{\varepsilon}_k$ on $\tilde{S}_k$ by $\tilde{\varepsilon}_k(\tilde{\tau}) = \varepsilon_k(\tau)$ for $\tau \in S_k$. The function χ is defined as follows. Since any element h of H can be uniquely written in the form $\tilde{\tau} \cdot (\tau_1 \tau_2 \cdots \tau_k)$, where each $\tau_s \in S_n(s)$ for $s = 1, 2, \ldots, k$, and $\tilde{\tau} \in \tilde{S}_k$,

$$\chi(h) = \tilde{\varepsilon}_k(\tilde{\tau}) \varepsilon_H(\tau_1 \tau_2 \cdots \tau_k).$$

Observe that $\chi(\tau) = -1$ for every 2-cycle $\tau \in H$.

THEOREM 3.1. [illegible] H [illegible] χ [illegible] $\det(C_k(A)) = d_\chi(A)$ [illegible] $N \times N$ [illegible] A [illegible]. We begin by examining the standard formula for $\det(C_k(A))$ as the generalized matrix function corresponding to the sign character ε_k of the symmetric group

S_k. That is,

$$\begin{aligned}\det(C_k(A)) &= \sum_{\tau\in S_k} \varepsilon_k(\tau)(\det A_{1,\tau(1)})\dots(\det A_{k,\tau(k)}) \\ &= \sum_{\tau\in S_k} \varepsilon_k(\tau)\prod_{s=1}^{k}\left(\sum_{\tau_s\in S_n(s)}\varepsilon(\tau_s)\prod_{u_s=1}^{n}(A_{s,\tau(s)})_{u_s,\tau_s(u_s)}\right).\end{aligned}$$

Note that the block submatrices $A_{s,\tau(s)}$ have entries $a_{(s-1)n+u_s,(\tau(s)-1)n+v_s}$, where $1\le u_s, v_s \le n$, for all $s = 1,\dots,k$. Therefore, we have

$$\det(C_k(A)) = \sum_{\tau\in S_k}\varepsilon_k(\tau)\prod_{s=1}^{k}\left(\sum_{\tau_s\in S_n}\varepsilon(\tau_s)\left(\prod_{u_s=1}^{n}a_{(s-1)n+u_s,(\tau(s)-1)n+\tau_s(u_s)}\right)\right).$$

Now, any element h of H can be uniquely written in the form $\tilde{\tau}\cdot(\tau_1\tau_2\dots\tau_k)$, where each $\tau_s \in S_n(s)$ for $s = 1,\dots,k$, and $\tilde{\tau}\in\tilde{S}_k$. If $i\in\{1,\dots,N\}$, then i can be uniquely written as $i = (s-1)n+u$ for some $s\in\{1,\dots,k\}$ and $u\in\{1,\dots,n\}$, and the value of $h(i)$ can be computed as follows:

$$\begin{aligned}h(i) &= h((s-1)n+u) \\ &= (\tilde{\tau}\cdot(\tau_1\tau_2\dots\tau_k))((s-1)n+u)) \\ &= \tilde{\tau}((s-1)n+\tau_s(u)) \\ &= (\tau(s)-1)n+\tau_s(u),\end{aligned}$$

where $\tau\in S_k$. Thus

$$\begin{aligned}\det(C_k(A)) &= \sum_{\tau\in S_k}\sum_{(\tau_1\tau_2\dots\tau_k)\in S_n(1)\times\cdots\times S_n(k)}\varepsilon_k(\tau)\varepsilon_H(\tau_1\tau_2\dots\tau_k) \\ &\quad\cdot\left(\prod_{s=1}^{k}\prod_{u_s=1}^{n}a_{(s-1)n+u_s,(\tau(s)-1)n+\tau_s(u_s)}\right) \\ &= \sum_{\tilde{\tau}\in\tilde{S}_k}\left(\sum_{(\tau_1\tau_2\dots\tau_k)\in S_n(1)\times\cdots\times S_n(k)}\tilde{\varepsilon}_k(\tilde{\tau})\varepsilon_H(\tau_1\tau_2\dots\tau_k)\left(\prod_{i=1}^{N}a_{i,\tilde{\tau}\cdot\tau_1\tau_2\dots\tau_k(i)}\right)\right) \\ &= \sum_{h\in H}\chi(h)A_h \\ &= \chi[A] \\ &= d_\chi(A),\end{aligned}$$

and the theorem follows. □

Note that if we let $K = \ker\chi$, then K is a subgroup of index 2 in H. In fact, since we are assuming $n > 1$, we have that the two distinct left cosets of K in H are K and σK, where σ can be chosen to be the permutation $(1,2)\in S_N$. We remark here that the above result is mentioned in passing in [9].

Since χ is a real valued character of degree 1, Theorem 3.1 allows us to apply the generalized Cauchy–Binet theorem of Marcus and Minc [13, Lemma 2.3] for the character corresponding to the determinant of the k-compression. We will use this in combination with the standard bidiagonal factorization of a totally positive matrix.

4. The generalized Cauchy–Binet theorem and the case $k = 3$. To state the version of the generalized Cauchy–Binet theorem needed, we first make some definitions. Let Γ_N denote the set of all N-multisets $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_N)$ with $\gamma_i \in \{1, 2, \ldots, N\}$ for all $i = 1, 2, \ldots, N$. When $N = nk$, we divide $\gamma \in \Gamma_N$ further into k blocks of n-multisets, so $\gamma = (\gamma^1, \gamma^2, \ldots, \gamma^k)$ with $\gamma^j = (\gamma_{(j-1)n+1}, \gamma_{(j-1)n+2}, \ldots, \gamma_{jn})$ for all $j = 1, 2, \ldots, k$. S_N acts naturally on Γ_N by permuting indices, so ${}^\sigma\gamma = (\gamma_{\sigma(1)}, \gamma_{\sigma(2)}, \ldots, \gamma_{\sigma(N)})$. By Theorem 3.1, the determinant of the k-compression of an $N \times N$ matrix can be identified with the generalized matrix function d_χ corresponding to the real-valued degree 1 character χ of the aforementioned subgroup H of S_N. Now, H acts naturally on Γ_N by restriction. Let Δ denote the set of representatives for the orbits of H on Γ_N whose elements are chosen to be minimal in lexicographic order among all the elements in their orbit. For each $\gamma \in \Delta$, let H_γ denote the stabilizer in H of γ, and let $v(\gamma) = |H_\gamma|$. Finally, let $\bar{\Delta} = \{\gamma \in \Delta : \chi_{H_\gamma} = 1_{H_\gamma}\}$. Then $\bar{\Delta}$ is the set of all those $\gamma = (\gamma^1, \gamma^2, \ldots, \gamma^k) \in \Gamma_N$ for which each γ^j is a strictly increasing n-multiset, and $\gamma^1 < \gamma^2 < \cdots < \gamma^k$ in the lexicographic order on n-multisets.

Now we state the version of the generalized Cauchy–Binet theorem that we need. We use $\star$ to denote the N-multiset, $\star = (1, 2, 3, \ldots, N)$.

THEOREM 4.1. *Let L and U be $N \times N$ [illegible] $A = LU$ [illegible]*

$$d_\chi(A) = \sum_{\alpha \in \bar{\Delta}} \frac{1}{v(\alpha)} d_\chi(L[\star|\alpha]) d_\chi(U[\alpha|\star]).$$

We apply this formula to the case of an $N \times N$ totally positive matrix A. Such a matrix has a decomposition of the form $A = LDU$, where L is a unipotent lower triangular ΔTP matrix, D is a diagonal matrix with positive main diagonal entries, and U is a unipotent upper triangular ΔTP matrix. By factoring appropriate positive scalars, we may assume for our purposes that $D = I$, so $A = LU$ for ΔTP matrices L and U (see [4]). It is interesting to observe the following property, which is true for all generalized matrix functions corresponding to real-valued characters of subgroups of S_N. We omit the straightforward proof.

LEMMA 4.2. *Let B be $N \times N$ [illegible] $\gamma, \delta \in \Gamma_N$ [illegible] $d_\chi(B[\gamma|\delta]) = d_\chi(B^T[\delta|\gamma])$*

For $1 \le i, j \le N$, let E_{ij} denote the $N \times N$ matrix whose only nonzero entry is a 1 in the (i, j) position. If t is a positive scalar, then let $E_r(t) = I + tE_{r,r-1}$. If L is a lower triangular ΔTP matrix with unit main diagonal, then the standard bidiagonal factorization of L is given by

$$L = \prod_{s=2}^{N} \left(\prod_{r=s}^{N} E_r(t_{s,r}) \right)$$

for uniquely chosen positive scalars $t_{s,r}$. Such a factorization is called an [illegible] and has proved to be a useful tool in the study of TN matrices (see [2, 4, 5, 8, 12]).

By repeatedly applying the generalized Cauchy–Binet theorem to this standard bidiagonal factorization of L, we can express $d_\chi(L[\star|\alpha])$ for any $\alpha \in \bar{\Delta}$ as a nonnegative linear combination of terms which are products of factors of the form $d_\chi(E_r(t_{s,r})[\gamma|\delta])$ for $\gamma, \delta \in \bar{\Delta}$. For any $r = 2, \ldots, N$, $t > 0$, and $\gamma, \delta \in \bar{\Delta}$, we have that $d_\chi(E_r(t)[\gamma|\delta]) = \det(C_k(E_r(t)[\gamma|\delta]))$, and $(C_k(E_r(t)[\gamma|\delta]))_{p,q} = \det(E_r(t)[\gamma^p|\delta^q])$ for all $1 \le p, q \le k$.

LEMMA 4.3. [illegible] $\gamma, \delta \in \bar{\Delta}$ [illegible] $p, q = 1, 2, \ldots, k$

$$\det(E_r(t)[\gamma^p|\gamma^q]) = \begin{cases} 1 & \text{[illegible] } \delta^q = \gamma^p, \\ t & \text{[illegible] } \delta^q = (\gamma^p \setminus \{r\}) \cup \{r-1\}, \\ 0 & \text{[illegible]} \end{cases}$$

[illegible]. Suppose $\det(E_r(t)[\gamma^p|\gamma^q]) \neq 0$. Since $\gamma, \delta \in \bar{\Delta}$, the n-multisets γ^p and δ^q are strictly increasing. If $r \notin \gamma^p$, then $E_r(t)[\gamma^p|\star]$ is an $n \times N$ matrix each of whose ith rows contain only one nonzero entry, which is a 1 in the $(\gamma^p)_i$th column. So in order to choose a strictly increasing n-multiset δ^q so that every row of $E_r(t)[\gamma^p|\delta^q]$ has a nonzero entry, we have to choose $(\delta^q)_i = (\gamma^p)_i$ for all $i = 1, \ldots, n$, so $\delta^q = \gamma^p$. In this case, $E_r(t)[\gamma^p|\delta^q]$ is a lower triangular matrix with 1's on the diagonal, and $\det(E_r(t)[\gamma^p|\gamma^q]) = 1$.

If $r \in \gamma^p$, then the rth row of $E_r(t)[\gamma^p|\star]$ contains two nonzero entries, a t in the $(r-1)$st column and a 1 in the rth column. If $i \neq r$, then the ith row of $E_r(t)[\gamma^p|\star]$ has only one nonzero entry, which is again a 1 in the $(\gamma^p)_i$th column. As in the previous case, if $(\gamma^p)_i < r-1$, then we are forced to choose $(\delta^q)_i = (\gamma^p)_i$. If $(\gamma^p)_i = r-1$, then $(\gamma^p)_{i+1}$ is automatically r, and in order for $\det(E_r(t)[\gamma^p|\gamma^q])$ to be nonzero we have to choose $(\gamma^q)_i = r-1$ and $(\gamma^q)_{i+1} = r$. If $r-1 \notin \gamma^p$ and $(\gamma^p)_i = r$, then we may choose $(\delta^q)_i$ to be either r or $r-1$, as both choices ensure a nonzero entry in the ith row of $E_r(t)[\gamma^p|\gamma^q]$. Whenever $(\gamma^p)_i > r$, we are again forced to choose $(\delta^q)_i = (\gamma^p)_i$ in order to have a nonzero entry in the ith row of $E_r(t)[\gamma^p|\gamma^q]$. So the two possible choices of a strictly increasing n-multiset δ^q resulting in a nonzero value for $\det(E_r(t)[\gamma^p|\gamma^q])$ are $\delta^q = \gamma^p$ or $\delta^q = (\gamma^p \setminus \{r\}) \cup \{r-1\}$. The first of these results in the determinant being 1, and for the latter the resulting determinant is t. This proves the lemma. □

The next lemma demonstrates an interesting property of the matrices $C_k(E_r(t)[\gamma|\delta])$ when $t > 0$ and $\gamma, \delta \in \bar{\Delta}$.

LEMMA 4.4. [illegible] $\gamma, \delta \in \bar{\Delta}$ [illegible] $\det C_k(E_r(t)[\gamma|\delta]) \neq 0$ [illegible] $C_k(E_r(t)[\gamma|\delta])$ [illegible]

[illegible]. By Lemma 4.3, the pth row of $C_k(E_r(t)[\gamma|\delta])$ contains at most two nonzero entries, and when it contains two nonzero entries, we must have $r \in \gamma^p$, $r-1 \notin \gamma^p$, $\delta^q = (\gamma^p \setminus \{r\}) \cup \{r-1\}$, and $\delta^{q+1} = \gamma^p$. Note that the two nonzero entries in the pth row must lie in adjacent positions. If $p = 1$, then, in the lexicographic order on n-multisets, we have $\delta^q < \delta^{q+1} = \gamma^1 < \gamma^\ell$ for all $1 \leq \ell \leq k$. Therefore, $\gamma^\ell \neq \delta^q, \delta^{q+1}$, so the qth and $(q+1)$st columns of $C_k(E_r(t)[\gamma|\delta])$ are all 0's after the first row is deleted. This implies that $\det C_k(E_r(t)[\gamma|\delta]) = 0$, a contradiction. Therefore, when $\det C_k(E_r(t)[\gamma|\delta]) \neq 0$, the first row of $C_k(E_r(t)[\gamma|\delta])$ contains only one nonzero entry. If this entry lies in the qth column, then we can apply induction to the matrix obtained by deleting the first row and qth column of $C_k(E_r(t)[\gamma|\delta])$, and the lemma follows. □

An easy consequence of the above lemma is that $C_k(E_r(t)[\gamma|\gamma]) = 1$ for all $\gamma \in \bar{\Delta}$, because $C_k(E_r(t)[\gamma|\gamma])$ is guaranteed to have 1's down the main diagonal.

We now investigate the possibility of applying our results about $\det C_k(E_r(t)[\gamma|\delta])$ for $\gamma, \delta \in \bar{\Delta}$ to the expression for $\det C_k(L[\star|\alpha])$, $\alpha \in \bar{\Delta}$, that results from repeatedly applying the generalized Cauchy–Binet expansion to the factors in the standard bidiagonal factorization of the lower triangular totally positive matrix L. This provides an expression for $\det C_k(L[\star|\alpha])$ as nonnegative linear combinations of products of the

form

$$\prod_{s=2}^{N-1}\left(\prod_{r=s}^{N}\det C_k(E_r(t_{s,r})[\gamma_{s,r}|\delta_{s,r}]\right), \tag{4.1}$$

where all $t_{s,r}>0$ and $\gamma_{s,r},\delta_{s,r}\in\bar{\Delta}$ are chosen so that (i) $\gamma_{2,2}=\star$; (ii) $\gamma_{s,r+1}=\delta_{s,r}$; (iii) $\gamma_{s+1,s+1}=\delta_{s,N}$; and (iv) $\delta_{N-1,N}=\alpha$.

For convenience, we will refer to those $\gamma_{s,r}$ for $r=s,\dots,N$ as being the elements of $\bar{\Delta}$ chosen in the sth round for (4.1). It follows from Lemma 4.3 that when (4.1) is nonzero, then every $\delta_{s,r}$ is either equal to $\gamma_{s,r}$ or is obtained from $\gamma_{s,r}$ by changing some of the r's in $\gamma_{s,r}$ to $r-1$'s. The first lemma gives a restriction on those $\gamma\in\bar{\Delta}$ that can occur in such a product when (4.1) is nonzero.

LEMMA 4.5. [illegible] $\gamma=\gamma_{s,r}\in\bar{\Delta}$ [illegible] (4.1) [illegible] [illegible] γ [illegible]

(i) [illegible] $m\in\{1,\dots,n\}$ [illegible] $\gamma_m=m$

(ii) [illegible] $m\in\{n+1,\dots,N\}$ [illegible] $\lceil\frac{m}{2}\rceil\le\gamma_m\le m$

[illegible]. It is easy to see that γ_m can be at most m. Since $\gamma_{2,2}=\star$ at the start, reducing any of the first n entries from r to $r-1$ would cause γ^1 to have a repeated index, thus forcing it out of Δ. This proves (i).

Now suppose $n+1\le m$. Then the index γ_m is obtained by reducing the mth position of $\star$ exactly $m-\gamma_m$ times. However, this index can be reduced only once in each round, at the factor involving $E_r(t_{s,r})$ when $r=(\gamma_{s,r})_m$. If we reduce as often as possible, the mth position can be reduced to $m-1$ in the 2-round, $m-2$ in the 3-round, $m-3$ in the 4-round, etc. Continuing in this manner, the first round in which the mth postion cannot be reduced is the sth round when the mth position is $m-(s-2)$ and this is less than s. This occurs when $m-(s-2)<s$ but $m-(s-3)\ge(s-1)$, so we have $2(s-2)\le m<2(s-1)$. If m is even and $m=2\ell$, then this forces $\ell=s-2$, so the mth position can be reduced by one in each of the rounds $2,3,\dots,s-1$, a total of $s-2=\ell$ times. This forces $\gamma_m\ge\ell=\frac{m}{2}=\lceil\frac{m}{2}\rceil$. If m is odd and $m=2\ell+1$, then this forces $\ell=s-2$, so the mth position can be reduced by one in each of the rounds $2,3,\dots,s-1$, a total of ℓ times, so again $\gamma_m\ge m-\ell=\lceil\frac{m}{2}\rceil$. □

The above lemma can be used to give another proof that $\det C_2(A)>0$ for all totally positive matrices A. In this case, every pair $\gamma,\delta\in\bar{\Delta}$ involved in a factor of the product (4.1) when it is nonzero satisfies $\gamma^1=\delta^1<\delta^2$. This implies that every factor in every nonzero product (4.1) is the determinant of a lower triangular matrix $C_2(E_r(t)[\gamma|\delta])$ with nonnegative entries, from which the result follows. In fact, it also follows that $\det C_2(A)\ge\det(A)$, since $\det(A)$ is the term of the generalized Cauchy–Binet expansion obtained when every $\gamma\in\bar{\Delta}$ is chosen to be $\star$.

For $k=3$, a similar approach also works.

THEOREM 4.6. [illegible] A [illegible] $3n\times 3n$ [illegible] $\det C_3(A)>0$

[illegible]. Again it suffices to show that every factor in the product (4.1) when it is nonzero is the determinant of the lower triangular matrix $C_3(E_r(t)[\gamma|\delta])$. From the proof of Lemma 4.4, we can see that the first row of $C_3(E_r(t)[\gamma|\delta])$ has only one nonzero entry, which has to be a 1 in the $(1,1)$-position because $\gamma^1=\delta^1=(1,2,\dots,n)$. So it suffices to show that the $(2,3)$-entry of $C_3(E_r(t)[\gamma|\delta])$ is always 0. Suppose that one of the $(2,3)$-entries of the $C_3(E_r(t)[\gamma|\delta])$ matrices involved in a nonzero term of the expansion of (4.1) is nonzero. In the first such occurrence, the γ that occurs must have an r in position $n+i$, with $1\le i\le 2n$. More precisely,

$$\gamma^2=(s_1,\dots,s_{i-1},r-1,t_2,u_1,\dots,u_{n-(i+1)}) \text{ and}$$
$$\gamma^3=(s_1,\dots,s_{i-1},r,t_1,v_1,\dots,v_{n-(i+1)})$$

with $s_1 < \cdots < s_{i-1} < r-1$, $r < t_1 < t_2 < u_1 < \cdots u_{n-(i+1)} \leq 2n$ and $t_1 < v_1 < \cdots < v_{n-(i+1)} \leq 3n$. We obtain δ from γ by replacing r by $r-1$ in γ^3, which reverses the lexicographic order of γ^2 and γ^3, making

$$\delta^2 = (s_1, \ldots, s_{i-1}, r-1, t_1, v_1, \ldots, v_{n-(i+1)}) \text{ and}$$
$$\delta^3 = \gamma^2.$$

(This reversing of the lexicographic order is needed to ensure that the (2,3)-entry of $C_3(E_r(t)[\gamma|\delta])$ nonzero.) Since $t_2 = \gamma_{n+i+1}$, we have that $t_2 \leq n+i+1$. If $i > 1$, then this forces $s_1 \leq n-1$. However, $s_1 = \gamma_{2n+1}$, so we must have $s_1 \geq n+1$ by Lemma 4.5, a contradiction. If $i = 1$, then $t_2 \leq n+2$, so $r \leq n$. But γ has an r in the $(2n+1)$-position, which again contradicts Lemma 4.5. This completes the proof. □

COROLLARY 4.7. [illegible] A [illegible] $3n \times 3n$ [illegible] $C_3(A)$ [illegible]

5. The case $k \geq 4$. We now verify by example that Theorem 4.6 does not extend to $k \geq 4$. Specifically, we show that there exists an 8×8 totally nonnegative matrix, in which the corresponding 4×4 compression consisting of 2×2 blocks is not totally nonnegative. It is evident how to extend this example for larger values of k by embedding this example into the upper left corner of a $2k \times 2k$ totally nonnegative matrix.

Consider the 8×8 matrix given by

$$A = \begin{bmatrix} 1 & t & t & 0 & 0 & 0 & 0 & 0 \\ t & t^2+1 & t^2+1 & 2t & 2t & t^2 & t^2 & 0 \\ t & t^2+1 & t^2+1 & 2t & 2t & t^2 & t^2 & 0 \\ t^2 & t^3+2t & t^3+2t & 1+4t^2 & 1+4t^2 & t(1+2t^2) & t(1+2t^2) & 0 \\ t^2 & t^3+2t & t^3+2t & 1+4t^2 & 1+4t^2 & t(1+2t^2) & t(1+2t^2) & 0 \\ 0 & t^2 & t^2 & 2t^3+2t & 2t^3+2t & 1+t(t^3+2t) & 1+t(t^3+2t) & t \\ 0 & t^2 & t^2 & 2t^3+2t & 2t^3+2t & 1+t(t^3+2t) & 1+t(t^3+2t) & t \\ 0 & 0 & 0 & t^2 & t^2 & t^3+t & t^3+t & t^2 \end{bmatrix}.$$

Then for any $t \geq 0$, the matrix A above is totally nonnegative. Furthermore, its compressed 4×4 matrix, namely, $C_4(A)$, is given by

$$C_4(A) = \begin{bmatrix} 1 & 2t^2 & 0 & 0 \\ t^2 & t^2+2t^4+1 & t^2 & 0 \\ t^4 & 2t^6+2t^4+3t^2 & 1+3t^4+4t^2 & t^2(1+2t^2) \\ 0 & t^4 & t^2+t^6+2t^4 & t^6+t^4 \end{bmatrix}.$$

It is not difficult to verify that

$$\det(C_4(A)) = t^8 - t^{12} + t^6,$$

which is negative for large t (even $t \geq \sqrt{2}$ suffices). A basic continuity argument can be applied to this example, using the fact that the TP matrices are dense in the TN matrices, to conclude that there exists an 8×8 TP matrix whose 4×4 compression matrix is not TP.

The above example represents another case of a generalized matrix function which can take negative values on the set of TN matrices (see also [16]).

6. Further results. Here we extend Corollaries 2.2 and 4.7 for $k = 2, 3$ by verifying that, in fact, the inequality $\det(C_k(A)) \geq \det(A)$ is strict if A is TP or, more generally, an oscillatory matrix. Recall [5, 6] that a matrix is called [illegible] if it is totally nonnegative and some positive integer power of it is totally positive. To proceed, we need additional background information.

As in the previous section we use an existence of a factorization of any invertible TN matrix into a product of totally nonnegative bidiagonal matrices of the form $E_r(t) = I + tE_{r,r-1}, E_{-r}(t) = E_r(t)^T$ and a positive diagonal factor.

An elementary bidiagonal factorization of an $N \times N$ invertible TN matrix A is not unique. Factorizations involving the minimal possible number of factors can be described as follows (see [5]). Each A belongs to one of the disjoint subsets $G_{u,v}$ of the set of invertible TN matrices uniquely parametrized by a pair of permutations $u, v \in S_n$. Let $u = (i_1, i_1 + 1) \cdots (i_l, i_l + 1)$ and $v = (j_1, j_1 + 1) \cdots (j_m, j_m + 1)$ be any pair of reduced (i.e., shortest possible) factorizations of u and v into products of elementary transpositions. Recall that l is called [illegible] of u and the ordered tuple $(i_1, \ldots, i_l)$ is called [illegible] for u. Further let $(k_1, \ldots, k_{l+m})$ be an arbitrary shuffle of ordered tuples $(-i_1, \ldots, -i_l)$ and $(j_1, \ldots, j_m)$. Then, for every $s \in \{1, \ldots, l + m\}$, there is a unique factorization of A of the form

$$A = E_{k_1}(t_1) \cdots E_{k_{s-1}}(t_{s-1}) D E_{k_s}(t_s) \cdots E_{k_{l+m}}(t_{l+m}), \tag{6.1}$$

where $t_1, \ldots, t_{l+m}$ are positive numbers and D is a positive diagonal matrix that does not depend on s. Furthermore, A is TP if and only if both u and v are elements of maximal length in S_N and A is oscillatory if and only if, for every $i \in \{1, \ldots, N-1\}$, indices $\pm i$ are present in the vector $(k_1, \ldots, k_{l+m})$ (cf. [5]). Recall that a permutation w is called a [illegible] of S_N if any reduced word for w contains exactly one copy of every index $i \in \{1, \ldots, N-1\}$.

One of the tools used in [2, 5] (see also [3]) is a graphical representation of the bidiagonal factorization in terms of planar diagrams (or networks). A [illegible] n is a planar acyclic digraph $\mathcal{D}$ with all edges oriented from left to right and $2n$ distinguished boundary vertices: n sources on the left and n sinks on the right with both sources and sinks labeled $1, \ldots, n$ from bottom to top. To each edge of a planar diagram $\mathcal{D}$ we assign a positive weight. We denote a collection of all assigned weights by W and call the pair $(\mathcal{D}, W)$ a [illegible] n. The weight of a path in $\mathcal{D}$ is defined as a product of weights assigned to edges that form this path.

Now, if $(\mathcal{D}, W)$ is an arbitrary weighted planar diagram of order n, then we define an $n \times n$ matrix $A = A(\mathcal{D}, W)$ by letting the (i, j)-entry of A be equal to the sum of the weights of all paths joining the vertex i on the left side of the obtained diagram $\mathcal{D}$ with the vertex j on the right side.

If we let $A = A(\mathcal{D}, W)$, then we can calculate any minor of A as follows (see [11], for example). For index sets $\alpha = \{i_1, i_2, \ldots, i_t\}$ and $\beta = \{j_1, j_2, \ldots, j_t\}$, consider a collection $P(\alpha, \beta)$ of all families of vertex-disjoint paths joining the vertices $\{i_1, i_2, \ldots, i_t\}$ on the left of the diagram $\mathcal{D}$ with the vertices $\{j_1, j_2, \ldots, j_t\}$ on the right. For $\pi \in P(\alpha, \beta)$, let $w(\pi)$ be the product of all the weights assigned to edges that form a family π. Then

$$\det A[\alpha|\beta] = \sum_{\pi \in P(\alpha,\beta)} w(\pi) \ .$$

In particular, $A(\mathcal{D}, W)$ is TN.

As an example, consider a planar diagram that corresponds to the factorization (6.1). It is obtained by concatenation, left to right, of diagrams that correspond to elementary factors. In the diagram representing D the only edges present are horizontal edges joining i on the left with i on the right ($i = 1, \dots, N$) with some positive weight assigned to each edge. In the diagram representing $E_r(\alpha)$ there is an additional edge from r to $r+1$ if r is positive and from $r+1$ to r otherwise. The weight of this edge is α, while the weights of all horizontal edges are equal to 1. The result looks like a collection of N horizontal lines with additional inclined edges between them directed either southeast or northeast.

Suppose now that $N = nk$ ($n \geq 2$) and A is a $N \times N$ TN matrix.

LEMMA 6.1. [illegible] $t > 0$ $l \in \{1, \dots, N-1\}$ [illegible] $k = 2, 3$

$$\det C_k(E_{\pm l}(t)A) \geq \det C_k(A), \quad \det C_k(AE_{\pm l}(t)) \geq \det C_k(A). \tag{6.2}$$

[illegible]. Define $B = E_l(t)A$. Then B is totally nonnegative. Moreover, B is obtained from A by replacing the lth row of A by itself plus t times row $l-1$. If both $l-1, l$ belong to some Δ_i, then evidently $C_k(B) = C_k(A)$. On the other hand, if $l \in \Delta_i$ and $l-1 \in \Delta_{i-1}$, then $C_k(B)$ and $C_k(A)$ are equal entrywise except in row i (which involves the interval Δ_i). Observe that any entry in row i of $C_k(B)$ is given by

$$\det B[\Delta_i|\Delta_j] = \det A[\Delta_i|\Delta_j] + t \det A'[\Delta_i|\Delta_j], \ j = 1, 2, \dots, k,$$

where A' is obtained from A by replacing row l with row $l-1$ (i.e., row $l-1$ is repeated in A). Clearly A' is TN. Hence by the linearity of the determinant it follows that

$$\det C_k(B) = \det C_k(A) + t \det C_k(A').$$

By Corollaries 2.2 or 4.7, $\det C_k(A') \geq 0$ and so

$$\det C_k(B) \geq \det C_k(A).$$

The remaining three inequalities in (6.2) can be proved in a similar way. □

Now suppose A is oscillatory and u, v is a pair of permutations that corresponds to A.

LEMMA 6.2. [illegible] u', v' [illegible] $A' \in G_{u',v'}$ [illegible] (6.1) [illegible] A [illegible] A' [illegible] $\det C_k(A) \geq \det C_k(A')$ [illegible] $k = 2, 3$

[illegible]. We will use induction on the length of u and v. If the length of both u and v is $N-1$, then both permutations are Coxeter and there is nothing to prove. If the length of v is at least N, we will show that v has a reduced factorization of the form $v = (i, i+1)v_1$ or $v = v_1(i, i+1)$, where any reduced word for v_1 contains every index $j \in \{1, \dots, N-1\}$. Then, by (6.1), A can be factored as $A = E_i(t)A_1$ (resp., $A = A_1E_i(t)$), where A_1 is oscillatory, has the same diagonal factor and belongs to G_{u,v_1} with the length of v_1 strictly less than the length of v. Moreover, by Lemma 6.1, $\det C_k(A) \geq \det C_k(A_1)$. Next, one can apply the same strategy to u and then use the induction hypothesis.

To obtain the required factorization of v, first recall that among reduced words for v there is one of the form $[m_1, n_1][m_2, n_2] \dots [m_s, n_s]$, where $m_i \leq n_i$ and $N-1 \geq n_1 > n_2 > \cdots > n_1 \geq 1$ (see, e.g., [17]). Since A is oscillatory, we have $n_1 = N-1$ and $\min\{m_1, \dots, m_s\} = 1$. Note that v is Coxeter if and only if $m_1 = n_2 + 1,\ m_2 = n_3 + 1, \dots,\ m_{s-1} = n_s + 1,\ m_s = 1$. If the length of v is greater than $N-1$, then

there exists $r \in \{1, s-1\}$ such that $m_1 = n_2 + 1,\ m_2 = n_3 + 1, \ldots,\ m_{r-1} = n_r + 1$, but $m_r \leq n_{r+1}$. Then

$$m_r < n_{r+1} + 1 \leq n_r = m_{r-1} - 1$$

and thus the transpositions $(m_r, m_r + 1)$ and $(m_{r-1}, m_{r-1} + 1)$ commute.

If $m_r > 1$, we can write $v = (m_r, m_r + 1)v_1$, where v_1 corresponds to the reduced word $[m_1, N-1] \ldots [m_{r-1}, n_{r-1}][m_r+1, n_r] \ldots [m_s, n_s]$. Otherwise, $v = v_1(n_s, n_s+1)$, where v_1 corresponds to the reduced word $[m_1, N-1] \ldots [m_{r-1}, n_{r-1}][1, n_r] \ldots [m_s, n_s - 1]$. In both cases, the reduced word for v_1 contains every index $j \in \{1, \ldots, N-1\}$ and we are done. □

LEMMA 6.3. [illegible] $A \in G_{u,v}$ [illegible] u, v [illegible] $\det C_k(A) > \det(A)$ [illegible] $k = 2, 3$

[illegible] Since u, v are Coxeter, their reduced words $(i_1, \ldots, i_{N-1})$ and $(j_1, \ldots, j_{N-1})$ are two permutations of indices $1, \ldots, N-1$. Consider a factorization (6.1) of A that corresponds to $s = 1$ and a shuffle $(-i_1, \ldots, -i_{N-1}, j_1, \ldots, j_{N-1})$. Let $(\mathcal{D}, W)$ be the weighted diagram defined by this factorization and let $D = \operatorname{diag}(d_1, \ldots, d_N)$ be the diagonal factor in the factorization. Note that for every $i \in \{1, \ldots, N-1\}$, $\mathcal{D}$ contains exactly one edge running southeast between $(i+1)$st and ith horizontal lines. This edge is followed by exactly one edge running northeast between ith and $(i+1)$st horizontal lines.

For $i, j \in \{1, \ldots, N\}$, consider the minor $\det A[\Delta_i|\Delta_j]$ We claim that for $i < j$, $\det A[\Delta_i|\Delta_j] = 0$. Indeed, since intervals Δ_i and Δ_j are disjoint, any collection of k nonintersecting paths from Δ_i on the left to Δ_j on the right in $\mathcal{D}$ must contain at least n northeast oriented edges between (in)th and $(in+1)$st horizontal lines. But only one such edge is contained in $\mathcal{D}$. Similarly, $\det A[\Delta_i|\Delta_j] = 0$ for $i > j$.

On the other hand, for $i > 1$, there exist at least two collections of nonintersecting paths from Δ_i to Δ_i: one that consists of n horizontal paths and another one consisting of $n-1$ horizontal paths from $(i-1)n+2$ to $(i-1)n+2, \ldots, in$ to in, and the path from $(i-1)n+1$ to $(i-1)n+1$ by going down on the left to $(i-1)n$ and then back up to $(i-1)n+1$ on the right. Thus, for $i > 1$, $\det A[\Delta_i] > d_{(i-1)n+1} \cdots d_{in}$. It is also easy to see that $\det A[\Delta_1] = d_1 \cdots d_n$.

It follows from the arguments above that

$$\begin{aligned} \det C_k(A) &= \prod_{i=1}^{k} \det A[\Delta_i] \\ &> \prod_{r=1}^{N} d_r = \det D = \det A. \end{aligned}$$ □

Combining the two preceeding Lemmas, we obtain the next result.

THEOREM 6.4. [illegible] $nk \times nk$ [illegible] A $n \geq 2$, $k = 2, 3$ $\det(C_k(A)) > \det(A)$

Our next generalization deals with compressions involving submatrices with overlapping row and column index sets.

For arbitrary N with $n < N$, let $I_1 = [i_1, i_1 + n - 1], I_2 = [i_2, i_2 + n - 1], \ldots, I_k = [i_k, i_k + n - 1]$ and $J_1 = [j_1, j_1 + n - 1], J_2 = [j_2, j_2 + n - 1], \ldots, J_k = [j_k, j_k + n - 1]$ be two collections of intervals of size n in $\{1, 2, \ldots, N\}$ such that $i_1 < \cdots < i_k$ and $j_1 < \cdots < j_k$. Suppose A is any $N \times N$ totally nonnegative matrix. Consider the "compressed matrix" $\hat{A} = (\det A[I_s|J_t])_{s,t=1}^{k}$, and let $d = \det \hat{A}$.

Let B_I be the $nk \times N$ totally nonnegative matrix represented by the planar diagram in which the only edges are those going from 1 to $i_1, \ldots, n$ to $i_1 + n - 1$; $n+1$

to i_2, ..., $2n$ to $i_2 + n - 1$; ...; $(k-1)n + 1$ to i_k, ..., nk to $i_k + n - 1$, and all edge weights are equal to one. Let B_J be constructed in a similar manner for $J_1, \ldots, J_k$. Define $A' = B_I A B_J^T$. Then A' is totally nonnegative and

$$\det A[I_s|J_t] = \det A'[\Delta_s|\Delta_t].$$

Thus it follows that

$$d = \det \hat{A} = \det C_k(A') \geq 0.$$

Consequently, the matrix $\hat{A} = (\det A[I_s|J_t])$ for $s, t = 1, 2, \ldots, k$, is totally nonnegative whenever A is totally nonnegative. We summarize the above analysis in the next theorem.

THEOREM 6.5. [illegible] n, N $n < N$ [illegible] $k = 2, 3$ [illegible] $I_1, I_2, \ldots, I_k$ $J_1, J_2, \ldots, J_k$ [illegible] n [illegible] $\{1, 2, \ldots, N\}$ [illegible] $I_1, I_2, \ldots, I_k$ (. [illegible] $J_1, J_2, \ldots, J_k$) [illegible] A [illegible] $N \times N$ [illegible] $\hat{A} = (\det A[I_s|J_t])_{s,t=1}^k$ [illegible] $\hat{A}$ [illegible]

Acknowledgment. We thank the referees of a previous version of the paper for their helpful comments and suggestions and, especially, for pointing out a gap in our earlier argument. The rethinking of that argument led us to the discovery of the example presented in section 5.

REFERENCES

[1] T. ANDO, *Totally positive matrices*, Linear Algebra Appl., 90 (1987), pp. 165–219.
[2] A. BERENSTEIN, S. FOMIN, AND A. ZELEVINSKY, *Parameterizations of canonical bases and totally positive matrices*, Adv. Math., 122 (1996), pp. 49–149.
[3] F. BRENTI, *Combinatorics and total positivity*, J. Combin. Theory Ser. A, 71 (1995), pp. 175–218.
[4] C. W. CRYER, *Some properties of totally positive matrices*, Linear Algebra Appl., 15 (1976), pp. 1–25.
[5] S. FOMIN AND A. ZELEVINSKY, *Total positivity: Tests and parameterizations*, Math. Intelligencer, 22 (2000), pp. 23–33.
[6] F. R. GANTMACHER AND M. G. KREIN, *Oscillation Matrices and Kernels and Small Vibrations of Mechanical Systems*, AMS, Providence, RI, 2002.
[7] M. GASCA AND C. A. MICCHELLI, *Total Positivity and Its Applications*, Math. Appl. 359, Kluwer, Dordrecht, The Netherlands, 1996.
[8] M. GASCA AND J. M. PEÑA, *On factorizations of totally positive matrices*, in Total Positivity and Its Applications, Math. Appl. 359, Kluwer, Dordrecht, the Netherlands, 1996, pp. 109–130.
[9] C. R. JOHNSON AND T. L. MARKHAM, *Compression and Hadamard inequalities*, Linear Multilinear Algebra, 18 (1985), pp. 23–34.
[10] S. KARLIN, *Total Positivity*, Vol. I, Stanford University Press, Stanford, CA, 1968.
[11] S. KARLIN AND G. MCGREGOR, *Coincidence probabilities*, Pacific J. Math., 9 (1959), pp. 1141–1164.
[12] C. LOEWNER, *On Totally Positive Matrices*, Math. Z., 63 (1955), pp. 338–340.
[13] M. MARCUS AND H. MINC, *Generalized matrix functions*, Trans. Amer. Math. Soc., 116 (1965), pp. 316–329.
[14] R. MERRIS, *Multilinear Algebra*, Gordan and Breach, Amsterdam, 1997.
[15] G. MUHLBACH AND M. GASCA, *A generalization of Sylvester's identity on determinants and some applications*, Linear Algebra Appl., 66 (1985), pp. 221-234.
[16] J. R. STEMBRIDGE, *Immanants of totally positive matrices are nonnegative*, Bull. London Math. Soc., 23 (1991), pp. 422–428.
[17] J. R. STEMBRIDGE, *Some combinatorial aspects of reduced words in finite Coxeter groups*, Trans. Amer. Math. Soc., 349 (1997), pp. 1285–1332.
[18] R. C. THOMPSON, *A determinantal inequality for positive definite matrices*, Canad. Math. Bull., 4 (1961), pp. 57–62.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 1, pp. 81–88

PERTURBATION OF MATRICES DIAGONALIZABLE UNDER CONGRUENCE*

SUSANA FURTADO† AND CHARLES R. JOHNSON‡

Abstract. A matrix $A \in M_n(\mathbb{C})$ is called unitoid if it is congruent to a diagonal matrix. Necessary and sufficient conditions are given on the canonical angles of a unitoid matrix so that sufficiently small perturbations remain unitoid. This, in particular, resolves the question of when simultaneous diagonalizability of two Hermitian matrices is retained under perturbation.

Key words. canonical angles, congruence, perturbation, simple canonical lines, unitoid

AMS subject classifications. 15A04, 15A21, 15A60

DOI. 10.1137/S089547980343775X

1. Introduction. Matrices A and $B \in M_n(\mathbb{C})$, M_n for short, are said to be if there is a nonsingular $C \in M_n$ such that $B = C^*AC$. Of course, congruence is an equivalence relation on M_n and, in addition to the classical motivation of change of variables in a quadratic form, it arises in many ways, such as study of the algebraic Riccati equation and indefinite scalar products [GLR]. We note also that our main result shows which pairs of Hermitian matrices remain simultaneously diagonalizable under congruence, a phenomenon important in several applications, including mechanics, computation, and control.

In [JF], a matrix that is diagonalizable by congruence is called ; by use of an auxiliary diagonal congruence, $A \in M_n$ is nonsingular and unitoid if and only if it is congruent to a diagonal unitary matrix, and a general unitoid matrix is necessarily congruent to a direct sum of a zero matrix and a diagonal unitary matrix (the "nonsingular part" of the diagonal form). The arguments of the nonzero diagonal entries of a diagonal matrix D to which a unitoid matrix A is congruent are a congruential invariant. Thus, these angles are called for A; the zero eigenvalues of D are referred to as . The canonical angles play an important role in the understanding of properties of unitoid matrices. The lines through the origin of the complex plane along which canonical angles (for A) lie are called (for A). Note that canonical angles may be multiple and that canonical angles may lie in both directions (from the origin) along a canonical line. Whether the latter occurs (canonical angles that differ by π) is quite important. We call a canonical line for the unitoid matrix A if it has canonical angles in only one direction; otherwise the line is . A nonsingular unitoid matrix A is called if all its canonical lines are simple; otherwise, it is called . A nonsingular $A \in M_n$ is unitoid if and only if $A^{-1}A^*$ is similar to a unitary matrix [DJ, FJ2]. The canonical lines of the nonsingular A are determined by the spectrum of $A^{-1}A^*$, although the canonical angles are not. In particular, canonical angles that lie on the same canonical line lead to multiple eigenvalues of $A^{-1}A^*$.

*Received by the editors November 24, 2003; accepted for publication (in revised form) by R. Nabben September 20, 2005; published electronically March 17, 2006.
http://www.siam.org/journals/simax/28-1/43775.html

†Faculdade de Economia, Rua Dr. Roberto Frias, 4200-464 Porto, Portugal (sbf@fep.up.pt). This work was done within the activities of Centro de Estruturas Lineares e Combinatórias da Universidade de Lisboa.

‡Department of Mathematics, College of William and Mary, Williamsburg, VA 23187-8795 (crjohnso@math.wm.edu).

Our interests here lay in describing a perturbation theory for unitoid matrices, and the remarkable structure of it is rather surprising compared to classical eigenvalue perturbation theory. In classical eigenvalue perturbation theory, small changes in a matrix with multiple eigenvalues can lead to a change in Jordan structure and sufficiently small perturbations of a matrix with distinct eigenvalues remain diagonalizable. We shall see that the situation for unitoid matrices is quite different. Small perturbations of a simple unitoid matrix with multiple canonical angles remain unitoid and small perturbations of a nonsimple unitoid matrix can lead to a nonunitoid one, even if all canonical angles have multiplicity one. Also, perturbations of a nonsingular A give very structured perturbations of $A^{-1}A^*$. For example, if $A^{-1}A^*$ is similar to a unitary matrix, even with repeated eigenvalues, small perturbations B of A almost always leave $B^{-1}B^*$ similar to a unitary.

Though it is not central to our theoretical development, it is relevant to applications that it is not difficult to see that $A \in M_n$ is unitoid if and only if two certain Hermitian matrices are simultaneously diagonalizable by congruence, namely, $H(A) = \frac{1}{2}(A + A^*)$ and $\frac{1}{i}S(A) = \frac{1}{2i}(A - A^*)$. Since these may be any two Hermitian matrices, a corollary to our results will be the identification of pairs of Hermitian matrices, arbitrarily small independent Hermitian perturbations of which will be simultaneously diagonalizable by congruence. In addition, perturbation of unitoid matrices is relevant to perturbation of Hermitian matrix pencils [T1, T2].

Recall that the [illegible] of $A \in M_n$ is

$$F(A) \equiv \{x^*Ax : x \in \mathbb{C}^n, x^*x = 1\},$$

a compact convex subset of the complex plane [HJ]. The location of 0 relative to $F(A)$ (in the interior, on the boundary or outside) is invariant under congruence.

It is known [DJ] that a matrix A for which $0 \notin F(A)$ is necessarily unitoid. Thus, the perturbation of such matrices is a special case of the results herein.

Given $A \in M_n$ we denote the spectrum of A by $\sigma(A)$. If A is nonsingular, we denote $A^{-1}A^*$ by $\Phi(A)$.

2. Preliminary results. We first need to identify a fact that follows immediately from Theorem 1 and Lemmas 3, 4, and 5 in [FJ2] and will be useful in the technical part of our development. These results in [FJ2] have been proved using the simultaneous canonical form for a pair of Hermitian matrices which may be found, for example, in [T1, T2].

LEMMA 1. [illegible] $A \in M_n$ [illegible] $C \in M_n$ [illegible] $C^*AC = E_1 \oplus \cdots \oplus E_{n_1} \oplus F_1 \oplus \cdots \oplus F_{n_2}$, [illegible] $E_i \in M_{p_i}$, $F_j \in M_{2q_j}$, $\Phi(E_i)$ [illegible] $i = 1, \ldots, n_1$, [illegible] $\Phi(F_j)$ [illegible] q_j [illegible] $\lambda_j e^{i\theta_j}$ $\frac{1}{\lambda_j}e^{i\theta_j}$ [illegible] $\lambda_j > 1$, $j = 1, \ldots, n_2$.

Many links between congruential structure of A and similarity structure of $\Phi(A)$ have been recognized, beginning with [DJ]. Our next lemma, crucial for our later results, more precisely relates reducibility in one setting to that in the other. Restrictions are necessary, as the extent to which similarity reduction of $\Phi(A)$ translates into congruential reduction of A is limited.

LEMMA 2. [illegible] $A \in M_n$ [illegible] $P \in M_n$ [illegible] $P^{-1}\Phi(A)P = Q_1 \oplus Q_2$ [illegible] $Q_1 \in M_r$, $Q_2 \in M_{n-r}$, [illegible] $\lambda e^{i\theta} \in \sigma(Q_1)$, $\lambda > 0$, [illegible] $\frac{1}{\lambda}e^{i\theta} \notin \sigma(Q_2)$. [illegible] $P^*AP = A_1 \oplus A_2$, [illegible] $A_1 \in M_r$ [illegible] $A_2 \in M_{n-r}$ [illegible] $\Phi(A_1) = Q_1$ [illegible] $\Phi(A_2) = Q_2$.

. It follows from Lemma 1 that there is a nonsingular $C \in M_n$ such that $C^*AC = B_1 \oplus \cdots \oplus B_m$, in which, for $j = 1, \ldots, m$, B_j is such that either $\Phi(B_j)$ is similar to a Jordan block associated with an eigenvalue on the unit circle or $\Phi(B_j)$ is similar to a direct sum of two Jordan blocks of the same size associated with eigenvalues $\lambda_j e^{i\theta_j}$, $\frac{1}{\lambda_j} e^{i\theta_j}$ for some $\lambda_j > 1$. Since $\Phi(C^*AC) = C^{-1}\Phi(A)C$ is similar to $Q_1 \oplus Q_2$, and because of our assumption on the spectra of Q_1 and Q_2, we have $\sigma(Q_1) \cap \sigma(Q_2) = \emptyset$ and we may suppose, without loss of generality, that $\sigma(\Phi(B_1 \oplus \cdots \oplus B_k)) = \sigma(Q_1)$, $1 \leq k \leq m$. Let $X_1 = B_1 \oplus \cdots \oplus B_k$ and $X_2 = B_{k+1} \oplus \cdots \oplus B_m$. Then $\Phi(C^*AC) = \Phi(X_1) \oplus \Phi(X_2)$ is similar to $Q_1 \oplus Q_2$. Moreover, since $\sigma(\Phi(X_1)) = \sigma(Q_1)$, $\sigma(\Phi(X_2)) = \sigma(Q_2)$, and $\sigma(Q_1) \cap \sigma(Q_2) = \emptyset$, there are $P_1 \in M_r$ and $P_2 \in M_{n-r}$ nonsingular such that $P_1^{-1}\Phi(X_1)P_1 = Q_1$ and $P_2^{-1}\Phi(X_2)P_2 = Q_2$. Then

$$(P_1 \oplus P_2)^{-1} C^{-1}\Phi(A)C\,(P_1 \oplus P_2) = Q_1 \oplus Q_2,$$

or, equivalently,

$$\left[(P_1 \oplus P_2)^{-1} C^{-1} P\right] [Q_1 \oplus Q_2] \left[P^{-1} C\,(P_1 \oplus P_2)\right] = Q_1 \oplus Q_2.$$

Because $\sigma(Q_1) \cap \sigma(Q_2) = \emptyset$, it follows easily from [HJ, Theorem 4.4.6] that $P^{-1}C(P_1 \oplus P_2) = R_1 \oplus R_2$ for some nonsingular $R_1 \in M_r$ and $R_2 \in M_{n-r}$. Thus, $P = C\,(P_1 \oplus P_2)\left(R_1^{-1} \oplus R_2^{-1}\right)$ and

$$P^*AP = \left(\left(R_1^{-1}\right)^* P_1^* X_1 P_1 R_1^{-1}\right) \oplus \left(\left(R_2^{-1}\right)^* P_2^* X_2 P_2 R_2^{-1}\right).$$

Since, by hypothesis, $\Phi(P^*AP) = P^{-1}\Phi(A)P = Q_1 \oplus Q_2$, it follows that $\Phi\big(\left(R_1^{-1}\right)^* P_1^* X_1 P_1 R_1^{-1}\big) = Q_1$ and $\Phi\big(\left(R_2^{-1}\right)^* P_2^* X_2 P_2 R_2^{-1}\big)$, completing the proof. □

The following examples show that without the restrictions imposed on the spectra of Q_1 and Q_2, Lemma 2 is not generally true.

. 1. Let

$$A = \begin{bmatrix} 1 & 3 \\ 0 & 1 \end{bmatrix}.$$

Then $\Phi(A)$ has eigenvalues $\lambda_1 = -\frac{7}{2} + \frac{3}{2}\sqrt{5}$ and $\lambda_2 = 1/\lambda_1$. The matrix $P^{-1}\Phi(A)P$ is diagonal with

$$P = \begin{bmatrix} -\frac{3}{2} + \frac{1}{2}\sqrt{5} & -\frac{3}{2} - \frac{1}{2}\sqrt{5} \\ 1 & 1 \end{bmatrix}.$$

However, P^*AP is not diagonal.

. 2. If $A \in M_n$, $n > 1$, is a nonsingular Hermitian matrix, all nonsingular $P \in M_n$ diagonalize $\Phi(A) = I_n$ by similarity. However, for almost all P, P^*AP is not diagonal.

Although it is not relevant for our work here, we now state a consequence of Lemma 2 concerning reducibility of unitoid matrices. We say that $B \in M_n$ is rotationally Hermitian if $B = e^{i\theta}H$ for some $0 \leq \theta < 2\pi$ and $H \in M_n$ Hermitian.

Corollary 3. $A \in M_n$ $C \in M_n$ $C^{-1}\Phi(A)C$ C^*AC A C^*AC

. Because $\Phi(A)$ is similar to a unitary matrix, by a possible unitary similarity via a permutation matrix, suppose, without loss of generality, that $C^{-1}\Phi(A)C = Q_1 \oplus \cdots \oplus Q_m$ with $Q_j = e^{-2i\gamma_j} I_{k_j}$, $\gamma_j \in \mathbb{R}$, $j = 1, \ldots, m$, and $e^{-2i\gamma_{j_1}} \neq e^{-2i\gamma_{j_2}}$ for $j_1 \neq j_2$. It follows easily from Lemma 2 that $C^*AC = A_1 \oplus \cdots \oplus A_m$ for some $A_j \in M_{k_j}$ such that $\Phi(A_j) = Q_j$, $j = 1, , \ldots, m$. Clearly, because $\Phi(A_j)$ is unitary, A_j is unitoid. Let A'_j be a diagonal unitary matrix congruent to A_j. Since $\Phi(A'_j)$ is similar to Q_j, then A'_j is unitarily similar (via a permutation matrix) to a matrix of the form $e^{i\gamma_j}\left(-I_{k_j-r_j} \oplus I_{r_j}\right)$ with $0 \leq r_j \leq k_j$. Therefore, A'_j (and, thus, A_j) is rotationally Hermitian, corresponding to the canonical line for A on which the angle γ_j lies. In case A is simple with distinct canonical angles, then $k_j = 1$, $j = 1, , \ldots, m$, completing the proof. □

As a practical matter in what follows we take $\|\!| \cdot |\!\|$ to be the spectral norm. However, any unitary similarity invariant submultiplicative matrix norm for which $\|\!|C|\!\| = \|\!|C^*|\!\|$ will do (e.g., the Frobenius norm).

For $A \in M_{n,m}$ and $\varepsilon > 0$, we denote by $\mathcal{V}_\varepsilon(A)$ the set

$$\{B \in M_{n,m} : \|\!|B - A|\!\| < \varepsilon\}.$$

LEMMA 4. $A \in M_{k_1}$ $B \in M_{k_2}$ $\sigma(A) \cap \sigma(B) = \emptyset$ $\varepsilon > 0$, $A_\varepsilon \in \mathcal{V}_\varepsilon(A)$, $B_\varepsilon \in \mathcal{V}_\varepsilon(B)$ $\Sigma_\varepsilon \in \mathcal{V}_\varepsilon(0_{k_1,k_2})$, $R_\varepsilon \in M_{k_1,k_2}$ $A_\varepsilon R_\varepsilon - R_\varepsilon B_\varepsilon = \Sigma_\varepsilon$. $R_\varepsilon \to 0$ $\varepsilon \to 0$.

. The uniqueness of the solution R_ε follows from [HJ, Theorem 4.4.6] because, by continuity, for sufficiently small $\varepsilon > 0$, $\sigma(A_\varepsilon) \cap \sigma(B_\varepsilon) = \emptyset$. Since for each ε, $A_\varepsilon R_\varepsilon - R_\varepsilon B_\varepsilon = \Sigma_\varepsilon$ is a linear system in the entries of R_ε, then the unique solution R_ε depends continuously on A_ε, B_ε and Σ_ε. Because $AR - RB = 0$ has the unique solution $R = 0$, it follows that $R_\varepsilon \to 0$ as $\varepsilon \to 0$. □

LEMMA 5. $Q = Q_1 \oplus Q_2$ $Q_1 \in M_r$ $Q_2 \in M_{n-r}$ $\sigma(Q_1) \cap \sigma(Q_2) = \emptyset$ $\varepsilon > 0$, $Q_\varepsilon \in \mathcal{V}_\varepsilon(Q)$

$$U_\varepsilon = \begin{bmatrix} U_{11}^\varepsilon & U_{12}^\varepsilon \\ U_{21}^\varepsilon & U_{22}^\varepsilon \end{bmatrix} \in M_n,$$

$U_{11}^\varepsilon \in M_r$, $U_\varepsilon^* Q_\varepsilon U_\varepsilon$ $\varepsilon \to 0$, $U_{12}^\varepsilon \to 0$, $U_{21}^\varepsilon \to 0$ $U_\varepsilon^* Q_\varepsilon U_\varepsilon \to Q$.

. For $\varepsilon > 0$, let $Q_\varepsilon \in \mathcal{V}_\varepsilon(Q)$. By continuity, the eigenvalues of Q_ε approach the eigenvalues of Q, as $\varepsilon \to 0$. Using Schur's unitary triangularization theorem, for each ε there is a unitary matrix $U_\varepsilon \in M_n$ such that

$$U_\varepsilon^* Q_\varepsilon U_\varepsilon = \begin{bmatrix} Q_1^\varepsilon & \Sigma_\varepsilon \\ 0 & Q_2^\varepsilon \end{bmatrix}, \tag{1}$$

in which $Q_1^\varepsilon \in M_r$ and $Q_2^\varepsilon \in M_{n-r}$ are upper triangular and the main diagonal of Q_1^ε (respectively, Q_2^ε) approaches the main diagonal of Q_1 (respectively, Q_2) as $\varepsilon \to 0$. Because $\|\!|U_\varepsilon^* Q_\varepsilon U_\varepsilon|\!\| = \|\!|Q_\varepsilon|\!\| \to \|\!|Q|\!\|$ and Q is diagonal, then $\Sigma_\varepsilon \to 0$, $Q_1^\varepsilon \to Q_1$ and $Q_2^\varepsilon \to Q_2$, as $\varepsilon \to 0$. Equality (1) is equivalent to

$$Q_\varepsilon U_\varepsilon = U_\varepsilon \begin{bmatrix} Q_1^\varepsilon & \Sigma_\varepsilon \\ 0 & Q_2^\varepsilon \end{bmatrix},$$

which, for

$$Q_\varepsilon = \begin{bmatrix} Q_{11}^\varepsilon & Q_{12}^\varepsilon \\ Q_{21}^\varepsilon & Q_{22}^\varepsilon \end{bmatrix} \text{ and } U_\varepsilon = \begin{bmatrix} U_{11}^\varepsilon & U_{12}^\varepsilon \\ U_{21}^\varepsilon & U_{22}^\varepsilon \end{bmatrix},$$

$Q_{11}^\varepsilon, U_{11}^\varepsilon \in M_r$, implies

$$Q_{11}^\varepsilon U_{12}^\varepsilon - U_{12}^\varepsilon Q_2^\varepsilon = U_{11}^\varepsilon \Sigma_\varepsilon - Q_{12}^\varepsilon U_{22}^\varepsilon, \tag{2}$$

$$Q_{22}^\varepsilon U_{21}^\varepsilon - U_{21}^\varepsilon Q_1^\varepsilon = -Q_{21}^\varepsilon U_{11}^\varepsilon. \tag{3}$$

Since, as $\varepsilon \to 0$, $Q_{21}^\varepsilon U_{11}^\varepsilon \to 0$ (note that $Q_{21}^\varepsilon \to 0$ and, because $U_\varepsilon^* U_\varepsilon = I_n$, the norms of U_{11}^ε and U_{22}^ε are bounded) and, for sufficiently small ε, $\sigma(Q_{22}^\varepsilon) \cap \sigma(Q_1^\varepsilon) = \emptyset$, it follows from (3) and Lemma 4 that $U_{21}^\varepsilon \to 0$. Analogously, from (2) it follows that $U_{12}^\varepsilon \to 0$. □

Our first key perturbation result is a technical prelude to our main results and indicates the importance of canonical lines.

THEOREM 6. [illegible] $A \in M_n$ [illegible] $A = A_1 \oplus \cdots \oplus A_m$, [illegible] $m \geq 2$, $A_i \in M_{k_i}$ [illegible] $i = 1, \ldots, m$, [illegible] A_i [illegible] A_j, $i \neq j$, [illegible] $\varepsilon > 0$, [illegible] $A_\varepsilon \in \mathcal{V}_\varepsilon(A)$. [illegible] ε, [illegible] $C_\varepsilon \in M_n$ [illegible] $C_\varepsilon^* A_\varepsilon C_\varepsilon = A_1^\varepsilon \oplus \cdots \oplus A_m^\varepsilon$ [illegible] $\varepsilon \to 0$, $A_i^\varepsilon \to A_i$, $i = 1, \ldots, m$, [illegible] $C_\varepsilon \to I_n$.

[illegible]. First, note that each block $\Phi(A_i)$ is $\lambda_i I_{k_i}$, for some complex number λ_i on the unit circle, and $\lambda_i \neq \lambda_j$ for $i \neq j$. For $\varepsilon > 0$, let $A_\varepsilon \in \mathcal{V}_\varepsilon(A)$. The proof is by induction on m. Let $D = A_2 \oplus \cdots \oplus A_m$. By the continuity of $\Phi(\cdot)$, $\Phi(A_\varepsilon) \to \Phi(A)$, as $\varepsilon \to 0$. Bearing in mind Lemma 5, for each ε, there is a unitary matrix

$$U_\varepsilon = \begin{bmatrix} U_{11}^\varepsilon & U_{12}^\varepsilon \\ U_{21}^\varepsilon & U_{22}^\varepsilon \end{bmatrix} \in M_n,$$

$U_{11}^\varepsilon \in M_{k_1}$, such that

$$U_\varepsilon^* \Phi(A_\varepsilon) U_\varepsilon = \begin{bmatrix} Q_1^\varepsilon & \Sigma_\varepsilon \\ 0 & Q_2^\varepsilon \end{bmatrix},$$

with $Q_1^\varepsilon \in M_{k_1}$ and $Q_2^\varepsilon \in M_{n-k_1}$ upper triangular matrices, and, as $\varepsilon \to 0$, $Q_1^\varepsilon \to \Phi(A_1)$, $Q_2^\varepsilon \to \Phi(D)$, $\Sigma_\varepsilon \to 0$, $U_{12}^\varepsilon \to 0$, and $U_{21}^\varepsilon \to 0$. According to Lemma 4, for sufficiently small ε, because $\sigma(Q_1^\varepsilon) \cap \sigma(Q_2^\varepsilon) = \emptyset$, there is $R_\varepsilon \in M_{k_1, n-k_1}$ such that

$$Q_1^\varepsilon R_\varepsilon - R_\varepsilon Q_2^\varepsilon = -\Sigma_\varepsilon,$$

and $R_\varepsilon \to 0$ as $\varepsilon \to 0$. Let

$$B_\varepsilon = U_\varepsilon \begin{bmatrix} I_{k_1} & R_\varepsilon \\ 0 & I_{n-k_1} \end{bmatrix}.$$

Then

$$B_\varepsilon^{-1} \Phi(A_\varepsilon) B_\varepsilon = \begin{bmatrix} Q_1^\varepsilon & 0 \\ 0 & Q_2^\varepsilon \end{bmatrix}.$$

Because $Q_1^\varepsilon \to \Phi(A_1)$, $Q_2^\varepsilon \to \Phi(D)$, the arguments of the eigenvalues of $\Phi(A_1)$ and $\Phi(D)$ are on the unit circle and $\sigma(\Phi(A_1)) \cap \sigma(\Phi(D)) = \emptyset$, for sufficiently small ε, if $\lambda e^{i\theta} \in \sigma(Q_1^\varepsilon)$, $\lambda > 0$, then $\frac{1}{\lambda} e^{i\theta} \notin \sigma(Q_2^\varepsilon)$. Therefore, by Lemma 2, for sufficiently small ε,

$$B_\varepsilon^* A_\varepsilon B_\varepsilon = \begin{bmatrix} Z_1^\varepsilon & 0 \\ 0 & Z_2^\varepsilon \end{bmatrix},$$

for some $Z_1^\varepsilon \in M_{k_1}$ and $Z_2^\varepsilon \in M_{n-k_1}$ such that $\Phi(Z_1^\varepsilon) = Q_1^\varepsilon$ and $\Phi(Z_2^\varepsilon) = Q_2^\varepsilon$. Let $P_\varepsilon = B_\varepsilon((U_{11}^\varepsilon)^* \oplus (U_{22}^\varepsilon)^*)$. Then

$$P_\varepsilon^* A_\varepsilon P_\varepsilon = \begin{bmatrix} A_1^\varepsilon & 0 \\ 0 & Y_2^\varepsilon \end{bmatrix}$$

with $A_1^\varepsilon = U_{11}^\varepsilon Z_1^\varepsilon (U_{11}^\varepsilon)^*$ and $Y_2^\varepsilon = U_{22}^\varepsilon Z_2^\varepsilon (U_{22}^\varepsilon)^*$. A calculation shows that $P_\varepsilon \to I_n$, $A_1^\varepsilon \to A_1$, and $Y_2^\varepsilon \to D$, as $\varepsilon \to 0$. If $m = 2$ the proof is complete, with $A_2^\varepsilon = Y_2^\varepsilon$. Now suppose that $m > 2$. By the induction hypothesis, for sufficiently small ε, there is a nonsingular $F_\varepsilon \in M_{n-k_1}$ such that $F_\varepsilon^* Y_2^\varepsilon F_\varepsilon = A_2^\varepsilon \oplus \cdots \oplus A_m^\varepsilon$, with $A_i^\varepsilon \to A_i$, $i = 2, \ldots, m$, and $F_\varepsilon \to I_{n-k_1}$, as $\varepsilon \to 0$. Then $P_\varepsilon (I_{k_1} \oplus F_\varepsilon) \to I_n$ and $[P_\varepsilon (I_{k_1} \oplus F_\varepsilon)]^* A_\varepsilon [P_\varepsilon (I_{k_1} \oplus F_\varepsilon)] = A_1^\varepsilon \oplus \cdots \oplus A_m^\varepsilon$, completing the proof. ∎

Thus far our results imagine perturbing a unitoid matrix in diagonal form. The following observation facilitates a natural transition.

PROPOSITION 7. [illegible] $A \in M_n$, $B = C^* A C$ [illegible] $C \in M_n$, [illegible] $\varepsilon > 0$. [illegible] $\delta > 0$ [illegible] $B_\delta \in \mathcal{V}_\delta(B)$, $(C^*)^{-1} B_\delta C^{-1} \in \mathcal{V}_\varepsilon(A)$.

[illegible] For $\varepsilon > 0$ let $\delta = \varepsilon / \|C^{-1}\|^2$ and $B_\delta \in \mathcal{V}_\delta(B)$. We have

$$\| (C^*)^{-1} B_\delta C^{-1} - A \| \leq \|C^{-1}\|^2 \|B_\delta - B\| \leq \|C^{-1}\|^2 \delta = \varepsilon. \quad ∎$$

3. Perturbation of simple unitoid matrices. We first investigate the important special case in which $A \in M_n$ is such that $0 \notin F(A)$. This includes, in particular, the case A is simple with just one canonical line (A is congruent to a scalar matrix).

LEMMA 8. [illegible] $A \in M_n$ [illegible] $0 \notin F(A)$. [illegible] $\varepsilon > 0$ [illegible] $A_\varepsilon \in \mathcal{V}_\varepsilon(A)$ [illegible] A_ε [illegible] A [illegible] $\varepsilon \to 0$.

[illegible] Since $0 \notin F(A)$, there is $\varepsilon > 0$ such that $0 \notin F(A_\varepsilon)$ for every $A_\varepsilon \in \mathcal{V}_\varepsilon(A)$. According to [DJ], each A_ε is unitoid. Suppose that A is congruent to $\operatorname{diag}(e^{i\theta_1}, \ldots, e^{i\theta_n})$ and A_ε is congruent to $\operatorname{diag}(e^{i\theta_1^\varepsilon}, \ldots, e^{i\theta_n^\varepsilon})$. Then the eigenvalues of $\Phi(A)$ and $\Phi(A_\varepsilon)$ are $e^{-2i\theta_1}, \ldots, e^{-2i\theta_n}$ and $e^{-2i\theta_1^\varepsilon}, \ldots, e^{-2i\theta_n^\varepsilon}$, respectively. By continuity, as $\varepsilon \to 0$, the eigenvalues of $\Phi(A_\varepsilon)$ approach the eigenvalues of $\Phi(A)$. Thus, suppose, without loss of generality, that

$$e^{-2i\theta_j^\varepsilon} \to e^{-2i\theta_j}, \ j = 1, \ldots, n. \tag{4}$$

Because there is an open half-plane determined by a line through the origin in which, for sufficiently small ε, the fields of values (and, thus, the canonical angles) of both A and A_ε lie, then (4) implies that

$$e^{i\theta_j^\varepsilon} \to e^{i\theta_j}, \ j = 1, \ldots, n,$$

completing the proof. ∎

The next theorem is the main result of this section.

THEOREM 9. [illegible] $A \in M_n$ [illegible] $\varepsilon > 0$ [illegible] $A_\varepsilon \in \mathcal{V}_\varepsilon(A)$ [illegible] A_ε [illegible] A [illegible] $\varepsilon \to 0$.

[illegible] Because of Proposition 7, suppose, without loss of generality, that $A = A_1 \oplus \cdots \oplus A_m$, in which $A_j \in M_{k_j}$ is diagonal with just one canonical line, $j = 1, \ldots, m$, and the canonical lines of A_{j_1} and A_{j_2}, $j_1 \neq j_2$, are distinct. Note that because A is simple, each A_j is congruent, via a diagonal real matrix, to a scalar unitary matrix, say, $e^{i\gamma_j} I_{k_j}$. Clearly, the distinct canonical angles for A are $\gamma_1, \ldots, \gamma_m$. According to

Theorem 6, for sufficiently small ε, each $A_\varepsilon \in \mathcal{V}_\varepsilon(A)$ is congruent to a matrix of the form $A_1^\varepsilon \oplus \cdots \oplus A_m^\varepsilon$ with $A_j^\varepsilon \to A_j$ as $\varepsilon \to 0$. Because of Lemma 8, for sufficiently small ε, A_j^ε, $j = 1, \ldots, m$, is simple unitoid and, as $\varepsilon \to 0$, its canonical angles approach γ_j. Then A_ε is unitoid and, since the canonical angles are unique, they are the union of the canonical angles of each block A_j^ε. Clearly, because A is simple, for sufficiently small ε, no two canonical angles of A_ε differ by π and, thus, A_ε is simple. ∎

4. Perturbation of nonsimple unitoid matrices.

PROPOSITION 10. [illegible] $A = \operatorname{diag}(a, b)$, [illegible] $a \in \{-1, 0\}$ [illegible] $b \in \{0, 1\}$ [illegible] A [illegible] [illegible]. For $\varepsilon > 0$,

$$A_1 = \begin{bmatrix} a + \varepsilon i & 0 \\ 0 & b + \varepsilon \end{bmatrix}$$

is simple unitoid ($0 \notin F(A_1)$). Because of Theorem 9, there is a neighborhood of A_1 in which all matrices are unitoid. Thus, in any neighborhood of A the unitoid matrices occur with positive density. The matrix

$$A_2 = \begin{bmatrix} a - \varepsilon & 0 \\ 0 & b + \varepsilon \end{bmatrix}$$

is nonsimple unitoid. (We observe that the nonsimple unitoid matrices occur with zero density.) It is easily checked that $0 \in \operatorname{int} F(A_3)$ with

$$A_3 = \begin{bmatrix} a - \varepsilon & \varepsilon \\ 0 & b + \varepsilon \end{bmatrix}.$$

But a 2-by-2 matrix with 0 in the interior of its field of values is necessarily nonunitoid. This conclusion also follows from the fact that the eigenvalues of $\Phi(A_3)$ are not on the unit circle. By continuity, in a sufficiently small neighborhood of A_3 every matrix B is such that $\Phi(B)$ has no eigenvalues on the unit circle. Therefore, every such B is nonunitoid and, thus, in any neighborhood of A nonunitoid matrices occur with positive density. ∎

We extend the notion of canonical angles of unitoid matrices to nonsingular nonunitoid matrices. We say that θ is a canonical angle for a nonsingular $A \in M_n$ if A is congruent to a matrix of the form $\left[e^{i\theta}\right] \oplus B$ for some $B \in M_{n-1}$. It follows from the work in [FJ2] that, also for this extension, the canonical angles are a congruential invariant.

THEOREM 11. [illegible] $A \in M_n$ [illegible] $\varepsilon > 0$ [illegible] $\mathcal{V}_\varepsilon(A)$ [illegible] ε, $A_\varepsilon \in \mathcal{V}_\varepsilon(A)$ [illegible] A [illegible] $\varepsilon \to 0$, [illegible] A_ε [illegible] A [illegible].

. Because of Proposition 7, suppose, without loss of generality, that $A = A_1 \oplus \cdots \oplus A_m$, in which A_i is diagonal unitary with just one canonical line, $i = 1, \ldots, m$, and the canonical lines of A_i and A_j, $i \neq j$, are distinct. Note that since A is nonsimple, there is at least a nonsimple block A_i. Without loss of generality, suppose that A_1 is nonsimple. It follows easily from Proposition 10 that for any $\varepsilon > 0$ there are unitoid (simple and nonsimple) and nonunitoid matrices in $\mathcal{V}_\varepsilon(A_1)$, both occurring

with positive density. Because for $A_1^\varepsilon \in \mathcal{V}_\varepsilon(A_1)$, $A_1^\varepsilon \oplus A_2 \oplus \cdots \oplus A_m \in \mathcal{V}_\varepsilon(A)$ is unitoid if and only if A_1^ε is, the first part of the claim follows. By Theorem 6, for sufficiently small ε, each $A_\varepsilon \in \mathcal{V}_\varepsilon(A)$ is congruent to a matrix of the form $A_1^\varepsilon \oplus \cdots \oplus A_m^\varepsilon$, with $A_i^\varepsilon \to A_i$, as $\varepsilon \to 0$. If A_i is simple, by Theorem 9, for sufficiently small ε, A_i^ε is simple and, as $\varepsilon \to 0$, the canonical angles of A_i^ε approach the canonical angles of A_i, completing the proof. □

We conclude with some observations concerning singular unitoid matrices. This allows the statement of a general result about the perturbation of unitoid matrices. If A is a singular unitoid matrix, then A is congruent to a matrix of the form $A' = 0 \oplus D$, with D diagonal unitary. As follows from Proposition 10, nonunitoid as well as unitoid matrices can be obtained by perturbing a singular principal submatrix of A'. Also, if $D_\varepsilon \in \mathcal{V}_\varepsilon(D)$, $\varepsilon > 0$, then $0 \oplus D_\varepsilon \in \mathcal{V}_\varepsilon(A')$, and, thus, according to Theorem 11, if D is nonsimple, unitoid and nonunitoid matrices in any neighborhood of A' can be obtained by perturbing only the summand D.

Our main result is then an immediate consequence of Theorems 9 and 11 and the observations above.

THEOREM 12. [illegible]

Because a nonsingular $A \in M_n$ is unitoid if and only if $A^{-1}A^*$ is similar to a unitary matrix, Theorem 12 implies the following.

COROLLARY 13. [illegible]

(a) [illegible]

(b) [illegible]

REFERENCES

[DJ] C. R. DEPRIMA AND C. R. JOHNSON, *The range of $A^{-1}A^*$ in $GL(n, C)$*, Linear Algebra Appl., 9 (1974), pp. 209–222.

[FJ2] S. FURTADO AND C. R. JOHNSON, *Congruence and $A^{-1}A^*$*, preprint.

[GLR] I. GOHBERG, P. LANCASTER, AND L. RODMAN, *Operator Theory: Advances and Applications*, Birkhäuser Verlag, Basel, Boston, Stuttgart, 1983.

[HJ] R. HORN AND C. R. JOHNSON, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.

[JF] C. R. JOHNSON AND S. FURTADO, A generalization of Sylvester's law of inertia, Linear Algebra Appl., 338 (2001), pp. 287–290.

[T1] R. C. THOMPSON, *The characteristic polynomial of a principal subpencil of a Hermitian matrix pencil*, Linear Algebra Appl., 14 (1976), pp. 135–177.

[T2] R. C. THOMPSON, *Pencils of complex and real symmetric and skew matrices*, Linear Algebra Appl., 147 (1991), pp. 323–371.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 1, pp. 89–104

A PERTURBATION ANALYSIS FOR NONLINEAR SELFADJOINT OPERATOR EQUATIONS*

A. C. M. RAN†, M. C. B. REURINGS‡, AND L. RODMAN§

Abstract. Perturbation analysis, including perturbation bounds, is developed for nonlinear operator equations of the form $X = Q \pm A^*F(X)A$, under perturbations of the given operators Q (which is assumed to be positive definite) and A and of the given operator function $F(X)$ which takes self-adjoint operator values. Stability of fixed points under suitable map perturbations serves as the main technical tool. More detailed analysis is provided in the particular cases where $F(X)$ is a power map.

Key words. operator equations, perturbation bounds, fixed points

AMS subject classifications. 47J05, 47H14, 47H10

DOI. 10.1137/05062873

1. Introduction.

Let $\mathcal{H}$ and $\mathcal{G}$ be two Hilbert spaces. We consider operator equations of the form

$$X = Q \pm A^*F(X)A. \tag{1.1}$$

Here X is the unknown positive semidefinite (or positive definite) matrix or operator, Q is a positive definite operator from $\mathcal{H}$ into itself, A is a linear operator from $\mathcal{H}$ into $\mathcal{G}$, and $F(\cdot)$ is a given (possibly nonlinear) function.

The equation (1.1), especially for matrices, has been extensively studied in the literature; see the papers [5, 9, 10, 11, 8, 16, 17, 18, 19, 20, 21] and the book [13]. The interest to study (1.1) arose, in particular, in connection with algebraic Riccati equations and interpolation; see, for example, [5, 23]. Fixed point theory techniques play a key role in many recent developments (see [18, 21]).

Of particular interest in applications are results on perturbations and bounds for solutions of (1.1). Several approaches have been explored in the literature. In [19] a perturbation theory was developed for the matrix problem with $\mathcal{H} = \mathcal{G}$, where the map F is kept fixed and perturbations of A and Q are allowed. The result obtained was [illegible], i.e., nearby equations have nearby solutions (provided they exist), where the difference between the solutions is of perturbed equations and the original solution is of the same magnitude as the difference between the coefficients of the perturbed equation and the coefficients of the original equation. In [22] the same results are stated in case $\mathcal{H}$ and $\mathcal{G}$ are not necessarily equal but still are finite dimensional. In [4] it was assumed that F and Q are kept fixed and perturbations of A are allowed.

*Received by the editors April 7, 2005; accepted for publication (in revised form) by H. J. Woerdeman September 21, 2005; published electronically March 17, 2006.
http://www.siam.org/journals/simax/28-1/62873.html

†Department of Mathematics, Faculteit der Exacte Wetenchappen, Vrije Universiteit, De Boelelaan 1083a, 1081 HV Amsterdam, The Netherlands (acm.ran@few.vu.nl).

‡Burg. Hogguerstraat 495, 1064 CT Amsterdam, The Netherlands (martine.reurings@swov.nl).
The research of this author was supported by The Netherlands Organization for Scientific Research (NWO).

§Department of Mathematics, College of William and Mary, Williamsburg, VA 23187-8795 (lxrodm@math.wm.edu). The research of this author was supported in part by an NWO grant, by an NSF grant, and by a Faculty Research Assignment of the College of William and Mary.

In the present paper we develop a perturbation analysis, including perturbation bounds, for operator equations of type (1.1). In contrast with the previous work, we allow changes in all three constituents A, Q, and F. Our approach is based on stability results in the abstract framework of maps on complete metric spaces, taking the cue from [18, 21]. Although our primary interest lies in matrix equations (1.1), i.e., with finite dimensional $\mathcal{H}$ and $\mathcal{G}$, it turns out that our results do not depend on finite dimensionality of the underlying Hilbert spaces. Therefore, we formulate and prove the results in the context of linear bounded operators acting on (possibly infinite dimensional) Hilbert spaces. In the terminology of [13], we consider situations when solutions of (1.1) are proper and give explicit error bounds. The book [13] contains detailed perturbation analysis and error bounds for symmetric and nonsymmetric matrix equations of degree at most two, which are developed using other techniques. Here, (1.1) may be of degree higher than two; on the other hand, (1.1) is assumed to be symmetric.

In the next section we review the needed abstract results on fixed points and their perturbations. Our main theorems, Theorems 3.1 and 3.3, are stated in section 3. In section 4 we specialize to the particular cases of power maps that are frequently encountered in applications and provide more detailed analysis.

2. Stability of fixed points with respect to map perturbation. We state the results here in an abstract framework of metric spaces. Let $\mathcal{X}$ be a complete metric space with the distance function $d(\cdot,\cdot)$, and let Ω be a closed subset of $\mathcal{X}$. For every fixed α, $0 < \alpha < 1$, consider the set $\mathcal{M}(\Omega,\alpha)$ of all maps $\Phi : \Omega \to \Omega$ with the property that

$$d(\Phi(x),\Phi(y)) \le \alpha d(x,y) \qquad \forall \quad x,y \in \Omega.$$

The maps in $\mathcal{M}(\Omega,\alpha)$ are necessarily continuous and moreover they are contractions with the contraction constant α.

The next theorem is well known; for instance, see [12] or [14].

THEOREM 2.1. [illegible] $\Phi \in \mathcal{M}(\Omega,\alpha)$ [illegible] $x^* \in \Omega$, [illegible] Φ, [illegible] $x^* = \Phi(x^*)$ [illegible] $x_0 \in \Omega$ [illegible]

$$x^* = \lim_{m\to\infty} x_m,$$

[illegible] $\{x_m\}_{m=0}^{\infty}$ [illegible] $x_m = \Phi(x_{m-1})$ $m = 1,2,\dots,$ [illegible]

$$d(x_m,x^*) \le \frac{\alpha^m}{1-\alpha} d(x_1,x_0), \qquad m = 0,1,\dots .$$

Based on this theorem, we will formulate a perturbation result (Theorem 2.2). In this result two maps Φ and Ψ are involved. The unique fixed points of Φ and of Ψ will be denoted by $x^*(\Phi)$ and $x^*(\Psi)$, respectively.

THEOREM 2.2. [illegible] $\Phi \in \mathcal{M}(\Omega,\alpha)$ [illegible] $\varepsilon > 0$ [illegible] $\Psi \in \mathcal{M}(\Omega,\alpha)$ [illegible]

$$\sup_{x\in\Omega} d(\Psi(x),\Phi(x)) < \min\left\{\frac{1-\alpha}{3}\varepsilon, 1\right\},$$

[illegible]

$$d(x^*(\Psi),x^*(\Phi)) < \varepsilon.$$

. The proof uses a standard approach and is provided here for completeness. Let $\varepsilon > 0$ be given. Now fix an $x_0 \in \Omega$ and select an integer k such that

$$\frac{\alpha^k}{1-\alpha}(1 + d(\Phi(x_0), x_0)) < \frac{\varepsilon}{3}. \tag{2.1}$$

Further, take $\delta > 0$ such that

$$\delta < \min\left\{\frac{1-\alpha}{3}\varepsilon, 1\right\}.$$

It follows from this inequality that

$$\frac{1-\alpha^k}{1-\alpha}\delta < \frac{\varepsilon}{3}.$$

Finally, let Ψ be in $\mathcal{M}(\Omega, \alpha)$ such that

$$\sup_{x\in\Omega} d(\Psi(x), \Phi(x)) < \delta. \tag{2.2}$$

We will prove that this implies that $d(x^*(\Phi), x^*(\Psi)) < \varepsilon$.

First note that because of (2.2), $\delta < 1$, and (2.1) we have that

$$\frac{\alpha^k}{1-\alpha} d(\Psi(x_0), x_0) \le \frac{\alpha^k}{1-\alpha}(d(\Psi(x_0), \Phi(x_0)) + d(\Phi(x_0), x_0)) < \frac{\varepsilon}{3}.$$

Therefore, by Theorem 2.1, we also have the inequalities

$$\begin{aligned} d(\Phi^k(x_0), x^*(\Phi)) &\le \frac{\alpha^k}{1-\alpha} d(\Phi(x_0), x_0) < \frac{\varepsilon}{3}, \\ d(\Psi^k(x_0), x^*(\Psi)) &\le \frac{\alpha^k}{1-\alpha} d(\Psi(x_0), x_0) < \frac{\varepsilon}{3}. \end{aligned} \tag{2.3}$$

Next denote for $m \ge 2$

$$c_m := \sup_{x\in\Omega} d(\Phi^m(x), \Psi^m(x))$$

and note that for all integers $m \ge 2$ and for all $x \in \Omega$:

$$\begin{aligned} d(\Phi^m(x), \Psi^m(x)) &\le d(\Phi^{m-1}(\Phi(x)), \Phi^{m-1}(\Psi(x))) + d(\Phi^{m-1}(\Psi(x)), \Psi^{m-1}(\Psi(x))) \\ &\le \alpha^{m-1} d(\Phi(x), \Psi(x)) + c_{m-1} \\ &\le \alpha^{m-1}\delta + c_{m-1}. \end{aligned}$$

Taking the supremum over all $x \in \Omega$ gives us that

$$c_m \le \alpha^{m-1}\delta + c_{m-1}, \quad m = 2, 3, \dots,$$

and thus

$$c_m \le \alpha^{m-1}\delta + \alpha^{m-2}\delta + \cdots + \alpha\delta + c_1.$$

Because of (2.2) we know that $c_1 < \delta$, so we derive

$$c_m \le \frac{1-\alpha^m}{1-\alpha}\delta,$$

which implies that

$$d(\Phi^m(x_0), \Psi^m(x_0)) \leq \frac{1-\alpha^m}{1-\alpha}\delta.$$

For $m = k$ the right-hand side is smaller than $\frac{\varepsilon}{3}$. This all leads to

$$\begin{aligned} d(x^*(\Phi), x^*(\Psi)) &\leq d(\Phi^k(x_0), x^*(\Phi)) + d(\Psi^k(x_0), x^*(\Psi)) + d(\Phi^k(x_0), \Psi^k(x_0)) \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon, \end{aligned}$$

which we wanted to prove. □

3. Applications to nonlinear operator equations. Let $\mathcal{H}, \mathcal{G}$ be (complex) Hilbert spaces, and let $\mathcal{L}(\mathcal{G}, \mathcal{H})$ stand for the Banach space of linear bounded operators from $\mathcal{G}$ into $\mathcal{H}$, considered with the operator norm $\|\cdot\|$. The notation $\mathcal{L}(\mathcal{H})$ is used as an abbreviation for $\mathcal{L}(\mathcal{H}, \mathcal{H})$. We denote by $\mathcal{P}(\mathcal{H})$ the set of positive definite invertible operators in $\mathcal{L}(\mathcal{H})$ and by $\overline{\mathcal{P}}(\mathcal{H})$ the closure of $\mathcal{P}(\mathcal{H})$, i.e., the set of positive semidefinite operators in $\mathcal{L}(\mathcal{H})$. Different notations for $X \in \mathcal{P}(\mathcal{H})$ and $X \in \overline{\mathcal{P}}(\mathcal{H})$ will be $X > 0$ and $X \geq 0$, respectively. Also, $X > Y$ (resp., $X \geq Y$) is used to denote that $X - Y > 0$ (resp., $X - Y \geq 0$).

In this section we will consider the operator equation

$$Q = X - A^*F(X)A, \tag{3.1}$$

where $Q \in \mathcal{P}(\mathcal{H})$, $A \in \mathcal{L}(\mathcal{H}, \mathcal{G})$, and F is a (possible nonlinear) map defined on a subset of $\overline{\mathcal{P}}(\mathcal{H})$ and taking self-adjoint values in $\mathcal{L}(\mathcal{G})$. The solutions of (3.1) are exactly the fixed points of the map

$$\Phi(X) = Q + A^*F(X)A. \tag{3.2}$$

We set $\mathcal{X} = \overline{\mathcal{P}}(\mathcal{H})$, so $\mathcal{X}$, equipped with the distance function induced by the norm $\|\cdot\|$, is a complete metric space. Further, we let Ω be a closed subset of $\mathcal{X}$ such that $\Phi : \Omega \to \Omega$. The existence of such a set Ω is of course not automatic. In specific examples it will be necessary to impose certain conditions on Q, A, and F to guarantee existence of Ω. We shall pay careful attention to this point in the next section, where particular cases are discussed, but for now we just assume that Q, A, and F are such that there exists some closed subset Ω in $\mathcal{X}$ which is invariant under Φ. Corresponding to F and Ω we introduce the value $M_{F,\Omega} > 0$, which is the smallest number β satisfying

$$\|F(X) - F(Y)\| \leq \beta\|X - Y\| \tag{3.3}$$

for all $X, Y \in \Omega$, provided such a value exists. Again, existence of such a value depends heavily on the particular form of F, but we shall assume its existence in this section. It follows from (3.3) and the definition of $M_{F,\Omega}$ that

$$\begin{aligned} \|\Phi(X) - \Phi(Y)\| &= \|A^*(F(X) - F(Y))A\| \leq \|A\|^2\|F(X) - F(Y)\| \\ &\leq M_{F,\Omega}\|A\|^2\|X - Y\|, \end{aligned}$$

so if $M_{F,\Omega}\|A\|^2 < 1$, then $\Phi \in \mathcal{M}(\Omega, M_{F,\Omega}\|A\|^2)$ and (3.1) has a unique solution in Ω, which we will denote by $X(\Phi)$.

Now consider the perturbed equation

$$\widetilde{Q} = X - \widetilde{A}^* \widetilde{F}(X) \widetilde{A}, \tag{3.4}$$

where $\widetilde{A}$ and $\widetilde{Q}$ are small perturbations of A and Q, respectively, and $\widetilde{F}$ is the perturbation of F given by

$$\widetilde{F}(X) = F(X) + E(X),$$

where E is some map on Ω into the set of bounded self-adjoint operators on $\mathcal{G}$. Let Ψ be the map corresponding to this equation, i.e., the solutions of (3.4) are the fixed points of Ψ:

$$\Psi(X) = \widetilde{Q} + \widetilde{A}^* \widetilde{F}(X) \widetilde{A}. \tag{3.5}$$

Again we have that $\Psi \in \mathcal{M}(\Omega, M_{\widetilde{F},\Omega} \|\widetilde{A}\|^2)$ provided $M_{\widetilde{F},\Omega} \|\widetilde{A}\|^2 < 1$ and $\Psi : \Omega \to \Omega$. In this case also (3.4) has a unique solution in Ω, which we will denote by $X(\Psi)$. Moreover, if $\|\widetilde{A}\| \le \|A\|$ and $M_{\widetilde{F},\Omega} < M_{F,\Omega}$, then also $\Psi \in \mathcal{M}(\Omega, M_{F,\Omega} \|A\|^2)$. We will now apply Theorem 2.2 to Φ.

THEOREM 3.1. [illegible] Φ [illegible] (3.2) [illegible] $M_{F,\Omega} \|A\|^2 < 1$ [illegible] $\varepsilon > 0$ [illegible] Ψ [illegible] (3.5) [illegible] $\Psi \in \mathcal{M}(\Omega, M_{F,\Omega} \|A\|^2)$ [illegible]

$$\sup_{X \in \Omega} \|\Psi(X) - \Phi(X)\| < \min \left\{ \frac{1 - M_{F,\Omega} \|A\|^2}{3} \varepsilon, 1 \right\}, \tag{3.6}$$

[illegible]

$$\|X(\Phi) - X(\Psi)\| < \varepsilon.$$

Now note that

$$\begin{aligned}
\|\Phi(X) - \Psi(X)\| &= \|\widetilde{Q} - Q + \widetilde{A}^* \widetilde{F}(X) \widetilde{A} - A^* F(X) A\| \\
&\le \|\widetilde{Q} - Q\| + \|\widetilde{A}^* \widetilde{F}(X) \widetilde{A} - A^* F(X) A\| \\
&= \|\widetilde{Q} - Q\| \\
&\quad + \|\widetilde{A}^* \widetilde{F}(X)(\widetilde{A} - A) + \widetilde{A}^* (\widetilde{F}(X) - F(X)) A + (\widetilde{A}^* - A^*) F(X) A\| \\
&\le \|\widetilde{Q} - Q\| + (\|\widetilde{A}\| \|\widetilde{F}(X)\| + \|A\| \|F(X)\|) \|\widetilde{A} - A\| \\
&\quad + \|\widetilde{A}\| \|A\| \|\widetilde{F}(X) - F(X)\| \\
&= \|\widetilde{Q} - Q\| + (\|\widetilde{A}\| \|\widetilde{F}(X)\| + \|A\| \|F(X)\|) \|\widetilde{A} - A\| \\
&\quad + \|\widetilde{A}\| \|A\| \|E(X)\|,
\end{aligned}$$

so the left-hand side of (3.6) can be estimated as

$$\begin{aligned}
\sup_{X \in \Omega} \|\Psi(X) - \Phi(X)\| \le \|\widetilde{Q} - Q\| &+ \sup_{X \in \Omega} \|\widetilde{A}\| \|A\| \|E(X)\| \\
&+ \|\widetilde{A} - A\| \sup_{X \in \Omega} (\|\widetilde{A}\| \|\widetilde{F}(X)\| + \|A\| \|F(X)\|).
\end{aligned}$$

Hence we have the following corollary.

COROLLARY 3.2. [illegible] Φ [illegible] (3.2) [illegible] $\Phi \in \mathcal{M}(\Omega, M_{F,\Omega}\|A\|^2)$,

$$M_{F,\Omega}\|A\|^2 < 1.$$

[illegible] $\varepsilon > 0$ [illegible] Ψ [illegible] (3.5) [illegible]

$$\Psi \in \mathcal{M}(\Omega, M_{F,\Omega}\|A\|^2)$$

[illegible]

$$\|\widetilde{Q} - Q\| + \|\widetilde{A} - A\| \sup_{X\in\Omega}(\|\widetilde{A}\|\|\widetilde{F}(X)\| + \|A\|\|F(X)\|) + \sup_{X\in\Omega}\|\widetilde{A}\|\|A\|\|E(X)\| < \min\left\{\frac{1 - M_{F,\Omega}\|A\|^2}{3}\varepsilon, 1\right\},$$

[illegible]

$$\|X(\Phi) - X(\Psi)\| < \varepsilon.$$

The condition that $\Psi \in \mathcal{M}(\Omega, M_{F,\Omega}\|A\|^2)$ can be incorporated in the other hypotheses of Corollary 3.2, at the expense of making the other hypotheses more stringent, as the next theorem shows.

THEOREM 3.3. [illegible] Φ [illegible] (3.2) [illegible] $M_{F,\Omega}\|A\|^2 < 1$ [illegible] $\varepsilon > 0$ [illegible] $\Psi : \Omega \to \Omega$ [illegible] (3.4) [illegible]

$$\|\widetilde{Q} - Q\| + \|\widetilde{A} - A\| \sup_{X\in\Omega}(\|\widetilde{A}\|\|\widetilde{F}(X)\| + \|A\|\|F(X)\|) + \sup_{X\in\Omega}\|\widetilde{A}\|\|A\|\|E(X)\| < \min\left\{\frac{1 - M_{F,\Omega}\|A\|^2}{3}\varepsilon, \frac{1 - M_{\widetilde{F},\Omega}\|\widetilde{A}\|^2}{3}\varepsilon, 1\right\}$$

[illegible]

(i) $\|\widetilde{A}\| \le \|A\|, M_{\widetilde{F},\Omega} \le M_{F,\Omega}$,

(ii) $\|\widetilde{A}\| \ge \|A\|$ [illegible] $\|\widetilde{A} - A\| < \frac{1}{\sqrt{M_{\widetilde{F},\Omega}}} - \|A\|$,

[illegible]

$$\|X(\Phi) - X(\Psi)\| < \varepsilon.$$

[illegible]. If $\widetilde{A}$ and $M_{\widetilde{F},\Omega}$ satisfy condition (i), then it is easy to see that $M_{\widetilde{F},\Omega}\|\widetilde{A}\|^2 < 1$. In case (ii) holds we can use similar lines as in Proposition 4.2 in [19] to show that the inequality holds. Indeed,

$$M_{\widetilde{F},\Omega}\|\widetilde{A}\|^2 = M_{\widetilde{F},\Omega}\|A + (\widetilde{A} - A)\|^2 \le M_{\widetilde{F},\Omega}(\|A\| + \|\widetilde{A} - A\|)^2 < M_{\widetilde{F},\Omega}\left(\frac{1}{M_{\widetilde{F},\Omega}}\right)^2 = 1.$$

Now let α be defined by

$$\alpha = \max\left\{M_{F,\Omega}\|A\|^2, M_{\widetilde{F},\Omega}\|\widetilde{A}\|^2\right\},$$

then Φ and Ψ both are in $\mathcal{M}(\Omega, \alpha)$. Now apply Corollary 3.2 with $M_{F,\Omega}\|A\|^2$ replaced by α. □

A special case of the types of perturbations we are studying in the present paper was already discussed in [19] (for finite dimensional $\mathcal{H}$ and $\mathcal{G}$). In [19] the perturbed equation

$$\widetilde{Q} = X - \widetilde{A}^* F(X) \widetilde{A} \tag{3.7}$$

is considered, i.e., (3.4) with $\widetilde{F} = F$. However, if we apply the previous results to (3.7), assuming $\mathcal{H}$ is finite dimensional, the obtained perturbation bounds turn out to be weaker than those of [19]. We record only one corollary that can be obtained that way.

COROLLARY 3.4. [illegible] Φ [illegible] (3.2) [illegible] $\Phi \in \mathcal{M}(\Omega, M_{F,\Omega}\|A\|^2)$, [illegible] $M_{F,\Omega}\|A\|^2 < 1$. [illegible] $\varepsilon > 0$ [illegible] Ψ [illegible] (3.4) [illegible] $\Psi \in \mathcal{M}(\Omega, M_{F,\Omega}\|A\|^2)$ [illegible]

$$\|\widetilde{Q} - Q\| + \|\widetilde{A} - A\|(\|\widetilde{A}\| + \|A\|) \sup_{X \in \Omega} \|F(X)\| < \min\left\{ \frac{1 - M_{F,\Omega}\|A\|^2}{3}\varepsilon, 1 \right\},$$

[illegible]

$$\|X(\Phi) - X(\Psi)\| < \varepsilon.$$

The analysis of the equation

$$Q = X + A^* F(X) A$$

is the same as the one for the equation already considered, as one sees immediately by replacing F by $-F$. Recall that F is assumed only to take self-adjoint values, not necessarily positive definite values.

4. Power maps. In this section we will specialize the results of the previous section to particular cases of the power maps

$$F_1(X) = X^m, \qquad m \le 1 \text{ or } m \ge 2,$$

and

$$F_2(X) = -X^m, \qquad m \le 1 \text{ or } m \ge 2,$$

which are the cases frequently encountered in applications (see [15, 3, 20, 5], for example). Thus, we consider equations

$$X = Q + A^* X^m A, \quad m \le 1 \text{ or } m \ge 2, \tag{4.1}$$

and

$$X = Q - A^* X^m A, \quad m \le 1 \text{ or } m \ge 2. \tag{4.2}$$

Note that we take $\mathcal{H} = \mathcal{G}$ for both equations. As the cases $m \le 1$ and $m \ge 2$ in (4.1) require slightly different arguments, they are considered separately.

Recall that Q is assumed to be positive definite and invertible.

4.1. The equation $X = Q + A^*X^mA$ with $m \leq 1$. It is easy to see that the map

$$G_1(X) := Q + A^*X^mA$$

maps the set $\{X \in \mathcal{P}(\mathcal{H}) | X \geq Q\}$ into itself. However, this set depends on the map G_1. Thus, we cannot expect also that the map

$$\widetilde{G}_1 := \widetilde{Q} + \widetilde{A}^*(X^m + E(X))\widetilde{A}$$

corresponding to the perturbed equation

$$X = \widetilde{Q} + \widetilde{A}^*(X^m + E(X))\widetilde{A} \tag{4.3}$$

maps that set into itself. Therefore, we will take $\Omega = \Omega_1$ as follows:

$$\Omega_1 = \{X \in \mathcal{P}(\mathcal{H}) | X \geq P\},$$

where P is a positive definite invertible operator such that $Q \geq P$, and we will only consider perturbations $\widetilde{Q}$ of Q which satisfy $\widetilde{Q} \geq P$. If we assume in addition that E is such that

$$X^m + E(X) \geq 0 \quad \forall X \geq P, \tag{4.4}$$

then both G_1 and $\widetilde{G}_1$ map Ω_1 into itself.

Before continuing our arguments, let us briefly comment on the fact that at least in the finite dimensional case one can easily deduce the solvability of the equations under consideration by constructing a compact and convex subset of the positive definite matrices that is mapped into itself by G_1. As G_1 is continuous on $\mathcal{P}(\mathcal{H})$, it follows by Schauder's fixed point theorem that a solution will exist.

To determine M_{F_1,Ω_1}, we will make use of the generalized mean value theorem (see Theorem 1.1.8 in [1]), which is stated below.

THEOREM 4.1. Let $\widetilde{\mathcal{G}}$, $\widetilde{\mathcal{H}}$ be Hilbert spaces, $\mathcal{U}$ an open subset of $\widetilde{\mathcal{G}}$, $\Phi : \mathcal{U} \to \widetilde{\mathcal{H}}$ a Fréchet differentiable function on $\mathcal{U}$, $X, Y \in \mathcal{U}$ such that $L_{X,Y} \subset \mathcal{U}$, then

$$\|\Phi(X) - \Phi(Y)\|_{\widetilde{\mathcal{H}}} \leq \sup_{Z \in L_{X,Y}} \|D\Phi(Z)\| \|X - Y\|_{\widetilde{\mathcal{G}}}.$$

Here $L_{X,Y}$ is the line segment joining X and Y, $D\Phi(Z)$ denotes the Fréchet derivative of Φ at Z and

$$\|D\Phi(Z)\| = \sup_{H \in \widetilde{\mathcal{H}},\ \|H\|=1} \|D\Phi(Z)(H)\|.$$

We will also use the equality

$$\|DF_1(X)\| = |m| \|X^{m-1}\|, \quad X \in \mathcal{P}(\mathcal{H}),$$

where $\mathcal{H}$ is a Hilbert space. This equality is proved in [2] for $m \in (-\infty, 1] \cup [2, \infty)$.

Now we have for all $X, Y \in \Omega_1$,

$$\begin{aligned}
\|F_1(X) - F_1(Y)\| = \|X^m - Y^m\| &\leq \sup_{Z \in L_{X,Y}} \|DF_1(Z)\| \|X - Y\| \\
&= |m| \sup_{Z \in L_{X,Y}} \|Z^{m-1}\| \|X - Y\| \leq |m| \sup_{Z \in \Omega_1} \|Z^{m-1}\| \|X - Y\| \\
&\leq |m| \sup_{Z \in \Omega_1} \|Z^{-1}\|^{1-m} \|X - Y\| \leq |m| \|P^{-1}\|^{1-m} \|X - Y\|,
\end{aligned}$$

so we can take $M_{F_1,\Omega_1} = |m|\|P^{-1}\|^{1-m}$. Note that in the last inequality we used the assumption $m \le 1$.

Together with Theorem 3.3 the foregoing proves the following result.

THEOREM 4.2. [illegible] $Q \in \mathcal{P}(\mathcal{H})$, $m \le 1$ $A \in \mathcal{L}(\mathcal{H})$ [illegible] $P \in \mathcal{P}(\mathcal{H})$ [illegible] $Q \ge P$ [illegible] $|m|\|P^{-1}\|^{1-m}\|A\|^2 < 1$. [illegible] $\varepsilon > 0$ [illegible] $\widetilde{A}$, $\widetilde{Q}$ [illegible] E [illegible]

1. $\widetilde{Q} \ge P$
2. $X^m + E(X) \ge 0$ [illegible] $X \ge P$
3. $$\|\widetilde{Q} - Q\| + \sup_{X\in\Omega_1} (\|\widetilde{A}\|\|X^m + E(X)\| + \|A\|\|X^m\|)\|\widetilde{A} - A\| + \sup_{X\in\Omega_1} \|\widetilde{A}\|\|A\|\|E(X)\| < \min\left\{\frac{1 - |m|\|P^{-1}\|^{1-m}\|A\|^2}{3}\varepsilon, \frac{1 - M_{\widetilde{F}_1,\Omega_1}\|\widetilde{A}\|^2}{3}\varepsilon, 1\right\},$$
4. [illegible]
   (i) $\|\widetilde{A}\| \le \|A\|$,
   (ii) $\|\widetilde{A}\| \ge \|A\|$ [illegible] $\|\widetilde{A} - A\| < \frac{1}{\sqrt{M_{\widetilde{F}_1,\Omega_1}}} - \|A\|$,

[illegible]

$$\|X_S - \widetilde{X}_S\| < \varepsilon$$

[illegible] X_S [illegible] (4.1) [illegible] Ω_1 [illegible] $\widetilde{X}_S$ [illegible] (4.3) [illegible] Ω_1.

If we assume that E satisfies

$$\|E(X) - E(Y)\| \le M_{E,\Omega_1}\|X - Y\| \quad \forall X, Y \in \Omega_1, \tag{4.5}$$

for some $M_{E,\Omega_1} > 0$, then

$$\begin{aligned}\|\widetilde{F}_1(X) - \widetilde{F}_1(Y)\| &= \|F_1(X) + E(X) - F_1(Y) - E(Y)\| \\ &\le \|F_1(X) - F_1(Y)\| + \|E(X) - E(Y)\| \\ &\le (M_{F_1,\Omega_1} + M_{E,\Omega_1})\|X - Y\|.\end{aligned}$$

Hence we can take $M_{\widetilde{F}_1,\Omega_1} = M_{F_1,\Omega_1} + M_{E,\Omega_1}$.

An interesting special case of the perturbations studied above is $\widetilde{Q} = Q, \widetilde{A} = A$, and $E(X) = X^{\widetilde{m}} - X^m$, where $\widetilde{m} \le 1$ is a small perturbation of m. So we are interested in the relation between solutions of

$$X = Q + A^* X^{\widetilde{m}} A, \quad \widetilde{m} \le 1, \tag{4.6}$$

and solutions of (4.1) if $|m - \widetilde{m}|$ is small. Note that $\widetilde{F}_1(X) = X^{\widetilde{m}}$ in this case, so we have $M_{\widetilde{F}_1,\Omega_1} = |\widetilde{m}|\|P^{-1}\|^{1-\widetilde{m}}$. Hence, if we apply Theorem 4.2, we find that E has to satisfy

$$\sup_{X\in\Omega_1} \|A\|^2\|E(X)\| < \min\left\{\frac{1 - |m|\|P^{-1}\|^{1-m}\|A\|^2}{3}\varepsilon, \frac{1 - |\widetilde{m}|\|P^{-1}\|^{1-\widetilde{m}}\|A\|^2}{3}\varepsilon, 1\right\}$$

for $\|X_S - \widetilde{X}_S\| < \varepsilon$ to hold. However, it is not possible to find an upper bound for

$$\sup_{X\in\Omega_1} \|E(X)\| = \sup_{X\in\Omega_1} \|X^{\widetilde{m}} - X^m\|$$

for the Ω_1 we used before. Thus we will need a different Ω_1.

From now on in this subsection we will assume that $Q = I$. First we shall discuss the case where m and $\widetilde{m}$ are [illegible]. Under these assumptions we can choose $\Omega = \widehat{\Omega}_1$ as follows:

$$\widehat{\Omega}_1 = \{X \in \mathcal{P}(\mathcal{H}) | I \leq X \leq I + A^*A\}.$$

The following lemma is needed in the derivation for an upper bound for $\|E(X)\|$.

LEMMA 4.3. [illegible] $X \in \mathcal{P}(\mathcal{H})$ [illegible] $X \geq I$ [illegible] p [illegible] q [illegible] $p > q$. [illegible] $X^p - X^q$ [illegible]. Because $X \geq I$ we can write

$$X^p - X^q = \int_1^{\|X\|} (t^p - t^q)\, dE(t),$$

where $E(t)$ is the spectral measure of X. Clearly this is positive semidefinite as $t^p - t^q \geq 0$ for $t \geq 1$. □

Now let $X \in \widehat{\Omega}_1$ and assume that $\widetilde{m} > m$. Note that $X \geq I$ implies that $X^{\widetilde{m}} \leq I$, because $\widetilde{m} < 0$. Further, it follows from $X \leq I + A^*A$ that $X \leq (1 + \|A\|^2)I$, which implies that $X^m \geq (1 + \|A\|^2)^m I$. So we have that

$$0 \leq X^{\widetilde{m}} - X^m \leq I - (1 + \|A\|^2)^m I = (1 - (1 + \|A\|^2)^m)I.$$

Interchanging $\widetilde{m}$ and m we find that

$$0 \leq X^m - X^{\widetilde{m}} \leq I - (1 + \|A\|^2)^{\widetilde{m}} I = (1 - (1 + \|A\|^2)^{\widetilde{m}})I$$

if $\widetilde{m} < m$. Because of one of the assumptions on the norm $\|\cdot\|$ this leads to

$$\begin{aligned} \sup_{X \in \widehat{\Omega}_1} \|X^{\widetilde{m}} - X^m\| &\leq \sup_{I \leq X \leq (1+\|A\|^2)I} \|X^{\widetilde{m}} - X^m\| \\ &\leq \max\{1 - (1 + \|A\|^2)^m, 1 - (1 + \|A\|^2)^{\widetilde{m}}\}. \end{aligned}$$

Hence we have the following theorem.

THEOREM 4.4. [illegible] $m < 0$ [illegible] $A \in \mathcal{L}(\mathcal{H})$ [illegible] $|m|\|A\|^2 < 1$. [illegible] $Q = I$ [illegible] $\varepsilon > 0$ [illegible] $E(X) = X^{\widetilde{m}} - X^m, \widetilde{m} < 0,$ [illegible] $|\widetilde{m}|\|A\|^2 < 1$ [illegible]

$$\begin{aligned} \|A\|^2 \max\{1 - (1 + \|A\|^2)^m, 1 - (1 + \|A\|^2)^{\widetilde{m}}\} \\ \leq \min\left\{\frac{1 - |m|\|A\|^2}{3}\varepsilon, \frac{1 - |\widetilde{m}|\|A\|^2}{3}\varepsilon, 1\right\}, \end{aligned}$$

[illegible]

$$\|X_S - \widetilde{X}_S\| < \varepsilon$$

[illegible] X_S [illegible] (4.1) $\widehat{\Omega}_1$ [illegible] $\widetilde{X}_S$ [illegible] (4.6) $\widehat{\Omega}_1$.

Next we consider that case where m and $\widetilde{m}$ are strictly between 0 and 1. Since $m < 1$, the scalar equation $1 + \|A\|^2 x^m = x$ has a unique positive solution, which we shall denote by $\beta(m)$. Observe that if $\widetilde{m} < m$, then $1 < \beta(\widetilde{m}) < \beta(m)$ (unless $A = 0$,

a trivial case we shall not consider). We choose $\widehat{\Omega}_1 = [I, \beta(m)I]$, and we claim that this set is mapped into itself by both G_1 and $\widetilde{G}_1$. Indeed, let $I \le X \le \beta(m)I$; then

$$\|G_1(X)\| \le 1 + \|A\|^2\|X\|^m \le 1 + \|A\|^2\beta(m)^m = \beta(m),$$
$$\|\widetilde{G}_1(X)\| \le 1 + \|A\|^2\|X\|^{\widetilde{m}} \le 1 + \|A\|^2\beta(m)^{\widetilde{m}} \le \beta(m).$$

Next, we consider the difference $X^m - X^{\widetilde{m}}$ for X in $[I, \beta(m)I]$ and $m > \widetilde{m}$. We have

$$0 < X^m - X^{\widetilde{m}} \le (\beta(m)^m - 1)I.$$

So, for the case where $0 < m < 1$ and $\widetilde{m} < m$ we arrive at the following theorem.

THEOREM 4.5. [illegible] $A \in \mathcal{L}(\mathcal{H})$ [illegible] $0 < m < 1$ [illegible] $m\|A\|^2 < 1$ [illegible] $Q = I$ [illegible] $\varepsilon > 0$ [illegible] $0 < \widetilde{m} < m$ [illegible]

$$\|A\|^2(\beta(m)^m - 1) \le \min\left\{\frac{1 - m\|A\|^2}{3}\varepsilon, 1\right\},$$

[illegible]

$$\|X_S - \widetilde{X}_S\| < \varepsilon$$

[illegible] X_S [illegible] $X = I + A^*X^mA$ [illegible] $\widetilde{X}_S$ [illegible] $X = I + A^*X^{\widetilde{m}}A$, [illegible] $[I, \beta(m)I]$

We finish this section with a few remarks on the cases $m = 0$ and $m = 1$, which have not been treated here. In the case $m = 0$ the map is not an interesting one, while the case $m = 1$ amounts to the Stein equation. For the Stein equation there is a well-developed perturbation theory for perturbations of A and Q; see, e.g., [6, 7].

4.2. The equation $X = Q + A^*X^mA$ with $m \ge 2$. In this subsection we shall consider the same equation as in the previous one, but with the assumption that $m \ge 2$. Our first order of business is to find a subset Ω_1 of $\overline{\mathcal{P}}(\mathcal{H})$ that is mapped into itself by the map G_1 and for which we can obtain estimates comparable to the ones in the previous subsection. It is the latter point that requires us to consider sets that are bounded not only from below but also from above.

Consider

$$G_1(x \cdot I) = Q + x^mA^*A, \quad x \text{ real}.$$

We estimate this as follows:

$$\|G_1(xI)\| = \|Q + x^mA^*A\| \le \|Q\| + \|A\|^2x^m.$$

We shall assume that Q and A are such that the equation $x = \|Q\| + \|A\|^2x^m$ has two positive solutions. Let β be the largest of these two solutions. We consider the set

$$\Omega := [0, \beta I] := \{X \in \mathcal{P}(\mathcal{H}) : 0 \le X \le \beta I\}$$

and claim that Ω is mapped into itself by G_1. Indeed, let $0 \le X \le \beta I$; then $0 \le X^m \le \beta^m I$, and hence $Q = G_1(0) \le G_1(X) \le G_1(\beta I)$. However, by the definition of β we have that $G_1(\beta I) \le \beta I$. Indeed, $G_1(\beta I) = Q + \beta^m A^*A$. So

$$\|G_1(\beta)\| \le \|Q\| + \beta^m\|A\|^2 = \beta.$$

As in the previous subsection we comment on the consequence of this inclusion for the finite dimensional case. The set Ω is then a compact convex set, and G_1 being continuous must have a fixed point in Ω by Schauder's fixed point theorem. Hence solvability of the equation is guaranteed in the finite dimensional case.

For the infinite dimensional case we wish to apply the results of section 2, the conditions of which imply uniqueness of the solution. Notice that part of the necessary estimate to come to $M_{F_1,\Omega}$ was already done in the previous subsection. For all $X, Y \in \Omega$ we have

$$\|F_1(X) - F_1(Y)\| \leq m \sup_{Z\in\Omega} \|Z\|^{m-1}\|X - Y\| \leq m\beta^{m-1}\|X - Y\|.$$

Thus we can take $M_{F_1,\Omega} = m\beta^{m-1}$. The following theorem gives conditions under which G_1 is a contraction on the set $[0, \beta I]$.

THEOREM 4.6. $m \geq 2$ $Q \in \mathcal{P}(\mathcal{H})$ $A \in \mathcal{L}(\mathcal{H})$ $x = \|Q\| + \|A\|^2 x^m$ β $m\|A\|^2\beta^{m-1} < 1$ $X = Q + A^*X^mA$ $[0, \beta I]$

Now consider perturbations $\widetilde{G}_1$ given by (4.3). Once again, we need a set Ω_1 which is mapped into itself not only by G_1 but also by the perturbation $\widetilde{G}_1$. In order to achieve this, we slightly change the definition of β as follows: let $q > 0$ and $a > 0$ be such that the equation $x = q + a^2x^m$ has two positive solutions. Let β be the maximum of the largest of the solutions and 1. We consider the map G_1 and its perturbation $\widetilde{G}_1$ under the conditions $\|Q\| \leq q$, $\|\widetilde{Q}\| \leq q$, $\|A\| \leq a$, $\|\widetilde{A}\| \leq a$, and $E(X)$ such that $X^m + E(X) \geq 0$ and $\|X^m + E(X)\| \leq \beta^m$ for $0 < X \leq \beta I$. Under these conditions it is easily seen that both G_1 and $\widetilde{G}_1$ leave the set $\Omega_1 := [0, \beta I]$ invariant. We check this for $\widetilde{G}_1$: for $0 \leq X \leq \beta I$ we have $0 < \widetilde{Q} \leq \widetilde{G}_1(X)$ and

$$\|\widetilde{G}_1(X)\| \leq \|\widetilde{Q}\| + \|\widetilde{A}\|^2(\|X^m + E(X)\|) \leq q + a^2\beta^m \leq \beta.$$

As an interesting example, let $\widetilde{m} < m$ and consider the case where $E(X) = X^{\widetilde{m}} - X^m$. Then both conditions on $E(X)$ are satisfied, as $\|X^{\widetilde{m}}\| \leq \|X\|^{\widetilde{m}}$. Now if $\|X\| \leq 1$, then certainly $\|X\| \leq \beta^m$ as $\beta \geq 1$, while if $1 \leq \|X\| \leq \beta$, then $\|X\|^{\widetilde{m}} \leq \|X\|^m \leq \beta^m$.

We are now in a position to apply Theorem 3.3 to this situation.

THEOREM 4.7. $m \geq 2$ $q > 0$ $a > 0$ $x = q + a^2x^m$ β_0 $\beta = \max\{\beta_0, 1\}$ $\Omega_1 = [0, \beta I]$ $Q, \widetilde{Q} \in \mathcal{P}(\mathcal{H})$ $A, \widetilde{A} \in \mathcal{L}(\mathcal{H})$

$$\|Q\| \leq q, \quad \|\widetilde{Q}\| \leq q, \quad \|A\| \leq a, \quad \|\widetilde{A}\| \leq a,$$

$E(X)$ $X \in \mathcal{P}(\mathcal{H})$

$$X^m + E(X) \geq 0 \quad \text{and} \quad \|X^m + E(X)\| \leq \beta^m \quad \text{for} \quad 0 \leq X \leq \beta I.$$

$m\|A\|^2\beta^{m-1} < 1$ $\varepsilon > 0$ $\widetilde{Q}, \widetilde{A}$

$$\|\widetilde{Q} - Q\| + \sup_{X\in\Omega_1} (\|\widetilde{A}\|\|X^m + E(X)\| + \|A\|\|X^m\|)\|\widetilde{A} - A\| + \sup_{X\in\Omega_1} \|\widetilde{A}\|\|A\|\|E(X)\|$$
$$< \min\left\{\frac{1 - m\|\beta\|^{m-1}\|A\|^2}{3}\varepsilon, \frac{1 - M_{\widetilde{F}_1,\Omega_1}\|\widetilde{A}\|^2}{3}\varepsilon, 1\right\}$$

(i) $\|\widetilde{A}\| \leq \|A\|$,
(ii) $\|\widetilde{A}\| \geq \|A\|$ $\|\widetilde{A} - A\| < \frac{1}{\sqrt{M_{\widetilde{F}_1,\Omega_1}}} - \|A\|$,

$$\|X_S - \widetilde{X}_S\| < \varepsilon$$

X_S (4.1) Ω_1 $\widetilde{X}_S$ (4.3) Ω_1. $M_{\widetilde{F}_1,\Omega_1}\|\widetilde{A}\|^2 < 1$

To make the conditions more transparent we would need to have an estimate for $M_{\widetilde{F}_1,\Omega_1}$. For the case mentioned above, where $E(X) = X^{\widetilde{m}} - X^m$ with $\widetilde{m} < m$ this is easily done, in this case $M_{\widetilde{F}_1,\Omega_1} = \widetilde{m}\beta^{\widetilde{m}-1}$. With $Q = \widetilde{Q}, A = \widetilde{A}$, we get that the condition becomes $\|A\|^2 \sup_{0\leq X\leq \beta I}\|X^m - X^{\widetilde{m}}\| < \frac{1-\widetilde{m}\beta^{\widetilde{m}-1}\|A\|^2}{3}\varepsilon$ to conclude that $\|X_S - \widetilde{X}_S\| < \varepsilon$.

4.3. The equation $X = Q - A^*X^mA$ with $m \in (-\infty, 1] \cup [2, \infty)$. In this subsection we will consider (4.2). For this equation we need a condition on A and Q such that there exists an $\Omega_2 \subset \overline{\mathcal{P}}(\mathcal{H})$ which is mapped into itself by

$$G_2(X) = Q - A^*X^mA.$$

Let us assume that there is an $R \in \mathcal{P}(\mathcal{H})$ such that

$$Q - A^*X^mA \geq R \quad \forall X \geq R. \tag{4.7}$$

Then we can take

$$\Omega = \{X \in \mathcal{P}(\mathcal{H}) | X \geq R\}.$$

If the perturbations $\widetilde{A}, \widetilde{Q}$, and E are such that

$$\widetilde{Q} - \widetilde{A}^*(X^m + E(X))\widetilde{A} \geq R \quad \forall X \geq R, \tag{4.8}$$

then the map $\widetilde{G}_2$ corresponding to

$$X = \widetilde{Q} - \widetilde{A}^*(X^m + E(X))\widetilde{A} \tag{4.9}$$

also maps $\{X \in \mathcal{P}(\mathcal{H}) | X \geq R\}$ into itself. Moreover, it is obvious that $G_2(X) \leq Q$ for all $X \geq 0$ and if we assume in addition that E is such that

$$X^m + E(X) \geq 0 \quad \forall X \geq R, \tag{4.10}$$

then also $\widetilde{G}_2(X) \leq \widetilde{Q}$ for all $X \geq 0$. So if we let B be an operator such that $Q \leq B$ and we only consider perturbations $\widetilde{Q}$ of Q satisfying $\widetilde{Q} \leq B$, then it is clear that G_2 and $\widetilde{G}_2$ map

$$\Omega_2 = \{X \in \mathcal{P}(\mathcal{H}) | R \leq X \leq B\}$$

into itself.

Analogously to subsection 4.1 we find $M_{F_2,\Omega_2} = |m|\|R^{-1}\|^{1-m}$ in case $m \leq 1$. If $m \geq 2$, then M_{F_2,Ω_2} is slightly different. Indeed, in this case we have for all $X, Y \in \Omega_2$

$$\begin{aligned}\|F_2(X) - F_2(Y)\| = \|X^m - Y^m\| &\leq \sup_{Z\in L_{X,Y}} \|DF_2(Z)\|\|X - Y\| \\ &= |m| \sup_{Z\in L_{X,Y}} \|Z^{m-1}\|\|X - Y\| \leq |m| \sup_{Z\in\Omega_2} \|Z^{m-1}\|\|X - Y\| \\ &\leq |m| \sup_{Z\in\Omega_2} \|Z\|^{m-1}\|X - Y\| \leq |m|\|B\|^{m-1}\|X - Y\|.\end{aligned}$$

Thus we have $M_{F_2,\Omega_2} = |m|\|B\|^{m-1}$, and we obtain the following results.

THEOREM 4.8. $m \leq 1$ $Q \in \mathcal{P}(\mathcal{H}), A \in \mathcal{L}(\mathcal{H})$ $R \in \mathcal{P}(\mathcal{H})$ (4.7) $|m|\|R^{-1}\|^{1-m}\|A\|^2 < 1$ $\varepsilon > 0$ $\widetilde{A}, \widetilde{Q}$ E (4.8) (4.10)

$$\|\widetilde{Q} - Q\| + \sup_{X \in \Omega_2} (\|\widetilde{A}\|\|X^m + E(X)\| + \|A\|\|X^m\|)\|\widetilde{A} - A\| + \sup_{X \in \Omega_2} \|\widetilde{A}\|\|A\|\|E(X)\|$$
$$< \min\left\{\frac{1 - |m|\|R^{-1}\|^{1-m}\|A\|^2}{3}\varepsilon, \frac{1 - M_{\widetilde{F}_2,\Omega_2}\|\widetilde{A}\|^2}{3}\varepsilon, 1\right\}$$

(i) $\|\widetilde{A}\| \leq \|A\|$,
(ii) $\|\widetilde{A}\| \geq \|A\|$ $\|\widetilde{A} - A\| < \frac{1}{\sqrt{M_{\widetilde{F}_2,\Omega_2}}} - \|A\|$,

$$\|X_S - \widetilde{X}_S\| < \varepsilon$$

X_S (4.2) Ω_2 $\widetilde{X}_S$ (4.9) Ω_2.

THEOREM 4.9. $m \geq 2$ $Q \in \mathcal{P}(\mathcal{H}), A \in \mathcal{L}(\mathcal{H})$ $R \in \mathcal{P}(\mathcal{H})$ (4.7) $B \in \mathcal{P}(\mathcal{H})$ $Q \leq B$ $|m|\|B\|^{m-1}\|A\|^2 < 1$. $\varepsilon > 0$ $\widetilde{A}, \widetilde{Q}$ E $\widetilde{Q} \leq B$ (4.8) (4.10)

$$\|\widetilde{Q} - Q\| + \sup_{X \in \Omega_2} (\|\widetilde{A}\|\|X^m + E(X)\| + \|A\|\|X^m\|)\|\widetilde{A} - A\| + \sup_{X \in \Omega_2} \|\widetilde{A}\|\|A\|\|E(X)\|$$
$$< \min\left\{\frac{1 - |m|\|B\|^{m-1}\|A\|^2}{3}\varepsilon, \frac{1 - M_{\widetilde{F}_2,\Omega_2}\|\widetilde{A}\|^2}{3}\varepsilon, 1\right\}$$

(i) $\|\widetilde{A}\| \leq \|A\|$,
(ii) $\|\widetilde{A}\| \geq \|A\|$ $\|\widetilde{A} - A\| < \frac{1}{\sqrt{M_{\widetilde{F}_2,\Omega_2}}} - \|A\|$,

$$\|X_S - \widetilde{X}_S\| < \varepsilon$$

X_S (4.2) Ω_2 $\widetilde{X}_S$ (4.9) Ω_2.

Independently of m we can take $M_{\widetilde{F}_2,\Omega_2} = M_{F_2,\Omega_2} + M_{E,\Omega_2}$, if we assume again that E satisfies (4.5) for some $M_{E,\Omega_2} > 0$.

Also for (4.2) and the perturbed equation (4.9) we will discuss the special case $\widetilde{Q} = Q, \widetilde{A} = A$ and $E(X) = X^{\widetilde{m}} - X^m$, where $\widetilde{m} \in (-\infty, 1] \cup [2, \infty)$ is a small perturbation of m such that

$$Q - A^* X^{\widetilde{m}} A \geq R \quad \forall X \geq R. \tag{4.11}$$

So we are interested in the relation between solutions of (4.2) and

$$X = Q - A^* X^{\widetilde{m}} A \tag{4.12}$$

for $|\widetilde{m} - m|$ small. In the rest of this section we set $B = Q$.

Now let $X \in [R, Q]$, i.e., $X \in \mathcal{P}(\mathcal{H})$ and $R \le X \le Q$. Then

$$\|R^{-1}\|^{-1} I \le X \le \|Q\| I.$$

So

$$\|R^{-1}\|^{-m} I \le X^m \le \|Q\|^m I, \quad m > 0,$$
$$\|Q\|^m I \le X^m \le \|R^{-1}\|^{-m} I, \quad m < 0.$$

First assume that $\widetilde{m} > 0$. Then it follows that

$$0 \le X^{\widetilde{m}} - X^m \le \|Q\|^{\widetilde{m}} I - \|R^{-1}\|^{-m} I, \quad \widetilde{m} > m > 0,$$
$$0 \le X^{\widetilde{m}} - X^m \le \|R^{-1}\|^{-\widetilde{m}} I - \|Q\|^m I, \quad 0 > \widetilde{m} > m.$$

Interchanging the role of $\widetilde{m}$ and m gives

$$0 \le X^m - X^{\widetilde{m}} \le \|Q\|^m I - \|R^{-1}\|^{-\widetilde{m}} I, \quad m > \widetilde{m} > 0,$$
$$0 \le X^m - X^{\widetilde{m}} \le \|R^{-1}\|^{-m} I - \|Q\|^{\widetilde{m}} I, \quad 0 > m > \widetilde{m}.$$

Hence for all $X \in [R, Q]$ and all $m, \widetilde{m} \in (-\infty, 1] \cup [2, \infty)$ we have

$$\|X^{\widetilde{m}} - X^m\| \le \max\{|\|Q\|^{\widetilde{m}} - \|R^{-1}\|^{-m}|, |\|R^{-1}\|^{-\widetilde{m}} - \|Q\|^m|\}.$$

This proves the following results.

THEOREM 4.10. [illegible] $m \le 1$ [illegible] $Q \in \mathcal{P}(\mathcal{H})$ $A \in \mathcal{L}(\mathcal{H})$ [illegible] $R \in \mathcal{P}(\mathcal{H})$ [illegible] (4.7) [illegible] $|m|\|R^{-1}\|^{1-m}\|A\|^2 < 1$ [illegible] $\varepsilon > 0$ [illegible] $E(X) = X^{\widetilde{m}} - X^m, \widetilde{m} \le 1$ [illegible] $\widetilde{m}$ [illegible] m, [illegible] $|\widetilde{m}|\|A\|^2 < 1$, [illegible] (4.11)

$$\|A\|^2 \max\{|\|Q\|^{\widetilde{m}} - \|R^{-1}\|^{-m}|, |\|R^{-1}\|^{-\widetilde{m}} - \|Q\|^m|\} < \min\left\{\frac{1 - |m|\|R^{-1}\|^{1-m}\|A\|^2}{3}\varepsilon, \frac{1 - |\widetilde{m}|\|R^{-1}\|^{1-\widetilde{m}}\|A\|^2}{3}\varepsilon, 1\right\},$$

[illegible]

$$\|X_S - \widetilde{X}_S\| < \varepsilon$$

[illegible] X_S [illegible] (4.2) [illegible] Ω_2 [illegible] $\widetilde{X}_S$ [illegible] (4.12) [illegible] Ω_2.

THEOREM 4.11. [illegible] $m \ge 2$ [illegible] $Q \in \mathcal{P}(\mathcal{H})$ $A \in \mathcal{L}(\mathcal{H})$ [illegible]

$$|m|\|Q\|^{m-1}\|A\|^2 < 1$$

[illegible] $R \in \mathcal{P}(\mathcal{H})$ [illegible] (4.7) [illegible] $\varepsilon > 0$ [illegible] $E(X) = X^{\widetilde{m}} - X^m, \widetilde{m} \ge 2,$ [illegible] $|\widetilde{m}|\|A\|^2 < 1$, [illegible] (4.11)

$$\|A\|^2 \max\{|\|Q\|^{\widetilde{m}} - \|R^{-1}\|^{-m}, |\|R^{-1}\|^{-\widetilde{m}} - \|Q\|^m|\} < \min\left\{\frac{1 - |m|\|Q\|^{m-1}\|A\|^2}{3}\varepsilon, \frac{1 - |\widetilde{m}|\|Q\|^{\widetilde{m}-1}\|A\|^2}{3}\varepsilon, 1\right\},$$

$$\|X_S - \widetilde{X}_S\| < \varepsilon$$

X_S (4.2) Ω_2 $\widetilde{X}_S$ (4.12) Ω_2.

REFERENCES

[1] A. Ambrostetti and G. Prodi, *A Primer of Nonlinear Analysis*, Cambridge Stud. Adv. Math. 34, Cambridge University Press, Cambridge, UK, 1995.

[2] R. Bhatia and K. B. Sinha, *Variation of real powers of positive operators*, Indiana Univ. Math. J., 43 (1994), pp. 913–925.

[3] D. A. Bini, G. Latouche, and B. Meini, *Solving nonlinear matrix equations arising in tree-like stochastic processes*, Linear Algebra Appl., 366 (2003), pp. 39–64.

[4] M. Cheng and S. Xu, *Perturbation analysis of the Hermitian positive definite solution of the matrix equation* $X - A^*X^{-2}A = I$, Linear Algebra Appl., 394 (2005), pp. 39–51.

[5] 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.

[6] P. M. Gahinet, A. J. Laub, C. S. Kenney, and G. A. Hewer, *Sensitivity of the stable discrete-time Lyapunov equation*, IEEE Trans. Automat. Control, 35 (1990), pp. 1209–1217.

[7] A. R. Ghavimi and A. J. Laub, *Residual bounds for discrete-time Lyapunov equations*, IEEE Trans. Automat. Control, 40 (1995), pp. 1244–1249.

[8] V. I. Hasanov and I. G. Ivanov, *Solutions and perturbation estimates for the matrix equations* $X \pm A^*X^{-n}A = Q$, Appl. Math. Comput., 156 (2004), pp. 513–525.

[9] V. I. Hasanov and I. G. Ivanov, *Positive definite solutions of the equation* $X + A^*X^{-n}A = I$, in Numerical Analysis and Its Applications, Lecture Notes in Comput. Sci., 1988, Springer, Berlin, 2001, pp. 377–384.

[10] I. G. Ivanov, V. I. Hasanov, and B. V. Minchev, *On matrix equations* $X \pm A^*X^{-2}A = I$, Linear Algebra Appl., 326 (2001), pp. 27–44.

[11] I. G. Ivanov, B. V. Minchev, and V. I. Hasanov, *Positive definite solutions of the equation* $X - A^*\sqrt{X^{-1}}A = I$, in Applications of Mathematics in Engineering, Heron Press, Sofia, Bulgaria, 1999, pp. 113–116.

[12] L. V. Kantorovich and G. P. Akilov, *Functional Analysis*, 2nd ed., Pergamon Press, Elmsford, NY, 1982.

[13] M. M. Konstantinov, D. Gu, V. Mehrmann, and P. Petkov, *Perturbation Theory of Matrix Equations*, Elsevier, New York, 2003.

[14] M. A. Krasnosel'skiĭ, G. M. Vaĭnikko, P. P. Zabreĭko, Ya. B. Rutitskiĭ, and V. Ya. Stetsenko, *Approximate Solution of Operator Equations*, Wolters-Noordhoff Publishing, Groningen, The Netherlands, 1972.

[15] P. Lancaster and L. Rodman, *Algebraic Riccati Equations*, Oxford University Press, New York, 1995.

[16] X.-G. Liu and H. Gao, *On the positive definite solutions of the matrix equations* $X^s \pm A^TX^{-t}A = I_n$, Linear Algebra Appl., 368 (2003), pp. 83–97.

[17] A. C. M. Ran and S. M. El-Sayed, *On an iteration method for solving a class of nonlinear matrix equations*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 632–645.

[18] A. C. M. Ran and M. C. B. Reurings, *A fixed point theorem in partially ordered sets and some applications to matrix equations*, Proc. Amer. Math. Soc., 132 (2004), pp. 1435–1443.

[19] A. C. M. Ran and M. C. B. Reurings, *On the nonlinear matrix equation* $X + A^*F(X)A = Q$*: Solutions and perturbation theory*, Linear Algebra Appl., 346 (2002), pp. 15–26.

[20] A. C. M. Ran and M. C. B. Reurings, *A nonlinear matrix equation connected to interpolation theory*, Linear Algebra Appl., 379 (2004), pp. 289–302.

[21] M. C. B. Reurings, *Contractive maps on normed spaces and their applications to nonlinear matrix equations*, to appear.

[22] M. C. B. Reurings, *Symmetric Matrix Equations*, Ph.D. thesis, Vrije Universiteit, Amsterdam, 2003.

[23] L. A. Sakhnovich, *Interpolation Theory and Its Applications*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 1, pp. 105–117

EIGENVALUE PROBLEMS IN FIBER OPTIC DESIGN*

LINDA KAUFMAN†

Abstract. Maxwell's equation for modeling the guided waves in a circularly symmetric fiber leads to a family of partial differential equation–eigenvalue systems. Incorporating the boundary condition into a discretized system leads to an eigenvalue problem which is nonlinear in only one element. In fiber design one would like to determine the index profile which is involved in Maxwell's equation so that certain optical properties, which sometimes involve derivatives of the eigenvalues, are satisfied. This contribution discusses how to handle the nonlinear eigenvalue problem and how to determine derivatives of the eigenvalue problem.

Key words. eigenvalue, nonlinear, Rayleigh quotient

AMS subject classifications. 15A18, 3FP30, 3FQ60

DOI. 10.1137/S0895479803432708

1. Introduction.

In this paper we consider several computation issues arising in fiber optic modeling. Assuming a fiber is perfectly straight, circular, and uniform along its length, then Maxwell's equations for guided waves of the fiber can be reduced to a family in m of problems of the form

$$\left(\frac{1}{r}\frac{d}{dr}\left(r\frac{dx}{dr}\right)+\omega^2\eta^2(r,\omega)-\frac{m}{r^2}\right)x=\beta^2 x. \tag{1.1}$$

The index of refraction profile $\eta(r,\omega)$ is in some regions a piecewise constant function, as in Figure 1.1, and can be parameterized by several design parameters relating to the widths and heights of each region. In (1.1), ω is a specified frequency and m is a specified mode number. The finite element method converts this family of differential equations to a family of symmetric tridiagonal eigenvalue problems in ω and m

$$A(\omega,m)x=\mu x, \tag{1.2}$$

where one wishes to find the positive eigenvalues and their corresponding eigenvectors. The eigenvalue μ corresponds to β^2 in (1.1). To simplify our notation we let A represent one member of this family. Usually the waveform is truncated at some radius beyond the core of the fiber, and the boundary condition is expressed as the mth order modified Bessel function of the second kind [9]. This changes the eigenvalue problem in (1.2) to the form

$$(A+s(\mu)e_ne_n^T)x=\mu x, \tag{1.3}$$

where $s(\mu)$ involves the appropriate Bessel functions, A is an $n\times n$ matrix, and e_n is the last column of the $n\times n$ identity matrix. In section 2, we describe several algorithms that can be used to solve (1.3).

As explained in [6], for a particular index profile $\eta(r,\omega)$, one is interested in various integrals of the modes (the eigenvectors), the dispersion

$$\frac{\partial^2\beta^2}{\partial\lambda\partial\omega} \tag{1.4}$$

*Received by the editors August 1, 2003; accepted for publication (in revised form) by B. T. Kägström September 24, 2005; published electronically March 17, 2006.
http://www.siam.org/journals/simax/28-1/43270.html
†Computer Science Department, William Paterson University, Wayne, NJ 07470 (KaufmanL@wpunj.edu).

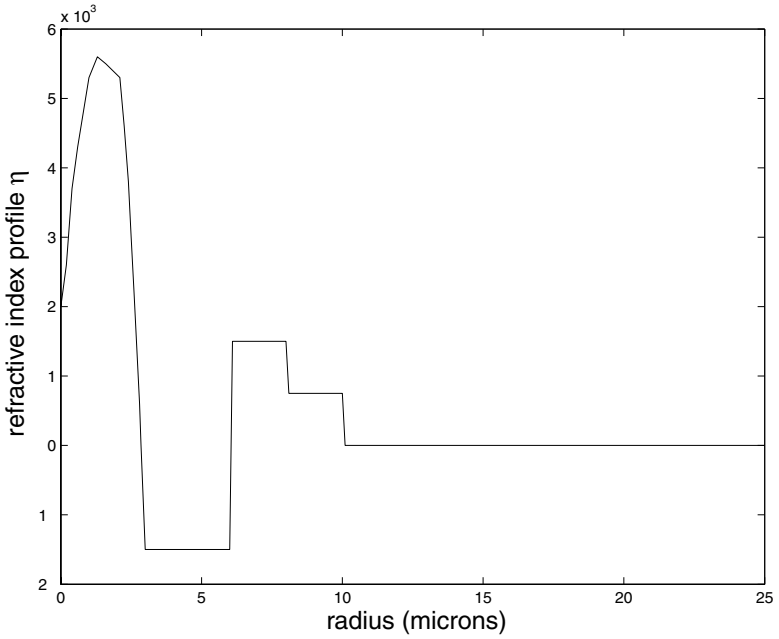


FIG. 1.1. *A typical index of refraction profile* $\eta(r, \omega)$.

and the dispersion slope

$$\frac{\partial^3 \beta^2}{\partial \lambda \partial \omega^2}, \tag{1.5}$$

where λ is the excitation energy and $2\pi c = \lambda\omega$ with the frequency ω is measured in radians per second, and c is the speed of light. More to the point, one is interested in finding the index profile η which might produce certain values of the dispersion and the dispersion slope at particular values of m and ω. If the index profile η is partially defined by a parameter p, then to use an optimization package to determine p typically requires derivatives such as

$$\frac{\partial^4 \beta^2}{\partial \lambda \partial \omega^2 \partial p}. \tag{1.6}$$

Numerical differentiation of these quantities has proved unsatisfactory as small changes in a variable are not necessarily accurately reflected in the computed dispersion. Thus, analytic differentiation is required. In section 4, we review formulas for derivatives of eigenvalues and indicate how the amount of computation can be reduced by noticing common subexpressions.

Note that by defining $\eta(r, \omega)$ we are actually specifying the chemical composition of the radial layers for a fiber. A fiber suitable for, say, underwater transmission would not be appropriate for a local area network or to splice into an existing network to restore a degraded signal. Also various manufacturers tend to emphasize different optical properties for a given application. When using an optimizer to determine $\eta(r, \omega)$, for each function evaluation one may have to solve up to 30 nonlinear eigenvalue problems for various values of m and ω in (1.1). Because there are sometimes multiple local minima to a particular optimization problem, the optimizer is often called several times with different starting guesses to find a design that is manufacturable and

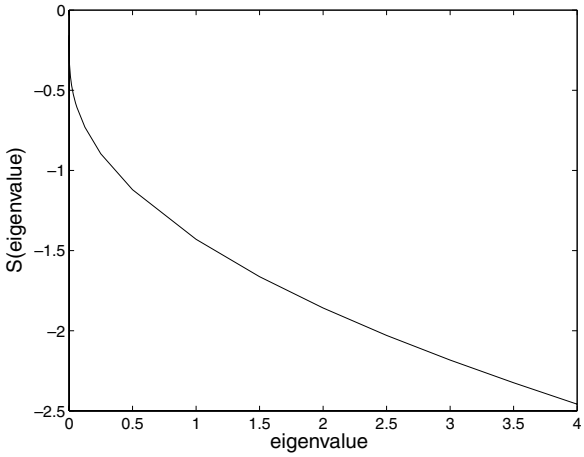


FIG. 2.1. *The function $s(\mu)$ for varying eigenvalue μ.*

more suitable. A production run may require solving thousands of nonlinear eigenvalue problems and then finding derivatives of the computed eigenvalues with respect to the parameters that define the matrix.

2. The nonlinear problem. In this section three methods are considered for finding the positive eigenvalues and their corresponding eigenvectors of the nonlinear eigenvalue problem (1.3). For $m = 0$, $s(\mu)$ is displayed in Figure 2.1. Usually only one or two eigenvalues are positive.

The first method is a simple Picard iteration.

Picard iteration
Let $B = A + s(0)e_n e_n^T$
Until convergence
    Find μ, a specific positive eigenvalue of B.
    Reset B to $A + s(\mu)e_n e_n^T$

The matrix B is also tridiagonal and differs from A only in its bottom right element. The eigenvalues of B were determined using the bisection code in EISPACK [12] modified as in Kaufman [7] so that the inertia of the complete matrix was not computed when the inertia of only part of the matrix sufficed. However, because there were such good guesses for the eigenvalues and eigenvectors from previous iterations of the optimization procedure it was much more efficient to use a Rayleigh quotient iteration to determine μ.

Linear Rayleigh quotient algorithm for a fixed B
(in the inner loop of the Picard iteration)
Determine a lower bound μ_l and an upper bound μ_u for the desired eigenvalue μ.
Set μ_0 to μ_l.
Set $x = e$, the vector of all 1's or a good guess if available.
Set k to 0.
Until convergence
    Solve $(B - \mu_k I)y = x$ and determine the inertia of $(B - \mu_k I)$
    If the inertia claims that μ_k is less than μ, reset μ_l to μ_k
    If the inertia claims that μ_k is greater than μ, reset μ_u to μ_k
    Set $\alpha = y^T By/(y^T y)$

TABLE 2.1
The growth of $|\, s'(\mu)\,|$ *as* μ *approaches zero.*

the eigenvalue μ	$\mid s'(\mu) \mid$
2.0000	0.3631
1.0000	0.5219
0.5000	0.7559
0.2500	1.1054
0.1250	1.6358
0.0625	2.4543

If $\alpha < \mu_l$, set $\mu_{k+1} = .9\mu_l + .1\mu_u$
If $\alpha > \mu_u$, set $\mu_{k+1} = .9\mu_u + .1\mu_l$
If $\mu_l \le \alpha \le \mu_u$ set $\mu_{k+1} = \alpha$
set $x = y/(y^T y)$
increment k.

In the algorithm outlined above, the inertia of the system and the solution y can be determined using the Bunch algorithm for tridiagonal systems [1]. The choice of .9 and .1 were arbitrary and just a heuristic.

Although the inner Rayleigh quotient iteration has cubic convergence, the outer iteration is a fixed point scheme and is generally linear convergent. If the outer iteration is as $\mu = f(\mu)$, then $f(\mu)$ is an eigenvalue of the B matrix. The rate of the convergence, and whether it converges, depends on $\frac{df}{d\mu}$. Let x be the eigenvector corresponding to f. Then $fx = (A + s(\mu)e_n e_n^T)x$, which implies $\frac{df}{d\mu}x = \frac{ds}{d\mu}e_n e_n^T x$, giving

$$\frac{df}{d\mu} = \frac{ds}{d\mu}x_n^2/(x^T x). \tag{2.1}$$

Because $x_n^2/(x^T x) < 1$, convergence is guaranteed as long as $|\frac{ds}{d\mu}| < 1$. For our fiber optic problem, Table 2.1 gives $\frac{ds}{d\mu}$ for various values of μ.

Thus as μ approaches zero, convergence is dubious.

In practice a major problem with the Picard iteration approach is speed. One potentially spends time in the inner loop of the Rayleigh quotient iteration for the wrong problem. This pitfall is especially noticeable for problems where the eigenvalue is near zero and the function $s(\mu)$ is rather steep, as in the example at the end of this section.

One way around the sluggishness of the Picard iteration is to mimic the derivation of the Rayleigh quotient iteration for the nonlinear problem. In the Rayleigh quotient iteration for a matrix A, the next iterate μ_{k+1} is chosen so that

$$\|(A - \mu I)y\|_2 \tag{2.2}$$

is minimized, which occurs when $2y^T Ay\mu + \mu^2$ is minimized or when $\mu = y^T Ay/y^T y$. For the nonlinear Rayleigh quotient case one would like to minimize

$$\|(A + s(\mu)e_n {e_n}^T - \mu I)y\|_2. \tag{2.3}$$

If $z = (A + s(\mu)e_n {e_n}^T)y$, then the required μ minimizes $z^T z - 2\mu z^T y + \mu^2 y^T y$, which occurs when

$$\mu = \frac{z^T y - z^T z'}{y^T y - {z'}^T y}. \tag{2.4}$$

Since $z' = s'(\mu)e_n e_n{}^T y = s'(\mu)y_n e_n$, from (2.4)

$$\mu = \frac{z^T y - s'(\mu)y_n z_n}{y^T y - s'(\mu){y_n}^2}. \tag{2.5}$$

With the linear Rayleigh quotient algorithm given above, the bounds on the eigenvalue estimates ensure that the algorithm is converging to the correct eigenvalue. The bounds can be obtained by examining the Bunch–Parlett decomposition [2], which dictates that one can find a lower triangular matrix L and a block diagonal matrix D with 1×1 blocks and 2×2 blocks such that

$$B - \mu I = LDL^T. \tag{2.6}$$

If D has i 2×2 blocks and j positive 1×1 blocks, then B has $i + j$ eigenvalues larger than μ. In a nonlinear Rayleigh quotient iteration one would like the same bounding mechanism to be preserved even if the element $B_{n,n}$ is changed in each iteration. The following theorem treats the largest eigenvalue and a similar proof can be used to handle other eigenvalues.

THEOREM 2.1. [illegible] $s(\mu)$ [illegible] μ_u [illegible] $B_0 = A + s(\mu_0)e_n e_n{}^T$ [illegible] μ_0 [illegible] $\mu_{\max}$ [illegible] $(A + s(\mu)e_n e_n{}^T)x = \mu x$ [illegible]. If $\mu_0 \geq \mu_{\max}$, the theorem is trivially true. So assume $\mu_{\max} > \mu_0$. Let D be the diagonal matrix in the Bunch–Parlett decomposition [2] of $B_0 - \mu_u I$. If μ_u is an upper bound of the largest eigenvalue of B_0, then D is diagonal with negative elements. Let $\tilde{B} = A + s(\mu_{\max})e_n e_n{}^T$, where $\mu_{\max}$ is the solution to the nonlinear problem, and let $\tilde{D}$ be the diagonal matrix in the Bunch–Parlett decomposition of $\tilde{B} - \mu_u I$.

The first $n-1$ diagonal elements of D and $\tilde{D}$ are identical and $\tilde{d}_n = d_n + s(\mu_{\max}) - s(\mu_0)$. If $\mu_{\max} > \mu_0$, then $s(\mu_{\max}) < s(\mu_0)$, so $\tilde{d}_n$ is negative and μ_u is an upper bound for the nonlinear problem. □

THEOREM 2.2. [illegible] $s(\mu)$ [illegible] μ_l [illegible] $B_0 = (A + s(\mu_0)e_n e_n{}^T)$ [illegible] $\mu_{\max}$ [illegible] $(A + s(\mu)e_n e_n{}^T)x = \mu x$ [illegible] $\mu_{\max} \geq 0$ [illegible]. Let D be the block diagonal matrix in the Bunch–Parlett decomposition of $B_0 - \mu_l I$. If there is a 2×2 block above the bottom two rows or if there is a positive element on the diagonal above the bottom right corner element, then no matter how s is changed, D will signal that μ_l always is a lower bound for the largest eigenvalue. Thus, one may assume that d_n is positive. Moreover, if $\mu_{\max} \geq \mu_0$, the theorem is also trivially true so assume $\mu_{\max} < \mu_0$. If $\tilde{D}$ is the diagonal matrix in the Bunch–Parlett decomposition of $A + s(\mu_{\max})e_n e_n{}^T$, where $\mu_{\max}$ is the solution to the nonlinear problem, then $\tilde{d}_n = d_n + s(\mu_{\max}) - s(\mu_0)$. If $0 \leq \mu_{\max} < \mu_0$, then $s(\mu_{\max}) > s(\mu_0)$, so $\tilde{d}_n$ will be positive and μ_u will be a lower bound of the nonlinear problem. □

From (2.5) one gets the following algorithm.

Nonlinear Rayleigh quotient

Determine a lower bound μ_l and an upper bound μ_u for the desired eigenvalue μ.
Set μ_0 to μ_l.
Set $x = e$, the vector of all 1's if a good guess is not available.
Set k to 0.
Until convergence

Set $B = A + s(\mu_k)e_n e_n^T$.
Solve $(B - \mu_k I)y = x$ and determine the inertia of $(B - \mu_k I)$
If the inertia claims that μ_k is less than μ, reset μ_l to μ_k
If the inertia claims that μ_k is greater than μ, reset μ_u to μ_k
Set $z = By$.
Set $\alpha = \frac{z^T y - s'(\mu) y_n z_n}{y^T y - s'(\mu) {y_n}^2}$
If $\alpha < \mu_l$, set $\mu_{k+1} = .9\mu_l + .1\mu_u$
If $\alpha > \mu_u$, set $\mu_{k+1} = .9\mu_u + .1\mu_l$
If $\mu_l <= \alpha <= \mu_u$ set $\mu_{k+1} = \alpha$
set $x = y/(y^T y)$
increment k.

A third method, suggested by Cowsar [4], is designed to find a root of the function

$$f(\gamma) = \gamma - s(\mu(\gamma)), \tag{2.7}$$

where μ satisfies

$$(A + \gamma e_n e_n^T - \mu I)x = 0.$$

Newton's method for determining the root of (2.7) determines the new approximation of γ_{new} using the formula

$$\gamma_{\text{new}} = \gamma - \frac{\gamma - s(\mu(\gamma))}{\frac{df}{d\gamma}}. \tag{2.8}$$

Note that

$$\frac{df}{d\gamma} = 1 - \frac{ds}{d\mu}\frac{d\mu}{d\gamma}$$

and if $\mu x = (A + \gamma e_n e_n^T)x$, then

$$\frac{d\mu}{d\gamma}x = e_n e_n^T x.$$

Because x can be chosen such that $x^T x = 1$,

$$\frac{d\mu}{d\gamma} = (x^T e_n)^2.$$

Thus each iterate of Newton's method for minimizing (2.7) would set the new approximation of γ_{new} to

$$\gamma_{\text{new}} = \gamma - \frac{\gamma - s(\mu(\gamma))}{1 - \frac{ds}{d\mu}(x^T e_n)^2}, \tag{2.9}$$

where x is the normalized eigenvector corresponding to the eigenvalue μ of $A + \gamma e_n e_n^T$.

Because Newton's method is not globally convergent, it is wise to determine initially an interval containing the root and to use a bisection technique if the new iterate does not fall within the interval. If one is seeking a positive eigenvalue μ, then 0 is an upper bound for γ because $f(0) > 0$ since $s(\mu)$ is negative if μ is positive. A lower bound for γ can be found if the Bunch–Parlett decomposition yields a diagonal

matrix D whose first $n-1$ diagonal elements are negative but whose right corner element d_n is positive. In this case the matrix has one nonnegative eigenvalue and if $\gamma = -d_n$, then the largest eigenvalue of $A + \gamma e_n e_n^T$ is 0. For $m = 0$, $s(0) = 0$, which means $f(-d_n) < 0$. If one of the first $n-1$ elements of the D matrix in the Bunch–Parlett decomposition is negative, then any very large negative number can be used as a lower bound for γ.

In the formal definition of Newton's method given below, to find one positive eigenvalue of the nonlinear eigenvalue problem an upper and a lower bound are first determined and subsequently the bounds are adjusted so that the root of f is always bracketed. If the Newton step is outside the bound, then a step that is 90% of the distance to the bound is taken. If the chosen step does not decrease the norm of f, the step size is reduced by 4. To determine several eigenvalues, the algorithm is used first to determine the largest eigenvalue, the bounds are adjusted to find subsequent eigenvalues, and the linear Rayleigh quotient solver is ask to find a specific numbered eigenvalue.

Safeguarded Newton's method

Set γ_u, an upper bound for γ to 0.
Find γ_l, a lower bound for γ as follows:
    Determine D, the diagonal matrix of the Bunch–Parlett decomposition of A.
    If none of the elements of D are positive, there is no positive eigenvalue.
    If one of the first $n-1$ elements of D is greater than 0, set γ_l to a very large negative number else set $\gamma_l = -d_n$
If an initial eigenvalue μ_0 is available, set $\gamma_0 = s(\mu_0)$, otherwise set $\gamma_0 = \gamma_u$;
Set k to 0.
Until convergence
    Find $\mu > 0$ and x such that $(A + \gamma_k e_n e_n^T)x = \mu x$ using the linear Rayleigh quotient algorithm
    Set $f = \gamma_k - s(\mu); r = abs(f)$
    If $f > 0$
        if $\gamma_k \le \gamma_u$ set $\gamma_u = \gamma_k$
    If $f < 0$
        if $\gamma_k \ge \gamma_l$ set $\gamma_l = \gamma_k$
    if ($k > 1$ and $r > minr$)
        Set $\gamma_{\text{change}} = \gamma_{\text{change}}/4$
        Set $\gamma_{k+1} = \gamma_k + \gamma_{\text{change}}$
    else
        Set $minr$ to r
        Set $\gamma_{k+1} = \gamma_k - \frac{\gamma_k - s(\mu)}{1 - \frac{ds}{d\mu}(x^T e_n)^2}$
        If $\gamma_{k+1} < \gamma_l$, set $\gamma_{k+1} = .9\gamma_l + .1\gamma_u$
        If $\gamma_{k+1} > \gamma_u$, set $\gamma_{k+1} = .9\gamma_u + .1\gamma_l$
        Set $\gamma_{\text{change}} = \gamma_{k+1} - \gamma_k$
        increment k

3. Computational experiments. To indicate the effectiveness of the three algorithms outlined above, we show their effectiveness on several problems. The problem in section 3.1 is a small example with a tunable parameter which indicates that the Picard iteration and the nonlinear Rayleigh quotient may converge linearly and slowly. The second example in section 3.2 is derived from a fiber optics problem

TABLE 3.1
The course of the Picard iteration, the nonlinear Rayleigh quotient, and the Newton approach with $z = .5$ in (3.1).

Picard iteration		Nonlinear Rayleigh		Newton	
iterate	inner iterations	iterate	inner iterations	derived eigenvalue	inner iterations
.3589622	4	.2865810	1	.3589622	4
.3589538	2	.3577595	1	.3589538	2
.3589538	1	.3589538	1	.3589538	1
		.3589538	1		

in the literature.

3.1. A small example. In this section we consider the following small problem:

$$a_{i,i} = \begin{cases} -2 + z, & 1 \le i \le 5, \\ -2 - z, & i = 6, 7, \\ -2, & 8 \le i \le 12, \end{cases} \tag{3.1}$$

and for $1 \le i \le 11$

$$a_{i+1,i} = a_{i,i+1} = (i + .5)/\sqrt{(i + 1) \times (i + 2)}. \tag{3.2}$$

The function $s(\mu)$ involved computing the modified Bessel function of the second kind calculated using ACM Algorithm 484 [3].

For $z = .5$, the largest eigenvalue is about .359, where $s(\mu)$ is not very steep, and all three methods behaved rather well, as Table 3.1 indicates. The rate of convergence for the Picard iteration is based on $\frac{ds}{d\mu} x_n^2/(x^T x)$. For this value of z, as the algorithm converged, $\frac{ds}{d\mu}$was about -1.08, x_n was about $-.00242$, so that $\frac{ds}{d\mu} x_n^2/(x^T x) \approx -6.36 \times 10^{-6}$, suggesting fast convergence. Note for Newton's method the eigenvalue corresponding to γ is given.

For $z = .081$, the largest eigenvalue is about 3.2×10^{-4}, where the s function is rather steep. For this value of z, as the Picard algorithm converged, $\frac{ds}{d\mu}$was about -3.23, x_n was about $-.047$, so that $\frac{ds}{d\mu} x_n^2/(x^T x) \approx -5.56 \times 10^{-3}$, which is approximately the square root of the previous example. The Picard iteration algorithm required 9 outer iterations and 42 inner iterations, which is far more than 14 required by the nonlinear Rayleigh quotient and the 13 required by the Newton approach. The course of the algorithms is shown in Table 3.2. Note that the first three iterations of the nonlinear Rayleigh quotient iteration produced iterates that were negative and then were forced to be positive.

Results like those obtained in Table 3.2 for the Picard iteration have stimulated the search for other algorithms that might not have linear convergence. It was hoped that the nonlinear Rayleigh quotient algorithm would be cubically convergent and one could mimic the proof of cubic convergence given in Parlett [11], but when an eigenvalue approached zero, one could not easily bound $\frac{ds}{d\mu}$. Moreover, problems like that given in (3.2) suggested that the algorithm was linearly convergent. Table 3.3 shows the rate of convergence of the Picard and nonlinear Rayleigh quotient iteration. In the table $\hat{\mu}$ refers to the solution.

3.2. A fiber optics problem. The nonlinear Rayleigh quotient algorithm and the Newton approach were also applied to a more realistic problem that was posed by Lenahan and Friedrichsen [10] for $m = 1$ involving a core region of silicon dioxide

TABLE 3.2
The course of the Picard iteration, the nonlinear Rayleigh quotient, and the Newton approach with $z = .081$ in (3.1).

Picard iteration		Nonlinear Rayleigh		Newton	
iterate	inner iterations	iterate	inner iterations	derived eigenvalue	inner iterations
$.6770570\times10^{-3}$	5	1.0	1	$.6770570\times10^{-3}$	5
$.2831247\times10^{-3}$	11	.1	1	$.3187768\times10^{-3}$	3
$.3294611\times10^{-3}$	3	.01	1	$.3229664\times10^{-3}$	3
$.3219722\times10^{-3}$	3	$.8701964\times10^{-4}$	1	$.3229683\times10^{-3}$	2
$.3231226\times10^{-3}$	2	$.3682030\times10^{-3}$	1		
$.3229445\times10^{-3}$	2	$.3161693\times10^{-3}$	1		
$.3229720\times10^{-3}$	2	$.3240255\times10^{-3}$	1		
$.3229677\times10^{-3}$	7	$.3228049\times10^{-3}$	1		
$.3229684\times10^{-3}$	2	$.3229936\times10^{-3}$	1		
$.3229683\times10^{-3}$	4	$.3229644\times10^{-3}$	1		
$.3229683\times10^{-3}$	1	$.3229689\times10^{-3}$	1		
		$.3229682\times10^{-3}$	1		
		$.3229683\times10^{-3}$	1		
		$.3229683\times10^{-3}$	1		

TABLE 3.3
Rates of convergence with $z = .081$.

Picard iteration	Nonlinear Rayleigh
$(\mu_k - \hat{\mu})/(\mu_{k-1} - \hat{\mu})$	$(\mu_k - \hat{\mu})/(\mu_{k-1} - \hat{\mu})$
$-.113$	.100
$-.163$	.097
$-.153$	$-.024$
$-.155$	$-.192$
$-.154$	$-.150$
$-.155$	$-.155$
$-.162$	$-.155$
$-.167$	$-.155$
	$-.154$
	$-.154$
	$-.167$

doped with germanium surrounded by a region of pure silicon dioxide. The index profile for the core region had the form

$$\eta(r,\omega) = e(\omega) + C(r)h(\omega, sign(C(r)), \tag{3.3}$$

where $e(\omega)$is the index of pure silicon dioxide, $C(r)$ denotes the dopant concentration, and h is a function of the Sellmeier coefficients [5] at ω and the reference frequency. We started with an example where the radius of the fiber R_f and the radius of the core R_c satisfied the relation $R_f = 6R_c$ and $C(r)$ was nonnegative and defined by

$$C(r) = \begin{cases} ((1 - 2\delta(r/R_c)^\alpha)/(1 - 2\delta))^{1/2} - 1), & 0 \le r \le R_c, \\ 0, & r > R_c. \end{cases} \tag{3.4}$$

We solved (3.4) for several values of $\lambda = 2\pi \times 4.0/(R_c\omega)$, δ, and α. In an optimization problem one may wish to vary α and δ and solve the eigenvalue problem for about 20 values of λ. A uniform grid was used in the finite element discretization and R_c was set at 100 units and R_f at 600 units. The matrix eigenvalue problem derived

TABLE 3.4
Inner iteration count for the nonlinear Rayleigh quotient method and the Newton approach on (3.4).

λ	α	δ	eigenvalue	nonlinear RQI	Newton
0.85	25	.003	2.636370×10^{-4}	6	7
1.1	25	.003	4.216316×10^{-6}	10	14
0.85	20	.003	2.525251×10^{-4}	6	7
1.1	20	.003	1.110249×10^{-6}	14	15
0.85	25	.004	5.486164×10^{-4}	5	6
1.1	25	.004	1.046444×10^{-4}	5	8

TABLE 3.5
Inner iteration count for the nonlinear Rayleigh quotient method and the Newton approach on the augmented model with $\lambda = .85$, $\alpha = 25$, and $\delta = .003$.

ρ	σ	eigenvalues	nonlinear RQI	Newton
−0.01	0.005	1.312023×10^{-3}	6	7
		2.824626×10^{-5}	10	19
−0.008	0.005	1.319503×10^{-3}	6	7
		4.442786×10^{-5}	16	20
		4.807390×10^{-6}	4	7
−0.01	0.008	2.366140×10^{-3}	7	8
		7.768107×10^{-4}	17	18
−0.01	0.001	5.740407×10^{-5}	11	15

from this problem involved a tridiagional matrix whose diagonal elements, $a_{i,i}$, were given by

$$a_{i,i} = \begin{cases} -2 + \omega^2(\eta^2 - e(\omega)^2) - (1/i)^2, & 1 \le i \le 100, \\ -2 - (1/i)^2, & 100 < i < 600, \\ -2 - (1/i)^2 + (1 + .5/600) \times 601.5/\sqrt{(600) \times (601)}, & i = 600, \end{cases}$$

and for $1 \le i < 600$

$$a_{i+1,i} = a_{i,i+1} = (i + 1.5)/\sqrt{(i) \times (i + 1)}. \tag{3.5}$$

For the values of λ, α, and δ given in Table 3.4, there was only one positive eigenvalue. With the smaller eigenvalues both algorithms had trouble. For $\lambda = .85$, we used the approximation $e = 1.45291$ and $h = 1.44943$ in (3.3). For $\lambda = 1.1$, we used $e = 1.4969$ and $h = 1.4201$.

We then augmented the model in (3.4) with a layer of fluorine doped silicon dioxide of constant concentration ρ and then a layer of germanium doped silicon dioxide of constant concentration σ. Each layer had the same width as the core region. We kept the same width of the fiber, used the same uniform finite element discretization, and set the wavelength λ at .85, $\alpha = 25$, and $\delta = .003$. As we varied the concentrations, the number of positive eigenvalues changed. For the fluorine layer the dopant concentration was negative and h in (3.3) was 1.456475. In Table 3.5 the number of Rayleigh quotient iterations is given for several values of σ and ρ. In general the nonlinear Rayleigh quotient seemed to be slightly faster than the Newton technique but the differences were rather inconsequential.

Usually when an eigenvalue was not close to zero, the nonlinear Rayleigh quotient algorithm required the same effort as the initial iteration of the Newton approach. In the production code a polyalgorithm was formed that initially used the Rayleigh

quotient algorithm, and if that did not work after a specified number of iterations, its best approximation was used to determine an initial γ in the Newton procedure.

Because we were solving a sequence of eigenvalue problems in which there were often only small changes in the matrix, good starting approximations for eigenvalues and eigenvectors were usually available from previous problems.

4. Analytic derivatives. The function η in (1.1) is defined by a number of parameters $p_k, k = 1, 2 \ldots K$, giving the shape of the initial region and the concentration and width of the various layers of the fiber. Often the parameters p_k are requested such that the dispersion $\frac{\partial^2 \beta^2}{\partial \lambda \partial \omega}$ and the dispersion slope $\frac{\partial^3 \beta^2}{\partial \lambda \partial \omega^2}$ have prescribed values where $\omega = 2\pi c/\lambda$ for specific value of ω and m in (1.1). In a typical optimization problem to find an optimal shape of η, one might have to supply the dispersion for about 30 combinations of ω and m for each function and derivative evaluation. A signal traveling down a fiber tends to spread out over various wavelengths. Traditionally, about every 40 kilometers the signal was restored electrically. The concept of a dispersion compensating fiber for existing fiber is to splice in a small length of fiber with negative dispersion that is specially created to restore the signal so that electrical restoration is not needed. A fiber optics company might also wish to lay new fiber which has zero dispersion.

If $\mu = \beta^2$, one needs first to solve $(A + s(\mu)e_n e_n^T)x = \mu x$ and then to determine the dispersion, $\frac{1}{2\mu^{1/2}} \frac{\partial^2 \mu}{\partial \lambda \partial \omega}$. Moreover, since $\frac{\partial \mu}{\partial \lambda} = -\frac{\partial \mu}{\partial \omega} \frac{2\pi c}{\lambda^2}$, $\frac{\partial^2 \mu}{\partial \omega^2}$ and $\frac{\partial^3 \mu}{\partial \omega^3}$ must be computed to evaluate the dispersion and dispersion slope, respectively.

Let us use the shorthand A', μ', and x' to denote the derivatives of the elements of A, μ, and x with respect to ω, respectively. To simplify our notation let $B = (A + s(\mu)e_n e_n^T)$, $\phi = x^T e_n$, and $\sigma = 1 - s_\mu \phi^2$. The following lemma provides a formula for μ', and x'.

LEMMA 4.1. [illegible] μ [illegible] ω [illegible]

$$\mu' = x^T A' x/\sigma, \tag{4.1}$$

[illegible] x' [illegible] $x'^T x = 0$ [illegible]

$$(\mu I - B)x' = -(\mu I - B)' x. \tag{4.2}$$

[illegible]. Differentiating the equation $Bx = \mu x$ yields

$$((\mu I - A)' - s_\mu \mu' e_n e_n^T)x + (\mu I - B)x' = 0, \tag{4.3}$$

which, when multiplied by x^T, implies $x^T(\mu I - A)' x - s_\mu \mu' \phi^2 = 0$. If x is chosen such that $x^T x = 1$, then

$$\mu' = x^T A' x/\sigma. \tag{4.4}$$

Because s_μ is always negative, $\sigma = 1 - s_\mu \phi^2$ is never zero. From (4.3) and (4.1) one gets (4.2). Unfortunately $(\mu I - B)$ is singular, but if μ is not a multiple eigenvalue of B, the vector x' can be determined uniquely by using the condition $x'^T x = 0$. □

Following the approach given in the proof of Lemma 4.1 it is shown in [8] that the dispersion is given by $-\frac{2\pi c}{\lambda^2}\mu''$, where

$$\mu'' = (x^T A'' x + 2x^T B' x' + \alpha)/\sigma, \tag{4.5}$$

where $\alpha = s_{\mu\mu} \mu'^2 \phi^2$.

Calculating the dispersion slope requires μ'''. If $\rho = \phi\phi((s''(\mu)\mu'^2)' + s''(\mu)\mu''\mu')$ and $\phi' = e_n^T x'$, then one can express μ''' as

$$\mu''' = (x^T A'''x + 6x^T B''x' + 6x'^T B'x' - 6\mu' x'^T x' + \rho)/\sigma. \tag{4.6}$$

Since our aim is to determine the parameters in η that give a prescribed dispersion, the derivatives of the dispersion with respect to these parameters are needed by most optimizers. Let us use the convention that μ_k is the derivative of μ with respect to the kth design parameter, A_k is the derivative of the A matrix with respect to the kth design parameter, x_k is the derivative of x with respect to its kth design parameter, etc. Mimicking the proof of Lemma 4.1 suggests that if $x^T x = 1$, then

$$\mu_k = x^T A_k x/\sigma. \tag{4.7}$$

It is shown in [8] that if θ contains all the terms in $s(\mu)''_k\phi^2$ except the one with μ''_k, then μ''_k can be expressed several ways, including

$$\mu''_k = (x^T A''_k x + 2(x^T B''x_k + x^T B'_k x' + x^T B'x'_k + x_k^T B'x') + \theta)/\sigma, \tag{4.8}$$

where x'_k satisfies $(\mu I - B)x'_k = -(\mu I - B)'_k x - (\mu I - B)'x_k - (\mu I - B)_k x'$ and the condition $x^T x'_k = -x'^T x_k$, and by the equation

$$\mu''_k = (x^T A''_k x + 4x^T B'_k x' + 2x'^T B_k x' + 2x^T B_k x'' + \theta)/\sigma, \tag{4.9}$$

where $(\mu I - B)x'' = -(\mu I - B)''x - 2(\mu I - B)'x'$ and $x^T x'' = -x'^T x'$.

The first formula for μ''_k in (4.8) comes from differentiating (4.5) with respect to the design parameter, but it can be the less efficient approach. The second formula in (4.9) reverses the order of differentiation and comes from twice differentiating with respect to ω the formula for μ_k in (4.7). The formula in (4.8) requires x, x', and for the kth design parameter x_k and x'_k. For 20 design parameters one must solve for 42 vectors. The second formula for μ''_k requires x, x', x''. Thus for 20 design parameters the formula in (4.9) requires 3 vectors. Theoretically, (4.8) requires at least twice as much work as (4.9).

Acknowledgments. I thank Bill Reed, formerly of Bell Labs, for bringing the fiber optics problem to my attention and I thank Lawrence Cowsar of Bell Labs for providing the program for determining a good discretization of the pde and the optical properties of the fiber.

REFERENCES

[1] J. R. Bunch, *Partial pivoting strategies for symmetric matrices*, SIAM J. Numer. Anal., 11 (1974), pp. 521–528.
[2] J. R. Bunch and B. N. Parlett, *Direct methods for solving symmetric indefinite systems of linear systems*, SIAM J. Numer. Anal., 8 (1971), pp. 539–655.
[3] K. H. Burrell, *Algorithm* 484: *Evaluation of the modified Bessel functions K0(Z) and K1(Z)for complex arguments*, Commun. ACM, 17 (1974), pp. 524–526.
[4] L. Cowsar, *personal communication*, 1999.
[5] J. W. Fleming, *Material dispersion in lightguide glasses*, Elect. Lett., 14 (1978), pp. 326–328.
[6] S. V. Kartalopoulous, *Introduction to DWDM Technology*, IEEE Press, Piscataway, NJ, 2000.
[7] L. Kaufman, *An observation on bisection software for the symmetric tridiagonal eigenvalue problem*, ACM Trans. Math. Software, 26 (2000), pp. 520–526.
[8] L. Kaufman, *Calculating Dispersion in Fiber Optics Design*, in preparation.

[9] T. A. Lenahan, *Calculation of modes in an optical fiber using the finite element method and EISPACK*, Bell System Technical J., 62 (1983), pp. 2663–2694.

[10] T. A. Lenahan and H. W. Friedrichsen, *Analysis of Propagation over Single-Mode Optical Fibers Using the Finite Element Method and EISPACK*, Technical Memorandum 82-54541-34, Bell Labs, Atlanta-Norcross, GA, Nov. 1982.

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

[12] B. T. Smith, J. M. Boyle, J. J. Dongarra, B. S. Garbow, Y. Ikebe, V. C. Klema, and C. B. Moler, *Matrix Eigenvalues Routines—Eispack Guide*. Springer-Verlag, Berlin, 1976

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 1, pp. 118–147

TWO VARIABLE ORTHOGONAL POLYNOMIALS AND STRUCTURED MATRICES*

ANTONIA M. DELGADO†, JEFFREY S. GERONIMO‡, PLAMEN ILIEV‡, AND FRANCISCO MARCELLÁN†

Abstract. We consider bivariate real valued polynomials orthogonal with respect to a positive linear functional. The lexicographical and reverse lexicographical orderings are used to order the monomials. Recurrence formulas are derived between polynomials of different degrees. These formulas link the orthogonal polynomials constructed using the lexicographical ordering with those constructed using the reverse lexicographical ordering. Relations between the coefficients in the recurrence formulas are established and used to give necessary and sufficient conditions for the existence of a positive linear functional. Links to the theory of matrix orthogonal polynomials are developed as well the consequences of a zero assumption on one of the coefficients in the the recurrence formulas.

Key words. bivariate orthogonal polynomials, positive linear functionals, moment problem, Hankel matrices.

AMS subject classifications. 42C05, 30E05, 47A57

DOI. 10.1137/05062946X

1. Introduction.

Bivariate orthogonal polynomials have been investigated by many authors. Special examples of these types of polynomials have arisen in studies related to symmetry groups [3], [14], [20], as extensions of one variable polynomials [5], [13], and as eigenfunctions of partial differential equations [12], [17], [11], [19] (see also the references in [4]). The general theory of these polynomials can trace its origins back to [10] and an excellent review of the theory can be found in [4] (see also [21]). A major difficulty encountered in the theory of orthogonal polynomials of more than one variable is which monomial ordering to use. Except for the special cases that have arisen from the subject mentioned above, the preferred ordering is the total degree ordering which is the one set by Jackson [4]. For polynomials with the same total degree the ordering is lexicographical. There is a good reason to use this ordering, which is that if new orthogonal polynomials of higher degree are to be constructed, then their orthogonality relations will not affect the relations governing the lower degree polynomials. This can be seen especially in Xu's vector formulation of the problem [22] (see also [2], [6], and [15], [16]). However, in their work on the Fejér–Riesz factorization problem, Geronimo and Woerdeman [8], [9] noticed that the most useful ordering was the lexicographical ordering and reverse lexicographical ordering. Important in their work were the relations of the orthogonal polynomials in these orderings. The reason for this is that in these orderings the moment matrix is a structured matrix, i.e., it is a block Toeplitz matrix where the blocks are themselves Toeplitz matrices. In the

*Received by the editors April 18, 2005; accepted for publication (in revised form) by H. J. Woerdeman October 3, 2005; published electronically March 17, 2006. The second and fourth authors were partially supported by NATO grant PST.CLG.979738. The second author was partially supported by an NSF grant.

http://www.siam.org/journals/simax/28-1/62946.html

†Departamento de Matemáticas, Universidad Carlos III de Madrid, 28911 Leganés, Spain (adelgado@math.uc3m.es, pacomar@ing.uc3m.es). These authors were partially supported by grant BFM2003-06335-C03-02 from the Dirección General de Investigación, Ministerio de Educación y Ciencia of Spain.

‡School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332–0160 (geronimo@math.gatech.edu, iliev@math.gatech.edu).

one variable case the connection between orthogonal polynomials and the Hankel or Toeplitz matrices associated with them plays an important role in the theory. The coefficients in the recurrence formulas for the orthogonal polynomials give a parameterization of positive definite Hankel or Toeplitz matrices. Furthermore, structured matrices come up in a variety of engineering and physics problems and so the orthogonal polynomials associated with them need to be investigated. The aim of this paper is to study orthogonal polynomials associated with positive definite block Hankel matrices whose entries are also Hankel and to develop methods for constructing such matrices. We proceed as follows. In section 2 we consider finite subspaces of monomials of the form $x^i y^j$, $0 \le i \le 2n$, $0 \le j \le 2m$, and show the connection between positive linear functionals defined on this space and positive doubly Hankel matrices, i.e., block Hankel matrices whose blocks are Hankel matrices. These structured matrices arise when using the lexicographical or reverse lexicographical ordering on the monomials. We then introduce certain matrix orthogonal polynomials and show how they give the Cholesky factors for the doubly Hankel matrix considered above. These polynomials may be thought of as arising from a parameterized moment problem. In section 3 we construct two variable orthogonal polynomials, where the monomials are ordered according to the lexicographical ordering. When these polynomials are organized into vector orthogonal polynomials they can be related to the matrix orthogonal polynomials constructed previously. From this relation it is shown that these vector polynomials are the minimizers of a certain quadratic functional. Using the orthogonality relation, recurrence relations satisfied by the vector polynomials and their counterparts in the reverse lexicographical ordering are derived and some elementary properties of the matrices entering these recurrence relations are deduced. Because of the size and shape of the coefficients in the recurrence formulas they must be related. In section 4 we derive and examine these relations, and in section 5 a number of Christoffel–Darboux-like formulas are derived. In section 6 we use the relations between the coefficients derived in section 4 to develop an algorithm to construct the coefficients in the recurrence formulas at a particular level, (n, m), say, in terms of the coefficients at the previous levels plus a certain number of unknowns. The collection of these unknowns is in one-to-one correspondence with the number of moments needed to construct the vector polynomials up to level (n, m). This is used in section 7 to construct a positive linear functional from the recurrence coefficients. The construction allows us to find necessary and sufficient conditions on the recurrence coefficients for the existence of a positive linear functional which is in one to one correspondence with the set of positive definite "doubly" Hankel matrices. In the above construction an important role is played by a set of matrices that must be contractions. In section 8 we explore the consequences of setting these contractive matrices equal to zero and show that this condition characterizes product measures. Finally in section 9 we give a numerical example for the case $n = 2,\ m = 2$, which illustrates the above algorithm. We also present an example for which the moment problem is not extendable.

2. Positive linear functionals and Hankel matrices. In this section we consider moment matrices associated with the lexicographical ordering which is defined by

$$(k, \ell) <_{\text{lex}} (k_1, \ell_1) \Leftrightarrow k < k_1 \text{ or } (k = k_1 \text{ and } \ell < \ell_1)$$

and the reverse lexicographical ordering defined by

$$(k, \ell) <_{\text{revlex}} (k_1, \ell_1) \Leftrightarrow (\ell, k) <_{\text{lex}} (\ell_1, k_1).$$

Both of these orderings are linear orders and in addition satisfy

$$(k,\ell) < (m,n) \Rightarrow (k+p,\ell+q) < (m+p,n+q).$$

Note that none of these orderings respects the total degree. Denote $\prod^{n,m}(x,y)$ as the span$\{x^i y^j,\ 0 \le i \le n,\ 0 \le j \le m\}$. Let $\mathcal{L}_{n,m}$ be a linear functional defined on $\prod^{2n,2m}(x,y)$ by

$$\mathcal{L}_{n,m}(x^i y^j) = h_{i,j}.$$

We will call $h_{i,j}$ the (i,j) moment of $\mathcal{L}_{n,m}$ and $\mathcal{L}_{n,m}$ a moment functional. If we form the $(n+1)(m+1)\times(n+1)(m+1)$ matrix $H_{n,m}$ for $\mathcal{L}_{n,m}$ in the lexicographical ordering then, as noted in the introduction, it has the special form

$$H_{n,m} = \begin{bmatrix} H_0 & H_1 & \cdots & H_n \\ H_1 & H_2 & & H_{n+1} \\ \vdots & & \ddots & \vdots \\ H_n & H_{n+1} & \cdots & H_{2n} \end{bmatrix}, \tag{2.1}$$

where each H_i is a $(m+1)\times(m+1)$ matrix of the form

$$H_i = \begin{bmatrix} h_{i,0} & h_{i1} & \cdots & h_{i,m} \\ h_{i1} & h_{i2} & \cdots & \\ \vdots & \vdots & & \vdots \\ h_{i,m} & & \cdots & h_{i,2m} \end{bmatrix}, \qquad i = 0,\dots,2n. \tag{2.2}$$

Thus $H_{n,m}$ is a block Hankel matrix where each block is a Hankel matrix so it has a doubly Hankel structure. If the reverse lexicographical ordering is used in place of the lexicographical ordering we obtain another moment matrix $\tilde{H}_{n,m}$ where the roles of n and m are interchanged. We have the following useful lemmas, which characterize doubly Hankel matrices. An analogous characterization of doubly Toeplitz matrices was given in [9].

LEMMA 2.1. [illegible] $k\times k$ [illegible] H^1 [illegible] H [illegible] H [illegible] $H = H^\top$ [illegible] $H^1 = H^{1\top}$ [illegible]. Recall that $H = (h_{i,j}) = (h_{i+j})$ characterizes a Hankel matrix. Thus the necessary conditions of the lemma follow immediately. To prove the converse, note that $H = H^\top$ implies that $h_{i,j} = h_{j,i}$. Since $H^1 = (h^1_{i,j}) = (h_{i+1,j})$, $i = 1,\dots,k-1$, $j = 1,\dots,k-1$, the second condition implies that

$$h_{i+1,j} = h_{j+1,i}.$$

Thus $h_{i+1,j} = h_{i,j+1}$, which completes the result. □

LEMMA 2.2. [illegible] $H = (H_{i,j})$ $i = 1,\dots,k$ $j = 1,\dots,k$ [illegible] $H_{i,j}$ [illegible] $m\times m$ [illegible] H [illegible] $H = H^\top$ $H^1 = H^{1\top}$ $H^2 = H^{2\top}$ [illegible] H^1 [illegible] H [illegible] H^2 [illegible] H [illegible] $H_{i,j}$ [illegible]. Again the necessary conditions follow from the structure of H. To see the converse, note that $H^\top = H$ implies that $H_{i,j} = H^\top_{j,i}$ so that $H_{i,i} = H^\top_{i,i}$. The

condition on H^1 shows that $H_{i+1,j} = H_{i,j+1}$. Thus H is block Hankel with each entry being symmetric. The result now follows from Lemma 2.1. □

REMARK 2.3. The above results are true if the roles of the rows and columns are interchanged.

We say that the moment functional $\mathcal{L}_{n,m} : \prod^{2n,2m} \to \mathbb{R}$ is positive definite if

$$\mathcal{L}_{n,m}(p^2) > 0 \tag{2.3}$$

for all nonzero $p \in \prod^{n,m}$. Likewise, the moment functional $\mathcal{L}_{n,m} : \prod^{2n,2m} \to \mathbb{R}$ is nonnegative definite if $\mathcal{L}_{n,m}(p^2) \geq 0$ for every $p \in \prod^{n,m}$. Note that it follows from a simple quadratic form argument that $\mathcal{L}_{n,m}$ is positive definite or nonnegative definite if and only if its moment matrix $H_{n,m}$ is positive definite or nonnegative definite, respectively.

We will say that $\mathcal{L}$ is positive definite or nonnegative definite if

$$\mathcal{L}(p^2) > 0 \qquad \text{or} \qquad \mathcal{L}(p^2) \geq 0,$$

respectively, for all nonzero polynomials. Again these conditions are equivalent to the moment matrices $H_{n,m}$ being positive definite or nonnegative definite for all positive integers n and m.

From the above remark we easily find the next lemma.

LEMMA 2.4. [illegible] $H_{n,m}$ [illegible] $(n+1)(m+1) \times (n+1)(m+1)$ [illegible] (2.1) [illegible] (2.2) [illegible] $\mathcal{L}_{n,m} : \prod^{2n,2m}(x,y) \to \mathbb{R}$ [illegible] $H_{n,m}$ [illegible]

$$h_{i,j} = \mathcal{L}_{n,m}(x^i y^j), \qquad 0 \leq i \leq 2n, \qquad 0 \leq j \leq 2m.$$

If the positive moment functional $\mathcal{L}_{n,m} : \prod^{2n,2m} \to \mathbb{R}$ is extended to two variable polynomials with matrix coefficients in the obvious way, we can associate to it a positive matrix function $\mathcal{L}_m : \prod_{m+1}^{n}(x) \times \prod_{m+1}^{n}(x) \to M^{m+1,m+1}$ defined by

$$\mathcal{L}_m(P(x), Q(x)) = \mathcal{L}_{n,m}(P(x,y)\ Q^{\top}(x,y)), \tag{2.4}$$

where

$$P(x,y) = P(x) \begin{bmatrix} 1 \\ \vdots \\ y^m \end{bmatrix} \text{ and } Q(x,y) = Q(x) \begin{bmatrix} 1 \\ \vdots \\ y^m \end{bmatrix}.$$

Here, $\prod_{m+1}^{n}(x)$ is the set of all $(m+1) \times (m+1)$ real valued matrix polynomials of degree n or less and $M^{m,n}$ is the space of $m \times n$ matrices. Because of the structure of $H_{n,m}$ we can associate to $\mathcal{L}_m$ matrix valued orthogonal polynomials in the following manner. Let $\{R_i(x)\}_{i=0}^{n}$ and $\{L_i(x)\}_{i=0}^{n}$ be $(m+1) \times (m+1)$ real valued matrix polynomials given by

$$R_i(x) = R_{i,i}x^i + R_{i,i-1}x^{i-1} + \cdots, \qquad i = 0, \ldots, n, \tag{2.5}$$

and

$$L_i(x) = L_{i,i}x^i + L_{i,i-1}x^{i-1} + \cdots, \qquad i = 0, \ldots, n, \tag{2.6}$$

satisfying

$$\mathcal{L}_m(R_i^\top, R_j^\top) = \delta_{ij} I_{m+1} \tag{2.7}$$

and

$$\mathcal{L}_m(L_i, L_j) = \delta_{ij} I_{m+1}, \tag{2.8}$$

respectively, where I_{m+1} denotes the $(m+1)\times(m+1)$ identity matrix. The above relations uniquely determine the sequences $\{R_i\}_{i=0}^n$ and $\{L_i\}_{i=0}^n$ up to a unitary factor and we fix this factor by requiring $R_{i,i}$ to be an upper triangular matrix with positive diagonal entries and $L_{i,i}$ to be a lower triangular matrix also with positive diagonal entries. From the defining equations (2.7) and (2.8) it follows that $R_i^\top = L_i$ hence we will concentrate on L_i. We write

$$L_i(x) = [L_{i,0}\ L_{i,1}\ \cdots\ L_{i,i}\ 0\ \cdots\ 0] \begin{bmatrix} I_{m+1} \\ xI_{m+1} \\ \vdots \\ x^n I_{m+1} \end{bmatrix} \tag{2.9}$$

and

$$L^n(x) = \begin{bmatrix} L_0(x) \\ L_1(x) \\ \vdots \\ L_n(x) \end{bmatrix} = L \begin{bmatrix} I_{m+1} \\ xI_{m+1} \\ \vdots \\ x^n I_{m+1} \end{bmatrix}, \tag{2.10}$$

where

$$L = \begin{bmatrix} L_{0,0} & 0 & \cdots & 0 \\ L_{1,0} & L_{1,1} & \cdots & 0 \\ \vdots & & \ddots & \\ L_{n,0} & & \cdots & L_{n,n} \end{bmatrix}. \tag{2.11}$$

By lower A (respectively, upper B) Cholesky factor of a positive definite matrix M we mean

$$M = AA^\top = BB^\top, \tag{2.12}$$

where A is a lower triangular matrix with positive diagonal elements and B is an upper triangular matrix with positive diagonal elements. With the above we have the following lemma.

LEMMA 2.5. [illegible] $L^\top$ [illegible] $H_{n,m}^{-1}$

[illegible] Note that (2.8) implies that

$$I = \mathcal{L}_m(L^n, L^n) = L\mathcal{L}_m\left(\begin{bmatrix} I_{m+1} \\ xI_{m+1} \\ \vdots \\ x^n I_{m+1} \end{bmatrix}, \begin{bmatrix} I_{m+1} \\ xI_{m+1} \\ \vdots \\ x^n I_{m+1} \end{bmatrix} \right) L^\top = LH_{n,m}L^\top,$$

where I is the $(n+1)(m+1) \times (n+1)(m+1)$ identity matrix. Thus

$$H_{n,m}^{-1} = L^\top L. \qquad \square$$

From this formula and the fact that L is upper triangular we see that $L_{n,n}^\top$ is the upper Cholesky factor of $[0, \dots, I_{m+1}] H_{n,m}^{-1} [0, \dots, I_{m+1}]^\top$. Hence from (2.11) we find

$$L_n(x) = \left[0, 0, \dots, 0, (L_{n,n}^\top)^{-1}\right] H_{n,m}^{-1} [I_{m+1}, xI_{m+1}, \dots, x^n I_{m+1}]^\top. \tag{2.13}$$

The theory of matrix orthogonal polynomials [1], [7], [18] can be applied to obtain the recurrence formula

$$\begin{aligned} xL_i(x) &= A_{i+1,m} L_{i+1}(x) + B_{i,m} L_i(x) + A_{i,m}^\top L_{i-1}(x), \quad i = 0, \dots, n-1, \\ L_{-1} &= 0, \end{aligned} \tag{2.14}$$

where

$$A_{i+1,m} = \mathfrak{L}_m(xL_i, L_{i+1}) = L_{i,i} L_{i+1,i+1}^{-1} \tag{2.15}$$

and

$$B_{i,m} = \mathfrak{L}_m(xL_i, L_i). \tag{2.16}$$

The above equations show that $B_{i,m}$ is an $(m+1) \times (m+1)$ real symmetric matrix and $A_{i,m}$ is an $(m+1) \times (m+1)$ real lower triangular matrix.

Routine manipulations of (2.14) using the fact that $B_{n,m}$ is self-adjoint give

$$\begin{aligned} &L_i^\top(x_1) A_{i+1,m} L_{i+1}(x) - L_{i+1}^\top(x_1) A_{i+1,m}^\top L_i(x) \\ &\quad = (x - x_1) L_i^\top(x_1) L_i(x) + L_{i-1}^\top(x_1) A_{i,m} L_i(x) - L_i^\top(x_1) A_{i,m}^\top L_{i-1}(x), \end{aligned} \tag{2.17}$$

and iteration of this formula to $i = 0$ yields the important Christoffel–Darboux formula.

We note that the same results hold for the reverse lexicographical ordering with x replaced by y and the roles of n and m interchanged.

As in the scalar case, matrix orthogonal polynomials satisfy a minimization principle [7]. Let sym $\mathbb{R}_{m+1}$ be the space of $(m+1) \times (m+1)$ real symmetric matrices and let $\mathfrak{L} : \prod_{m+1}^n \to$ sym $\mathbb{R}_{m+1}$ be given by

$$\mathfrak{L}(Y) = \mathfrak{L}_m(Y, Y) - 2\text{sym } Y_n. \tag{2.18}$$

Here

$$Y(x) = Y_n x^n + \cdots + Y_0 = [Y_0, Y_1, \dots, Y_n][I_{m+1}, xI_{m+1}, \dots, x^n I_{m+1}]^\top$$

and

$$\text{sym } Y_n = \frac{Y_n + Y_n^\top}{2}.$$

The equation (2.18) can be evaluated as

$$\mathfrak{L}(Y(x)) = [Y_0, Y_1, \dots, Y_n] H_{n,m} [Y_0, Y_1, \dots, Y_n]^\top - 2\text{sym } Y_n.$$

Set

$$X = [Y_0, Y_1, \dots, Y_n]H_{n,m}^{1/2} - V_n H_{n,m}^{-1/2},$$

where any square root of $H_{n,m}$ may be used and where $V_n = [0, 0, \dots, I_{m+1}]$. Then (2.18) becomes

$$\mathfrak{L}(Y) = XX^\top - V_n H_{n,m}^{-1} V_n^\top. \tag{2.19}$$

Thus there is a unique $W \in \prod_{m+1}^n$, corresponding to $X = 0$ given by

$$W(x) = V_n H_{n,m}^{-1} [I_{m+1}, xI_{m+1}, \dots, x^n I_{m+1}]^\top, \tag{2.20}$$

that minimizes $\mathfrak{L}$ in the sense that

$$\mathfrak{L}(W) \le \mathfrak{L}(Y) \tag{2.21}$$

for all $Y \in \prod_{m+1}^n$. From formula (2.13) we find

$$L_n(x) = (L_{n,n}^\top)^{-1} W(x). \tag{2.22}$$

3. Lexicographic order and orthogonal polynomials. In this section we examine the properties of two variable orthogonal polynomials where the monomial ordering is either the lexicographical or reverse lexicographical. Given a positive definite linear functional $\mathcal{L}_{N,M} : \prod^{2N,2M} \to \mathbb{R}$ we perform the Gram–Schmidt procedure using the lexicographical ordering and define the orthonormal polynomials $p_{n,m}^l(x,y)$, $0 \le n \le N$, $0 \le m \le M$, $0 \le l \le m$, by the equations

$$\begin{aligned} &\mathcal{L}_{N,M}(p_{n,m}^l, x^i y^j) = 0, \quad 0 \le i < n \text{ and } 0 \le j \le m \text{ or } i = n \text{ and } 0 \le j < l, \\ &\mathcal{L}_{N,M}(p_{n,m}^l, p_{n,m}^l) = 1, \end{aligned} \tag{3.1}$$

and

$$p_{n,m}^l(x,y) = k_{n,m,l}^{n,l} x^n y^l + \sum_{(i,j) <_{\text{lex}} (n,l)} k_{n,m,l}^{i,j} x^i y^j. \tag{3.2}$$

With the convention $k_{n,m,l}^{n,l} > 0$, the above equations uniquely specify $p_{n,m}^l$. Polynomials orthonormal with respect to $\mathcal{L}_{N,M}$ but using the reverse lexicographical ordering will be denoted by $\tilde{p}_{n,m}^l$. They are uniquely determined by the above relations with the roles of n and m interchanged.

Set

$$\mathbb{P}_{n,m} = \begin{bmatrix} p_{n,m}^0 \\ p_{n,m}^1 \\ \vdots \\ p_{n,m}^m \end{bmatrix} = K_{n,m} \begin{bmatrix} 1 \\ y \\ \vdots \\ x^n y^m \end{bmatrix}, \tag{3.3}$$

where the $(m+1) \times [(n+1)(m+1)]$ matrix $K_{n,m}$ is given by

$$K_{n,m} = \begin{bmatrix} k_{n,m,0}^{0,0} & k_{n,m,0}^{0,1} & \cdots & k_{n,m,0}^{n,0} & 0 & \cdots \\ k_{n,m,1}^{0,0} & k_{n,m,1}^{0,1} & \cdots & k_{n,m,1}^{n,0} & k_{n,m,1}^{n,1} & \ddots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots \\ k_{n,m,m}^{0,0} & k_{n,m}^{0,1} & \cdots & \cdots & \cdots & k_{n,m,m}^{n,m} \end{bmatrix}. \tag{3.4}$$

As indicated above denote

$$\tilde{\mathbb{P}}_{n,m} = \begin{bmatrix} \tilde{p}^0_{n,m} \\ \tilde{p}^1_{n,m} \\ \vdots \\ \tilde{p}^n_{n,m} \end{bmatrix} = \tilde{K}_{n,m} \begin{bmatrix} 1 \\ x \\ \vdots \\ x^n y^m \end{bmatrix}, \tag{3.5}$$

where the $(n+1) \times [(n+1)(m+1)]$ matrix $\tilde{K}_{n,m}$ is given similarly to (3.4) with the roles of n and m interchanged. In order to find recurrence formulas for the vector polynomials $\mathbb{P}_{n,m}$ we introduce the inner product

$$\langle X, Y \rangle = \mathcal{L}_{N,M}(XY^\top). \tag{3.6}$$

Let $\prod_{(k)}^{(n,m)}$ be the vector space of k dimensional vectors with entries in $\prod^{n,m}(x,y)$. Utilizing the orthogonality relations (3.1) we see as in the next lemma.

LEMMA 3.1. [illegible] $\mathbb{P} \in \prod_{(k)}^{(n,m)}$ [illegible]

$$\langle \mathbb{P}, x^i y^j \rangle = 0, \quad 0 \le i < n, \quad 0 \le j \le m, \tag{3.7}$$

[illegible] $\mathbb{P} = C\mathbb{P}_{n,m}$ [illegible] C [illegible] $k \times (m+1)$. [illegible] $k = m+1$ C [illegible] $\langle \mathbb{P}, \mathbb{P} \rangle = I_{m+1}$ [illegible] $C = I_{m+1}$

Likewise we have the next lemma.

LEMMA 3.2. [illegible] $\tilde{\mathbb{P}} \in \prod_{(k)}^{(n,m)}$ [illegible]

$$\langle \tilde{\mathbb{P}}, x^i y^j \rangle = 0, \quad 0 \le i \le n, \quad 0 \le j < m, \tag{3.8}$$

[illegible] $\tilde{\mathbb{P}} = C\tilde{\mathbb{P}}_{n,m}$ [illegible] C [illegible] $k \times (n+1)$. [illegible] $k = n+1$ C [illegible] $\langle \tilde{\mathbb{P}}, \tilde{\mathbb{P}} \rangle = I_{n+1}$ [illegible] $C = I_{n+1}$

The discussion above allows us to make contact with the matrix orthogonal polynomials introduced in section 2.

LEMMA 3.3. [illegible] $\mathbb{P}_{n,m}$ [illegible] (3.3) [illegible]

$$\mathbb{P}_{n,m} = L_n(x)[1, y, y^2, \dots, y^m]^\top \tag{3.9}$$

[illegible]

$$\begin{bmatrix} \mathbb{P}_{0,m}(x,y) \\ \mathbb{P}_{1,m}(x,y) \\ \\ \\ \mathbb{P}_{n,m}(x,y) \end{bmatrix} = \begin{bmatrix} L_0(x) \\ L_1(x) \\ \\ \\ L_n(x) \end{bmatrix} [1, y, \dots, y^m]^\top = L \begin{bmatrix} I_{m+1} \\ xI_{m+1} \\ \\ \\ x^n I_{m+1} \end{bmatrix} [1, y, \dots, y^m]^\top. \tag{3.10}$$

[illegible]. If we substitute the equation

$$\mathbb{P}_{n,m} = \hat{L}_n(x)[1 \ \cdots \ y^m]^\top = \sum_i \hat{L}_{n,i} x^i [1 \ \cdots \ y^m]^\top$$

into (3.7), where $\hat{L}_n(x)$ is some $(m+1) \times (m+1)$ matrix polynomial of degree n, we

find, for $j = 0, \dots, n-1$,

$$0 = \left\langle \mathbb{P}_{n,m},\, x^j \begin{bmatrix} 1 \\ \vdots \\ y^m \end{bmatrix} \right\rangle = \sum_{i=0}^{n} \hat{L}_{n,i} \left\langle x^i \begin{bmatrix} 1 \\ \vdots \\ y^m \end{bmatrix},\, x^j \begin{bmatrix} 1 \\ \vdots \\ y^m \end{bmatrix} \right\rangle$$

$$= \sum_{i=1}^{n} \hat{L}_{n,i} \begin{bmatrix} \mathcal{L}_{NM}(x^{i+j}) & \cdots & \mathcal{L}_{NM}(x^{i+j}y^m) \\ \vdots & & \vdots \\ \mathcal{L}_{NM}(x^{i+j}y^m) & \cdots & \mathcal{L}_{NM}(x^{i+j}y^{2m}) \end{bmatrix}$$

$$= \sum_{i=1}^{n} \hat{L}_{n,i} \mathcal{L}_m(x^i,\ x^j) = \mathcal{L}_m(\hat{L}_n(x),\ x^j).$$

Similarly,

$$\langle \mathbb{P}_{n,m},\, \mathbb{P}_{n,m} \rangle = I_{m+1} = \mathcal{L}_m \langle \hat{L}_n(x),\, \hat{L}_n(x) \rangle.$$

This coupled with (2.8) and the fact that (3.3) implies that $\hat{L}_{n,m}$ is lower triangular with positive diagonal entries, gives the result. □

As mentioned earlier, analogous formulas exist for orthogonal polynomials in the reverse lexicographical ordering with the roles of n and m interchanged.

THEOREM 3.4. [illegible] $\{\mathbb{P}_{n,m}\}$ [illegible] $\{\tilde{\mathbb{P}}_{n,m}\}$ $0 \le n \le N$ $0 \le m \le M$ [illegible]

$$x\mathbb{P}_{n,m} = A_{n+1,m}\mathbb{P}_{n+1,m} + B_{n,m}\mathbb{P}_{n,m} + A_{n,m}^{\top}\mathbb{P}_{n-1,m}, \tag{3.11}$$

$$\Gamma_{n,m}\mathbb{P}_{n,m} = \mathbb{P}_{n,m-1} - \mathcal{K}_{n,m}\tilde{\mathbb{P}}_{n-1,m}, \tag{3.12}$$

$$J_{n,m}^1\mathbb{P}_{n,m} = y\mathbb{P}_{n,m-1} + J_{n,m}^2\tilde{\mathbb{P}}_{n-1,m} + J_{n,m}^3\tilde{\mathbb{P}}_{n-1,m-1}, \tag{3.13}$$

$$\mathbb{P}_{n,m} = I_{n,m}\tilde{\mathbb{P}}_{n,m} + \Gamma_{n,m}^{\top}\mathbb{P}_{n,m-1}, \tag{3.14}$$

[illegible]

$$A_{n,m} = \langle x\mathbb{P}_{n-1,m}, \mathbb{P}_{n,m} \rangle \in M^{m+1,m+1}, \tag{3.15}$$

$$B_{n,m} = \langle x\mathbb{P}_{n,m}, \mathbb{P}_{n,m} \rangle \in M^{m+1,m+1}, \tag{3.16}$$

$$J_{n,m}^1 = \langle y\mathbb{P}_{n,m-1}, \mathbb{P}_{n,m} \rangle \in M^{m,m+1}, \tag{3.17}$$

$$J_{n,m}^2 = -\langle y\mathbb{P}_{n,m-1}, \tilde{\mathbb{P}}_{n-1,m} \rangle \in M^{m,n}, \tag{3.18}$$

$$J_{n,m}^3 = -\langle y\mathbb{P}_{n,m-1}, \tilde{\mathbb{P}}_{n-1,m-1} \rangle \in M^{m,n}, \tag{3.19}$$

$$\Gamma_{n,m} = \langle \mathbb{P}_{n,m-1}, \mathbb{P}_{n,m} \rangle \in M^{m,m+1}, \tag{3.20}$$

$$\mathcal{K}_{n,m} = \langle \mathbb{P}_{n,m-1}, \tilde{\mathbb{P}}_{n-1,m} \rangle \in M^{m,n}, \tag{3.21}$$

$$I_{n,m} = \langle \mathbb{P}_{n,m}, \tilde{\mathbb{P}}_{n,m} \rangle \in M^{m+1,n+1}. \tag{3.22}$$

[illegible] $\tilde{\mathbb{P}}_{n,m}(x,y)$ [illegible] ($\tilde{3}$.11), ($\tilde{3}$.12), [illegible] (3.11) follows from Lemma 3.3 and (2.14). To prove (3.12) note that because of the linear independence of the entries of $\mathbb{P}_{n,m}$, there is an $m \times (m+1)$ matrix $\Gamma_{n,m}$ such that $\Gamma_{n,m}\mathbb{P}_{n,m} - \mathbb{P}_{n,m-1} \in \prod_{(m)}^{(n-1,m)}(x,y)$. Furthermore,

$$\langle \Gamma_{n,m}\mathbb{P}_{n,m} - \mathbb{P}_{n,m-1}, x^i y^j \rangle = 0, \qquad 0 \le i \le n-1 \quad 0 \le j \le m-1.$$

Thus Lemma 3.2 implies that

$$\Gamma_{nm}\mathbb{P}_{n,m} - \mathbb{P}_{n,m-1} = \mathcal{K}_{n,m}\tilde{\mathbb{P}}_{n-1,m}.$$

The remaining recurrence formulas follow in a similar manner. □

REMARK 3.5. As indicated in the proof, formula (3.11) follows from the theory of matrix orthogonal polynomials and so allows us to move along a strip of size $m+1$. This formula does not mix the polynomials in the two orderings. However, to increase m by one for polynomials constructed in the lexicographical ordering, the remaining relations show that orthogonal polynomials in the reverse lexicographical ordering must be used.

REMARK 3.6. We saw in the previous section that $A_{n,m}$ is a lower triangular matrix with positive entries on the main diagonal, and $B_{n,m}$ is a symmetric matrix. From the orthogonality relations it follows immediately that $(J^1_{n,m})_{ij} = \langle y p^{i-1}_{n,m-1}, p^{j-1}_{n,m}\rangle = 0$ if $i+1 < j$, and $(J^1_{n,m})_{i,i+1} > 0$. Thus $J^1_{n,m}$ has the form

$$J^1_{n,m} = \begin{bmatrix} (J^1_{n,m})_{1,1} & (J^1_{n,m})_{1,2} & 0 & 0 & 0 \\ (J^1_{n,m})_{2,1} & (J^1_{n,m})_{2,2} & (J^1_{n,m})_{2,3} & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (J^1_{n,m})_{m,1} & (J^1_{n,m})_{m,2} & (J^1_{n,m})_{m,3} & \cdots & (J^1_{n,m})_{m,m+1} \end{bmatrix}.$$

Similarly, $(\Gamma_{n,m})_{ij} = \langle p^{i-1}_{n,m-1}, p^{j-1}_{n,m}\rangle = 0$ if $i < j$, and $(\Gamma_{n,m})_{i,i} > 0$, i.e., $\Gamma_{n,m}$ has the form

$$\Gamma_{n,m} = \begin{bmatrix} (\Gamma_{n,m})_{11} & 0 & \cdots & 0 & 0 \\ (\Gamma_{n,m})_{21} & (\Gamma_{n,m})_{22} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ (\Gamma_{n,m})_{m1} & (\Gamma_{n,m})_{m2} & \cdots & (\Gamma_{n,m})_{mm} & 0 \end{bmatrix}.$$

Finally notice that $p^m_{n,m} = \tilde{p}^n_{n,m}$ and therefore $(I_{n,m})_{m+1,n+1} = 1$, $(I_{n,m})_{m+1,j} = 0$ for $j \le n$ and $(I_{n,m})_{i,n+1} = 0$ for $i \le m$, i.e.,

$$I_{n,m} = \begin{bmatrix} * & * & \cdots & * & 0 \\ \vdots & \vdots & & \vdots & 0 \\ * & * & \cdots & * & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Using the orthogonality relations and Theorem 3.4 one can verify the following.

PROPOSITION 3.7. [illegible]

$$\tilde{\mathcal{K}}_{n,m} = \mathcal{K}^\top_{n,m}, \tag{3.23}$$

$$\mathcal{K}_{n,m}\mathcal{K}^\top_{n,m} + \Gamma_{n,m}\Gamma^\top_{n,m} = I_m, \tag{3.24}$$

$$J^3_{n,m} = -\mathcal{K}_{n,m}\tilde{A}^\top_{n-1,m}, \tag{3.25}$$

$$\tilde{I}_{n,m} = I^\top_{n,m}, \tag{3.26}$$

$$I_{n,m}I^\top_{n,m} + \Gamma^\top_{n,m}\Gamma_{n,m} = I_{m+1}. \tag{3.27}$$

PROOF. We prove only (3.25) since the others are obvious from their defining relations. Beginning with (3.19) we find

$$J^3_{n,m} = -\langle \mathbb{P}_{n,m-1}, y\tilde{\mathbb{P}}_{n-1,m-1}\rangle.$$

The result now follows by using the counterpart of (3.11) for $\tilde{\mathbb{P}}$ and the defining equation for $\mathcal{K}_{n,m}$. □

The function $\mathfrak{L}$ given by (2.18) can be used to show that $\mathbb{P}_{n,m}$ satisfies a certain minimization condition. Define $\hat{\mathfrak{L}} : \prod_{(m+1)}^{(n,m)} \to \operatorname{sym} \mathbb{R}_{m+1}$ by

$$\hat{\mathfrak{L}}(P) = \langle P, P \rangle - 2\operatorname{sym} K_n, \tag{3.28}$$

where $P \in \prod_{(m+1)}^{(n,m)}$ is written as

$$P(x,y) = K_n x^n \begin{bmatrix} 1 \\ \vdots \\ y^m \end{bmatrix} + R(x,y) \tag{3.29}$$

with $R \in \prod_{(m+1)}^{(n-1,m)}$. We find,

LEMMA 3.8. [illegible] $\hat{\mathbb{P}}_{n,m}(x,y) = L_{n,n}^{-1}\mathbb{P}_{n,m}(x,y)$ [illegible] $\hat{\mathfrak{L}}$ [illegible] (2.21)

[illegible]. Write P as

$$P(x,y) = \sum_{i=0}^{n} K_i x^i \begin{bmatrix} 1 \\ \vdots \\ y^m \end{bmatrix}.$$

Since

$$\left\langle x^i \begin{bmatrix} 1 \\ \vdots \\ y^m \end{bmatrix}, x^j \begin{bmatrix} 1 \\ \vdots \\ y^m \end{bmatrix} \right\rangle = \begin{bmatrix} \mathcal{L}_{N,M}(x^i,\, x^j) & \cdots & \mathcal{L}_{N,M}(x^i,\, x^j y^m) \\ \vdots & & \vdots \\ \mathcal{L}_{N,M}(x^i y^m,\, x^j) & \cdots & \mathcal{L}_{N,M}(x^i y^m,\, x^j y^m) \end{bmatrix} = \mathcal{L}_m(x^i,\, x^j),$$

we see that $\hat{\mathfrak{L}}(P)$ can be written as $\hat{\mathfrak{L}}(P) = \mathfrak{L}(K)$, where $K(x) = K_n x^n + \cdots \in \prod_{m+1}^{n}$. The result now follows from (2.22) and Lemma 3.3. □

4. Relations. As is evident from the previous section, there are relations between the various coefficients in (3.11)–(3.14) and their ($\tilde{3}$.11)–($\tilde{3}$.14) analogues. In this section we exhibit these relations.

LEMMA 4.1 (relations for $\mathcal{K}_{n,m}$).

$$\Gamma_{n,m-1}\mathcal{K}_{n,m}\tilde{A}_{n-1,m}^{\top} = -J_{n,m-1}^2 - \mathcal{K}_{n,m-1}\tilde{B}_{n-1,m-1}, \tag{4.1}$$

$$A_{n,m-1}\mathcal{K}_{n,m}\tilde{\Gamma}_{n-1,m}^{\top} = -\tilde{J}_{n-1,m}^{2\top} - B_{n-1,m-1}\mathcal{K}_{n-1,m}. \tag{4.2}$$

[illegible]. We have

$$\begin{aligned} \mathcal{K}_{n,m}\tilde{A}_{n-1,m}^{\top} &= \langle \mathbb{P}_{n,m-1}, \tilde{A}_{n-1,m}\tilde{\mathbb{P}}_{n-1,m} \rangle \\ &= \langle \mathbb{P}_{n,m-1}, y\tilde{\mathbb{P}}_{n-1,m-1} - \tilde{B}_{n-1,m-1}\tilde{\mathbb{P}}_{n-1,m-1} - \tilde{A}_{n-1,m-1}^{\top}\tilde{\mathbb{P}}_{n-1,m-2} \rangle \\ &= \langle \mathbb{P}_{n,m-1}, y\tilde{\mathbb{P}}_{n-1,m-1} \rangle. \end{aligned}$$

Thus,

$$\begin{aligned} \Gamma_{n,m-1}\mathcal{K}_{n,m}\tilde{A}_{n-1,m}^{\top} &= \langle \Gamma_{n,m-1}\mathbb{P}_{n,m-1}, y\tilde{\mathbb{P}}_{n-1,m-1} \rangle \\ &= \langle \mathbb{P}_{n,m-2} - \mathcal{K}_{n,m-1}\tilde{\mathbb{P}}_{n-1,m-1}, y\tilde{\mathbb{P}}_{n-1,m-1} \rangle \\ &= -J_{n,m-1}^2 - \mathcal{K}_{n,m-1}\tilde{B}_{n-1,m-1}, \end{aligned}$$

which completes the proof of (4.1). Writing (4.1) for $\tilde{\mathcal{K}}_{n,m}$ and using (3.23) we obtain (4.2). ∎

LEMMA 4.2 (relations for $J^2_{n,m}$).

$$\Gamma_{n,m-1}J^2_{n,m} = -J^1_{n,m-1}\mathcal{K}_{n,m} + \mathcal{K}_{n,m-1}\tilde{A}_{n-1,m}, \tag{4.3}$$

$$\begin{aligned} -A_{n,m-1}J^2_{n,m}\tilde{\Gamma}^\top_{n-1,m} &= J^1_{n-1,m}A^\top_{n-1,m}I_{n-2,m} - J^2_{n-1,m}\tilde{\Gamma}_{n-1,m}\tilde{J}^{1\,\top}_{n-1,m} \\ &\quad + J^2_{n-1,m}\mathcal{K}^\top_{n-1,m}\tilde{J}^{2\,\top}_{n-1,m} + J^3_{n-1,m}I^\top_{n-2,m-1}\tilde{J}^{3\,\top}_{n-1,m} \\ &\quad + B_{n-1,m-1}J^2_{n-1,m} - A^\top_{n-1,m-1}I_{n-2,m-1}\tilde{A}_{n-2,m} \\ &\quad + A_{n,m-1}J^3_{n,m}I^\top_{n-1,m-1}\mathcal{K}_{n-1,m}. \end{aligned} \tag{4.4}$$

PROOF. The first equation can be derived by multiplying (3.18) on the left by $\Gamma_{n,m-1}$ then using ($\tilde{3}$.15) to obtain

$$\Gamma_{n,m-1}J^2_{n,m} = -\langle y\mathbb{P}_{n,m-2}, \tilde{\mathbb{P}}_{n-1,m}\rangle + \mathcal{K}_{n,m-1}\tilde{A}_{n-1,m}.$$

Eliminating $y\mathbb{P}_{n,m-2}$ using (3.13), then using the orthogonality of the polynomials and (3.19) yields (4.3).

To derive (4.4) notice that

$$\begin{aligned} -A_{n,m-1}J^2_{n,m}\tilde{\Gamma}^\top_{n-1,m} &= A_{n,m-1}\langle y\mathbb{P}_{n,m-1}, \tilde{\Gamma}_{n-1,m}\tilde{\mathbb{P}}_{n-1,m}\rangle \\ &= A_{n,m-1}\langle y\mathbb{P}_{n,m-1}, \tilde{\mathbb{P}}_{n-2,m}\rangle \\ &\quad - A_{n,m-1}\langle y\mathbb{P}_{n,m-1}, \mathbb{P}_{n-1,m-1}\rangle\mathcal{K}_{n-1,m}. \end{aligned} \tag{4.5}$$

Using (3.11) in the first term on the right-hand side of (4.5) gives

$$\begin{aligned} A_{n,m-1}\langle y\mathbb{P}_{n,m-1}, \tilde{\mathbb{P}}_{n-2,m}\rangle &= \langle x\mathbb{P}_{n-1,m-1}, y\tilde{\mathbb{P}}_{n-2,m}\rangle - B_{n-1,m-1}J^2_{n-1,m} \\ &\quad - A^\top_{n-1,m-1}\langle \mathbb{P}_{n-2,m-1}, y\tilde{\mathbb{P}}_{n-2,m}\rangle. \end{aligned} \tag{4.6}$$

Interchanging the positions of x and y in the first term on the right-hand side of (4.6), then using (3.13) and its ($\tilde{3}$.13) analogue yield

$$\begin{aligned} \langle x\mathbb{P}_{n-1,m-1}, y\tilde{\mathbb{P}}_{n-2,m}\rangle &= J^1_{n-1,m}(I_{n-1,m}\tilde{J}^{1\,\top}_{n-1,m} - \Gamma^\top_{n-1,m}\tilde{J}^{2\,\top}_{n-1,m}) \\ &\quad - J^2_{n-1,m}\tilde{\Gamma}_{n-1,m}\tilde{J}^{1\,\top}_{n-1,m} + J^2_{n-1,m}\mathcal{K}^\top_{n-1,m}\tilde{J}^{2\,\top}_{n-1,m} + J^3_{n-1,m}I^\top_{n-2,m-1}\tilde{J}^{3\,\top}_{n-1,m}, \end{aligned}$$

where (3.21), (3.22), and their ($\tilde{3}$.21), ($\tilde{3}$.22) analogues have been used.

Substituting (3.26) as well as the transpose of ($\tilde{4}$.18) into the above equation yields

$$\begin{aligned} \langle x\mathbb{P}_{n-1,m-1}, y\tilde{\mathbb{P}}_{n-2,m}\rangle &= J^1_{n-1,m}A^\top_{n-1,m}I_{n-2,m} - J^2_{n-1,m}\tilde{\Gamma}_{n-1,m}\tilde{J}^{1\,\top}_{n-1,m} \\ &\quad + J^2_{n-1,m}\mathcal{K}^\top_{n-1,m}\tilde{J}^{2\,\top}_{n-1,m} + J^3_{n-1,m}I^\top_{n-2,m-1}\tilde{J}^{3\,\top}_{n-1,m}. \end{aligned} \tag{4.7}$$

The last term in (4.6) can be computed using ($\tilde{3}$.11) and (3.22), which gives

$$\langle \mathbb{P}_{n-2,m-1}, y\tilde{\mathbb{P}}_{n-2,m}\rangle = I_{n-2,m-1}\tilde{A}_{n-2,m}. \tag{4.8}$$

Substituting (4.7) and (4.8) into (4.6) we see that the first term of (4.5) is

$$\begin{aligned} A_{n,m-1}\langle y\mathbb{P}_{n,m-1}, \tilde{\mathbb{P}}_{n-2,m}\rangle &= J^1_{n-1,m}A^\top_{n-1,m}I_{n-2,m} - J^2_{n-1,m}\tilde{\Gamma}_{n-1,m}\tilde{J}^{1\,\top}_{n-1,m} \\ &\quad + J^2_{n-1,m}\mathcal{K}^\top_{n-1,m}\tilde{J}^{2\,\top}_{n-1,m} + J^3_{n-1,m}I^\top_{n-2,m-1}\tilde{J}^{3\,\top}_{n-1,m} \\ &\quad + B_{n-1,m-1}J^2_{n-1,m} - A^\top_{n-1,m-1}I_{n-2,m-1}\tilde{A}_{n-2,m}. \end{aligned} \tag{4.9}$$

Substituting (3.13) in the second term on the right-hand side of (4.5) and using the equations

$$\langle \mathbb{P}_{n,m-1}, \mathbb{P}_{n-1,m} \rangle = \mathcal{K}_{n,m} I_{n-1,m}^{\top} \tag{4.10}$$

and

$$\langle \mathbb{P}_{n,m-1}, \tilde{\mathbb{P}}_{n-2,m} \rangle = \mathcal{K}_{n,m} \tilde{\Gamma}_{n-1,m}^{\top}, \tag{4.11}$$

which follow easily from (3.12), yields

$$\begin{aligned} & A_{n,m-1} \langle y\mathbb{P}_{n,m-1}, \mathbb{P}_{n-1,m-1} \rangle \mathcal{K}_{n-1,m} \\ & \quad = A_{n,m-1} \mathcal{K}_{n,m} I_{n-1,m}^{\top} J_{n-1,m}^{1\,\top} \mathcal{K}_{n-1,m} - A_{n,m-1} \mathcal{K}_{n,m} \tilde{\Gamma}_{n-1,m}^{\top} J_{n-1,m}^{2\,\top} \mathcal{K}_{n-1,m} \\ & \quad = A_{n,m-1} \mathcal{K}_{n,m} (I_{n-1,m}^{\top} J_{n-1,m}^{1\,\top} - \tilde{\Gamma}_{n-1,m}^{\top} J_{n-1,m}^{2\,\top}) \mathcal{K}_{n-1,m} \\ & \quad = -A_{n,m-1} J_{n,m}^{3} I_{n-1,m-1}^{\top} \mathcal{K}_{n-1,m}. \end{aligned} \tag{4.12}$$

In the last equality we used (3.25) and (4.18). Finally, combining (4.9) and (4.12) we obtain (4.4). □

LEMMA 4.3 (relations for $J_{n,m}^{1}$).

$$\Gamma_{n,m-1} J_{n,m}^{1} = J_{n,m-1}^{1} \Gamma_{n,m}, \tag{4.13}$$

$$J_{n,m}^{1} \Gamma_{n,m}^{\top} \Gamma_{n,m-1}^{\top} = J_{n,m-1}^{1\,\top} + J_{n,m}^{3} \mathcal{K}_{n,m-1}^{\top} + J_{n,m}^{2} \mathcal{K}_{n,m}^{\top} \Gamma_{n,m-1}^{\top}. \tag{4.14}$$

PROOF. (4.13) can be derived by multiplying (3.17) by $\Gamma_{n,m-1}$ then using (3.13). For the second equality we multiply (3.17) by $\Gamma_{n,m}^{\top}$ then use (3.12) to obtain

$$J_{n,m}^{1} \Gamma_{n,m}^{\top} = \langle y\mathbb{P}_{n,m-1}, \Gamma_{n,m} \mathbb{P}_{n,m} \rangle = \langle y\mathbb{P}_{n,m-1}, \mathbb{P}_{n,m-1} \rangle + J_{n,m}^{2} \mathcal{K}_{n,m}^{\top}.$$

Multiplying on the right of the above formula by $\Gamma_{n,m-1}^{\top}$, then using (3.12) followed by (3.13) twice, leads to the result. □

LEMMA 4.4 (relations for $A_{n,m}$).

$$\Gamma_{n-1,m} A_{n,m} = A_{n,m-1} \Gamma_{n,m}, \tag{4.15}$$

$$J_{n-1,m}^{1} A_{n,m} = A_{n,m-1} J_{n,m}^{1}. \tag{4.16}$$

PROOF. First we compute

$$\langle \mathbb{P}_{n-1,m-1}, x\mathbb{P}_{n,m} \rangle = \langle \mathbb{P}_{n-1,m-1}, \mathbb{P}_{n-1,m} \rangle A_{n,m} = \Gamma_{n-1,m} A_{n,m}.$$

On the other hand,

$$\langle \mathbb{P}_{n-1,m-1}, x\mathbb{P}_{n,m} \rangle = A_{n,m-1} \Gamma_{n,m}.$$

This gives (4.15). If we use (3.17) with n changed to $n-1$, then multiply on the right by $A_{n,m}$ and use (3.11), we find

$$J_{n-1,m}^{1} A_{n,m} = \langle x\mathbb{P}_{n-1,m-1}, y\mathbb{P}_{n,m} \rangle.$$

Now eliminating $x\mathbb{P}_{n-1,m-1}$ using (3.11), then applying (3.17), yields (4.16). □

LEMMA 4.5 (relations for $I_{n,m}$).

$$\Gamma_{n,m} I_{n,m} = -\mathcal{K}_{n,m} \tilde{\Gamma}_{n,m}, \tag{4.17}$$

$$J^1_{n,m} I_{n,m} = I_{n,m-1} \tilde{A}_{n,m} + J^2_{n,m} \tilde{\Gamma}_{n,m}. \tag{4.18}$$

PROOF. The first relation follows in a straightforward manner by multiplying (3.22) on the left by $\Gamma_{n,m}$ and then using (3.12).

For (4.18) we first compute

$$I_{n,m-1} \tilde{A}_{n,m} = \langle \mathbb{P}_{n,m-1}, y\tilde{\mathbb{P}}_{n,m} \rangle,$$

which follows by using ($\tilde{3}$.11) to eliminate $y\tilde{\mathbb{P}}_{n,m}$. Next, in the above equation use (3.17) to obtain

$$\langle y\mathbb{P}_{n,m-1}, \tilde{\mathbb{P}}_{n,m} \rangle = J^1_{n,m} I_{n,m} - J^2_{n,m} \tilde{\Gamma}_{n,m},$$

which gives (4.18). □

LEMMA 4.6 (relations for $B_{n-1,m}$).

$$\begin{aligned} \Gamma_{n-1,m} B_{n-1,m} = & -\mathcal{K}_{n-1,m} I^{\top}_{n-2,m} A_{n-1,m} + B_{n-1,m-1} \Gamma_{n-1,m} \\ & + A_{n,m-1} \mathcal{K}_{n,m} I^{\top}_{n-1,m}, \end{aligned} \tag{4.19}$$

$$\begin{aligned} J^1_{n-1,m} B_{n-1,m} = & B_{n-1,m-1} J^1_{n-1,m} + J^2_{n-1,m} I^{\top}_{n-2,m} A_{n-1,m} \\ & + J^3_{n-1,m} I^{\top}_{n-2,m-1} A_{n-1,m-1} \Gamma_{n-1,m} - A_{n,m-1} J^2_{n,m} I^{\top}_{n-1,m} \\ & - A_{n,m-1} J^3_{n,m} I^{\top}_{n-1,m-1} \Gamma_{n-1,m}. \end{aligned} \tag{4.20}$$

PROOF. We begin by multiplying (3.20) on the left by $B_{n-1,m}$ and then using (3.11) to obtain

$$\Gamma_{n-1,m} B_{n-1,m} = \langle \mathbb{P}_{n-1,m-1}, x\mathbb{P}_{n-1,m} \rangle - \langle \mathbb{P}_{n-1,m-1}, \mathbb{P}_{n-2,m} \rangle A_{n-1,m}.$$

We see from (4.10) that the second term on the right-hand side of the above formula gives the first term on the right-hand side in (4.19). We can compute the first term on the right-hand side of the above formula by eliminating $x\mathbb{P}_{n-1,m}$ using (3.11) to find

$$\begin{aligned} \langle \mathbb{P}_{n-1,m-1}, x\mathbb{P}_{n-1,m} \rangle &= A_{n,m-1} \langle \mathbb{P}_{n,m-1}, \mathbb{P}_{n-1,m} \rangle + B_{n-1,m-1} \langle \mathbb{P}_{n-1,m-1}, \mathbb{P}_{n-1,m} \rangle \\ &= A_{n,m-1} \mathcal{K}_{n,m} I^{\top}_{n-1,m} + B_{n-1,m-1} \Gamma_{n-1,m}, \end{aligned}$$

where in the last equality we used again (4.10). This completes the proof of (4.19). Relation (4.20) can be derived as follows. First we multiply (3.16) with n reduced by one on the left by $J^1_{n-1,m}$ to obtain

$$\begin{aligned} J^1_{n-1,m} B_{n-1,m} = & \langle y\mathbb{P}_{n-1,m-1}, x\mathbb{P}_{n-1,m} \rangle + J^2_{n-1,m} \langle \tilde{\mathbb{P}}_{n-2,m}, x\mathbb{P}_{n-1,m} \rangle \\ & + J^3_{n-1,m} \langle \tilde{\mathbb{P}}_{n-2,m-1}, xP_{n-1,m} \rangle. \end{aligned} \tag{4.21}$$

Next we compute the three terms on the right-hand side of the above formula. For the first term we eliminate $x\mathbb{P}_{n-1,m-1}$ using (3.11) to find

$$\langle y\mathbb{P}_{n-1,m-1}, x\mathbb{P}_{n-1,m} \rangle = A_{n,m-1} \langle \mathbb{P}_{n,m-1}, y\mathbb{P}_{n-1,m} \rangle + B_{n-1,m-1} J^1_{n-1,m}.$$

The term $\langle \mathbb{P}_{n,m-1}, y\mathbb{P}_{n-1,m}\rangle$ can be computed using (3.13) so that

$$\begin{aligned}\langle y\mathbb{P}_{n-1,m-1}, x\mathbb{P}_{n-1,m}\rangle = &- A_{n,m-1}J^2_{n,m}I^\top_{n-1,m} - A_{n,m-1}J^3_{n,m}\tilde{I}_{n-1,m-1}\Gamma_{n-1,m}\\ &+ B_{n-1,m-1}J^1_{n-1,m}.\end{aligned} \tag{4.22}$$

Next we compute the second term in (4.21) using (3.13) to obtain

$$\langle \tilde{\mathbb{P}}_{n-2,m}, x\mathbb{P}_{n-1,m}\rangle = \tilde{J}^1_{n-1,m}I^\top_{n-1,m} - \tilde{J}^2_{n-1,m}\Gamma_{n-1,m}.$$

Using (4.18) for $\tilde{I}_{n-1,m}$ and (3.26) we obtain the following formula for the second term

$$J^2_{n-1,m}\langle \tilde{\mathbb{P}}_{n-2,m}, x\mathbb{P}_{n-1,m}\rangle = J^2_{n-1,m}I^\top_{n-2,m}A_{n-1,m}. \tag{4.23}$$

Finally, the third term in (4.21) can be computed with the help of ($\tilde{3}$.13),

$$\langle \tilde{\mathbb{P}}_{n-2,m-1}, x\mathbb{P}_{n-1,m}\rangle = \tilde{J}^1_{n-1,m-1}\langle \tilde{\mathbb{P}}_{n-1,m-1}, \mathbb{P}_{n-1,m}\rangle - \tilde{J}^2_{n-1,m-1}\langle \mathbb{P}_{n-1,m-2}, \mathbb{P}_{n-1,m}\rangle.$$

Notice that

$$\langle \tilde{\mathbb{P}}_{n-1,m-1}, \mathbb{P}_{n-1,m}\rangle = \tilde{I}_{n-1,m-1}\Gamma_{n-1,m},$$

and using (3.14) for $\mathbb{P}_{n-1,m}$, we get

$$\langle \mathbb{P}_{n-1,m-2}, \mathbb{P}_{n-1,m}\rangle = \Gamma_{n-1,m-1}\Gamma_{n-1,m}.$$

Thus we have

$$J^3_{n-1,m}\langle \tilde{\mathbb{P}}_{n-2,m-1}, x\mathbb{P}_{n-1,m}\rangle = J^3_{n-1,m}\tilde{I}_{n-2,m-1}A_{n-1,m-1}\Gamma_{n-1,m},$$

where in the last equality we used again ($\tilde{4}$.18). Combining the above equation, (4.22), and (4.23), we obtain (4.20). □

LEMMA 4.7 (relations for $\tilde{J}^1_{n,m}$).

$$\begin{aligned}\tilde{J}^1_{n,m} =& I^\top_{n-1,m}A_{n,m}I_{n,m} + I^\top_{n-1,m}B_{n-1,m}I_{n-1,m}\tilde{\Gamma}_{n,m}\\ &+ I^\top_{n-1,m}A^\top_{n-1,m}I_{n-2,m}\tilde{\Gamma}_{n-1,m}\tilde{\Gamma}_{n,m} + \tilde{\Gamma}^\top_{n-1,m}\tilde{J}^1_{n-1,m}\tilde{\Gamma}_{n,m}.\end{aligned} \tag{4.24}$$

Proof. We begin by eliminating $\tilde{\mathbb{P}}_{n-1,m}$ and $\tilde{\mathbb{P}}_{n,m}$ in ($\tilde{3}$.17) using ($\tilde{3}$.14) and ($\tilde{3}$.13),

$$\begin{aligned}\tilde{J}^1_{n,m} =& I^\top_{n-1,m}\langle x\mathbb{P}_{n-1,m}, I^\top_{n,m}\mathbb{P}_{n,m} + \tilde{\Gamma}^\top_{n,m}\tilde{\mathbb{P}}_{n-1,m}\rangle\\ &+ \tilde{\Gamma}^\top_{n-1,m}\langle \tilde{J}^1_{n-1,m}\tilde{\mathbb{P}}_{n-1,m} - \tilde{J}^2_{n-1,m}\mathbb{P}_{n-1,m-1} - \tilde{J}^3_{n-1,m}\mathbb{P}_{n-2,m-1}, \tilde{\mathbb{P}}_{n,m}\rangle,\end{aligned}$$

which simplifies to

$$\tilde{J}^1_{n,m} = I^\top_{n-1,m}A_{n,m}I_{n,m} + I^\top_{n-1,m}\langle x\mathbb{P}_{n-1,m}, \tilde{\mathbb{P}}_{n-1,m}\rangle\tilde{\Gamma}_{n,m} + \tilde{\Gamma}^\top_{n-1,m}\tilde{J}^1_{n-1,m}\tilde{\Gamma}_{n,m}.$$

In the above equation, (3.15) and ($\tilde{3}$.20) have been used. Next use (3.11) to eliminate $x\mathbb{P}_{n-1,m}$. All terms can be evaluated using (3.15)–(3.22) except the term $\langle \mathbb{P}_{n-2,m}, \tilde{\mathbb{P}}_{n-1,m}\rangle$, which can be evaluated by applying ($\tilde{3}$.14) to $\tilde{\mathbb{P}}_{n-1,m}$. □

Similar arguments show the next lemma.

LEMMA 4.8 (relations for $\tilde{J}^2_{n,m}$).

$$\begin{aligned}\tilde{J}^2_{n,m} = &-(I^\top_{n-1,m}A_{n,m}\Gamma^\top_{n,m} + I^\top_{n-1,m}B_{n-1,m}I_{n-1,m}\mathcal{K}^\top_{n,m}\\ &+ I^\top_{n-1,m}A^\top_{n-1,m}I_{n-2,m}\tilde{\Gamma}_{n-1,m}\mathcal{K}^\top_{n,m} + \tilde{\Gamma}^\top_{n-1,m}\tilde{J}^1_{n-1,m}\mathcal{K}^\top_{n,m}).\end{aligned} \tag{4.25}$$

LEMMA 4.9 (relations for $\tilde{A}_{n,m}$).

$$\tilde{A}_{n,m} = I^\top_{n,m-1}J^1_{n,m}I_{n,m} - I^\top_{n,m-1}J^2_{n,m}\tilde{\Gamma}_{n,m} + \tilde{\Gamma}^\top_{n,m-1}\tilde{A}_{n-1,m}\tilde{\Gamma}_{n,m}. \tag{4.26}$$

PROOF. From (3.15) we obtain

$$\tilde{A}_{n,m} = I^\top_{n,m-1}\langle y\mathbb{P}_{n,m-1}, \tilde{\mathbb{P}}_{n,m}\rangle + \tilde{\Gamma}^\top_{n,m-1}\langle y\tilde{\mathbb{P}}_{n-1,m-1}, \tilde{\mathbb{P}}_{n,m}\rangle.$$

Eliminate $y\mathbb{P}_{n,m-1}$ using (3.13) and $y\tilde{\mathbb{P}}_{n-1,m-1}$ using (3.11) leads to the result. ∎

Again in an analogous fashion, we have the next lemma.

LEMMA 4.10 (relations for $\tilde{B}_{n,m}$).

$$\begin{aligned}\tilde{B}_{n,m-1} = &\, I^\top_{n,m-1}\Gamma_{n,m}J^{1\top}_{n,m}I_{n,m-1} - I^\top_{n,m-1}\mathcal{K}_{n,m}J^{2\top}_{n,m}I_{n,m-1}\\ &- \tilde{\Gamma}^\top_{n,m-1}J^{3\top}_{n,m}I_{n,m-1} - I^\top_{n,m-1}J^3_{n,m}\tilde{\Gamma}_{n,m-1} + \tilde{\Gamma}^\top_{n,m-1}\tilde{B}_{n-1,m-1}\tilde{\Gamma}_{n,m-1}.\end{aligned} \tag{4.27}$$

5. Christoffel–Darboux-like formulas. It is well known that the Christoffel–Darboux formula plays an important role in the theory of orthogonal polynomials of one variable. Using the connection between matrix orthogonal polynomials and two variable orthogonal polynomials developed in section 3 we will present two variable analogues of this celebrated formula.

THEOREM 5.1 (Christoffel–Darboux formula).

$$\begin{aligned}&\frac{\mathbb{P}^\top_{n,m}(x_1,y_1)A_{n+1,m}\mathbb{P}_{n+1,m}(x,y) - \mathbb{P}^\top_{n+1,m}(x_1,y_1)A^\top_{n+1,m}\mathbb{P}_{n,m}(x,y)}{x-x_1}\\ &\qquad = \sum_{k=0}^{n}\mathbb{P}^\top_{k,m}(x_1,y_1)\mathbb{P}_{k,m}(x,y)\\ &\qquad = \sum_{j=0}^{m}\tilde{\mathbb{P}}^\top_{n,j}(x_1,y_1)\tilde{\mathbb{P}}_{n,j}(x,y).\end{aligned}$$

PROOF. The first equality follows from (3.11) and standard manipulations. The second equality follows since both sums are reproducing kernels for the same space. ∎

An analogous result holds for the reverse lexicographical ordering. The above Theorem also implies the next lemma.

LEMMA 5.2.

$$\begin{aligned}&\mathbb{P}^\top_{n,m}(x_1,y_1)A_{n+1,m}\mathbb{P}_{n+1,m}(x,y) - \mathbb{P}^\top_{n+1,m}(x_1,y_1)A^\top_{n+1,m}\mathbb{P}_{n,m}(x,y)\\ &\quad = (x-x_1)\tilde{\mathbb{P}}^\top_{n,m}(x_1,y_1)\tilde{\mathbb{P}}_{n,m}(x,y)\\ &+ \mathbb{P}^\top_{n,m-1}(x_1,y_1)A_{n+1,m-1}\mathbb{P}_{n+1,m-1}(x,y) - \mathbb{P}^\top_{n+1,m-1}(x_1,y_1)A^\top_{n+1,m-1}\mathbb{P}_{n,m-1}(x,y),\end{aligned} \tag{5.1}$$

$$\mathbb{P}^{\top}_{n+1,m+1}(x_1,y_1)\mathbb{P}_{n+1,m+1}(x,y) - \mathbb{P}^{\top}_{n+1,m}(x_1,y_1)\mathbb{P}_{n+1,m}(x,y)$$
$$= \tilde{\mathbb{P}}^{\top}_{n+1,m+1}(x_1,y_1)\tilde{\mathbb{P}}_{n+1,m+1}(x,y) - \tilde{\mathbb{P}}^{\top}_{n,m+1}(x_1,y_1)\tilde{\mathbb{P}}_{n,m+1}(x,y). \tag{5.2}$$

. To prove the first formula let

$$Z_{n,m}(x,y) = [1, y, \dots, y^m][I_{m+1}, xI_{m+1}, \dots, x^n I_{m+1}],$$

and let $\tilde{Z}_{n,m}(x,y)$ be given by a similar formula with the roles of x and y, and n and m interchanged. Then from the Christoffel–Darboux formula, Lemma 2.5, and (3.10) we find

$$\frac{\mathbb{P}^{\top}_{n,m}(x_1,y_1)A_{n+1,m}\mathbb{P}_{n+1,m}(x,y) - \mathbb{P}^{\top}_{n+1,m}(x_1,y_1)A^{\top}_{n+1,m}\mathbb{P}_{n,m}(x,y)}{x - x_1}$$
$$= Z_{n,m}(x_1,y_1)H^{-1}_{n,m}Z_{n,m}(x,y)^{\top} = \tilde{Z}_{n,m}(x_1,y_1)\tilde{H}^{-1}_{n,m}\tilde{Z}_{n,m}(x,y)^{\top}$$
$$= \tilde{\mathbb{P}}^{\top}_{n,m}(x_1,y_1)\tilde{\mathbb{P}}_{n,m}(x,y) + \tilde{Z}_{n,m-1}(x_1,y_1)\tilde{H}^{-1}_{n,m-1}\tilde{Z}_{n,m-1}(x,y)^{\top}.$$

Switching back to the lexicographical ordering in the second term in the last equation implies the result. (5.2) can be obtained by using the equality of the sums in Theorem 5.1 to find

$$\mathbb{P}^{\top}_{n+1,m+1}(x_1,y_1)\mathbb{P}_{n+1,m+1}(x,y) - \sum_{j=0}^{m}\tilde{\mathbb{P}}^{\top}_{n+1,j}(x_1,y_1)\tilde{\mathbb{P}}_{n+1,j}(x,y)$$
$$= \tilde{\mathbb{P}}^{\top}_{n+1,m+1}(x_1,y_1)\tilde{\mathbb{P}}_{n+1,m+1}(x,y) - \sum_{j=0}^{n}\mathbb{P}^{\top}_{j,m+1}(x_1,y_1)\mathbb{P}_{j,m+1}(x,y).$$

Switching to the lexicographical ordering in the sum on the left-hand side of the above equation and reverse lexicographical ordering in the sum on the right-hand side then extracting out the highest terms and using Theorem 5.1 gives the result. □

5.3. The above equations can be derived from the recurrence formulas in the previous sections. (5.2) follows easily from (3.12) and Proposition 3.7. However, the derivation of (5.1) is rather tedious.

6. Algorithm. With the use of the relations derived in the previous section we develop an algorithm that allows us to compute the coefficients in the recurrence formulas at higher levels in terms of those at lower levels plus some indeterminates that are equivalent to the moments (see Theorem 7.1). More precisely, at each level (n,m) we construct the matrices $\mathcal{K}_{n,m}$, $\Gamma_{n,m}$, $J^2_{n,m}$, $J^1_{n,m}$, $A_{n,m}$, $I_{n,m}$, $B_{n-1,m}$, $\tilde{A}_{n,m}$, $\tilde{J}^1_{n,m}$, $\tilde{J}^2_{n,m}$, $\tilde{B}_{n,m-1}$ and the polynomials $\mathbb{P}_{n,m}(x,y)$ and $\tilde{\mathbb{P}}_{n,m}(x,y)$ recursively, using the matrices at levels $(n-1,m)$ and $(n,m-1)$. In order to construct the above matrices we will have need of the $m \times (m+1)$ matrix U_m given by

$$U_m = [I_m|0] = [\delta_{i,j}] = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

and the $m \times m$ elementary matrix $E_{m,m}$ having just one nonzero entry at (m,m). The matrix norm used in that and the remaining sections is the l_2 norm.

At level $(0,0)$ we have just one free parameter (corresponding to $h_{0,0} = \langle 1 \rangle$), at level $(n,0)$ (resp., $(0,m)$) we have two new parameters corresponding to $h_{2n-1,0} = \langle x^{2n-1} \rangle$ and $h_{2n,0} = \langle x^{2n} \rangle$ (resp., $h_{0,2m-1} = \langle y^{2m-1} \rangle$, and $h_{0,2m} = \langle y^{2m} \rangle$), and, if $n > 0$ and $m > 0$, we have four new parameters corresponding to the moments $h_{2n-1,2m-1} = \langle x^{2n-1} y^{2m-1} \rangle$, $h_{2n-1,2m} = \langle x^{2n-1} y^{2m} \rangle$, $h_{2n,2m-1} = \langle x^{2n} y^{2m-1} \rangle$, $h_{2n,2m} = \langle x^{2n} y^{2m} \rangle$.

Level(0,0). When $n = m = 0$ we simply put

$$\mathbb{P}_{0,0}(x,y) = \tilde{\mathbb{P}}_{0,0}(x,y) = s_{0,0}, \tag{6.1}$$

where $s_{0,0}$ is the new parameter corresponding to the moment $\langle 1 \rangle$.

Level(n,0). When $m = 0$, $\mathbb{P}_{n,0} = (p^0_{n,0})$ is just a scalar valued function in x and clearly

$$\tilde{\mathbb{P}}_{n,0} = \begin{bmatrix} p^0_{0,0} \\ p^0_{1,0} \\ \vdots \\ p^0_{n,0} \end{bmatrix}, \text{ i.e., } \tilde{p}^k_{n,0} = p^0_{k,0}. \tag{6.2}$$

Thus $I_{n,0} = (0,0,\ldots,0,1)$ $(1 \times (n+1)$ matrix).

To construct all other matrices at level $(n,0)$ from $(n-1,0)$ we have two new parameters, $s_{2n-1,0}$ and $s_{2n,0}$, corresponding to the moments $\langle x^{2n-1} \rangle$ and $\langle x^{2n} \rangle$.

We take $A_{n,0} = s_{2n,0} > 0$ and $B_{n-1,0} = s_{2n-1,0}$. Then,

$$p^0_{n,0} = A^{-1}_{n,0}(x p^0_{n-1,0} - B_{n-1,0} p^0_{n-1,0} - A_{n-1,0} p^0_{n-2,0}), \tag{6.3}$$

and $\tilde{J}^1_{n,0}$ is the tridiagonal matrix

$$\tilde{J}^1_{n,0} = \begin{bmatrix} B_{0,0} & A_{1,0} & & & & & \\ A_{1,0} & B_{1,0} & A_{2,0} & & & & \\ & A_{2,0} & B_{2,0} & A_{3,0} & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & A_{n-1,0} & B_{n-1,0} & A_{n,0} \end{bmatrix}.$$

Level(0,m). In this case $\tilde{\mathbb{P}}_{0,m} = (\tilde{p}^0_{0,m})$ is a scalar function in y and

$$\mathbb{P}_{0,m} = \begin{bmatrix} \tilde{p}^0_{0,0} \\ \tilde{p}^0_{0,1} \\ \vdots \\ \tilde{p}^0_{0,m} \end{bmatrix}, \text{ i.e., } p^k_{0,m} = \tilde{p}^0_{0,k}. \tag{6.4}$$

We have two new parameters, $s_{0,2m-1}$ and $s_{0,2m}$, corresponding to the moments $\langle y^{2m-1} \rangle$ and $\langle y^{2m} \rangle$.

Clearly $\Gamma_{0,m} = U_m$, $I_{0,m}$ is the $(m+1) \times 1$ matrix $(0,0,\ldots,0,1)^\top$.

We take $\tilde{A}_{0,m} = s_{0,2m} > 0$ and $\tilde{B}_{0,m-1} = s_{2m-1,0}$. Then

$$\tilde{p}^0_{0,m} = \tilde{A}^{-1}_{0,m}(y \tilde{p}^0_{0,m-1} - \tilde{B}_{0,m-1} \tilde{p}^0_{0,m-1} - \tilde{A}_{0,m-1} \tilde{p}^0_{0,m-2}) \tag{6.5}$$

and

$$J^1_{0,m} = \begin{bmatrix} \tilde{B}_{0,0} & \tilde{A}_{0,1} & & & & \\ \tilde{A}_{0,1} & \tilde{B}_{0,1} & \tilde{A}_{0,2} & & & \\ & \tilde{A}_{0,2} & \tilde{B}_{0,2} & \tilde{A}_{0,3} & & \\ & & \ddots & \ddots & \ddots & \\ & & & \tilde{A}_{0,m-1} & \tilde{B}_{0,m-1} & \tilde{A}_{0,m} \end{bmatrix}.$$

Level(n,m). If $n \geq 1$ and $m \geq 1$ we have four new parameters $s_{2n-1,2m-1}$, $s_{2n-1,2m}$, $s_{2n,2m-1}$, and $s_{2n,2m}$, corresponding to the moments $\langle x^{2n-1}y^{2m-1}\rangle$, $\langle x^{2n-1}y^{2m}\rangle$, $\langle x^{2n}y^{2m-1}\rangle$, and $\langle x^{2n}y^{2m}\rangle$.

Step (n,m): $\mathcal{K}_{n,m}$ In $\mathcal{K}_{n,m}$ the only new entry is $(\mathcal{K}_{n,m})_{m,n}$. (Everything else can be recovered from the previous levels.) Indeed, if $m > 1$ we show below that all rows except the last one can be obtained from (4.1). Notice that $\Gamma_{n,m-1}E_{m,m} = 0$ and therefore

$$\Gamma_{n,m-1} = \Gamma_{n,m-1}(U_{m-1}^\top U_{m-1} + E_{m,m}) = \Gamma_{n,m-1}U_{m-1}^\top U_{m-1}.$$

But we know that $\Gamma_{n,m-1}U_{m-1}^\top$ is an invertible matrix, which allows us to rewrite (4.1) as follows:

$$U_{m-1}\mathcal{K}_{n,m} = -(\Gamma_{n,m-1}U_{m-1}^\top)^{-1}(J^2_{n,m-1} + \mathcal{K}_{n,m-1}\tilde{B}_{n-1,m-1})\tilde{A}^{\top\,-1}_{n-1,m}.$$

The matrix $U_{m-1}\mathcal{K}_{n,m}$ is the $(m-1)\times n$ matrix obtained from $\mathcal{K}_{n,m}$ by deleting the last row.

Similarly, if $n > 1$ we can write

$$\tilde{\Gamma}_{n-1,m} = \tilde{\Gamma}_{n-1,m}(U_{n-1}^\top U_{n-1} + E_{n,n}) = \tilde{\Gamma}_{n-1,m}U_{n-1}^\top U_{n-1},$$

i.e., $\tilde{\Gamma}^\top_{n-1,m} = U^\top_{n-1}(U_{n-1}\tilde{\Gamma}^\top_{n-1,m})$, and (4.2) can be rewritten as

$$\mathcal{K}_{n,m}U_{n-1}^\top = -A^{-1}_{n,m-1}(\tilde{J}^{2\,\top}_{n-1,m} + B_{n-1,m-1}\mathcal{K}_{n-1,m})(U_{n-1}\tilde{\Gamma}^\top_{n-1,m})^{-1}.$$

Thus the $m\times(n-1)$ matrix $\mathcal{K}_{n,m}U_{n-1}^\top$, which is obtained from $\mathcal{K}_{n,m}$ by deleting the last column, is known from the previous levels. This allows us to compute all entries in the last row of $\mathcal{K}_{n,m}$ except $(\mathcal{K}_{n,m})_{m,n}$. Finally we put $(\mathcal{K}_{n,m})_{m,n} = s_{2n-1,2m-1}$. From the computation of $\Gamma_{n,m}$ below we see that the parameters must be chosen so that $I - \mathcal{K}_{n,m}\mathcal{K}^\top_{n,m}$ is positive definite.

Step (n,m): $\Gamma_{n,m}$ If $||\mathcal{K}_{n,m}|| < 1$, the matrix $I - \mathcal{K}_{n,m}\mathcal{K}^\top_{n,m}$ is symmetric and positive definite. Rewriting (3.24) as

$$I - \mathcal{K}_{n,m}\mathcal{K}^\top_{n,m} = (\Gamma_{n,m}U_m^\top)(\Gamma_{n,m}U_m^\top)^\top,$$

we see that $\Gamma_{n,m}U_m^\top$ (which is $\Gamma_{n,m}$ except the last zero column) is the lower-triangular factor in the Cholesky factorization of the matrix $I - \mathcal{K}_{n,m}\mathcal{K}^\top_{n,m}$.

Step (n,m): $J^2_{n,m}$ The computation of $J^2_{n,m}$ is similar to the computation of $\mathcal{K}_{n,m}$. First, if $m > 1$ we can write (4.3) in the form

$$U_{m-1}J^2_{n,m} = (\Gamma_{n,m-1}U_{m-1}^\top)^{-1}(-J^1_{n,m-1}\mathcal{K}_{n,m} + \mathcal{K}_{n,m-1}\tilde{A}_{n-1,m}),$$

which gives all entries of $J^2_{n,m}$ except the entries in the last row. Rewriting equation (4.4) as

$$\begin{aligned} J^2_{n,m}U^\top_{n-1} = &- A^{-1}_{n,m-1}(J^1_{n-1,m}A^\top_{n-1,m}I_{n-2,m} - J^2_{n-1,m}\tilde{\Gamma}_{n-1,m}\tilde{J}^{1\top}_{n-1,m} \\ &+ J^2_{n-1,m}\mathcal{K}^\top_{n-1,m}\tilde{J}^{2\top}_{n-1,m} + J^3_{n-1,m}I^\top_{n-2,m-1}\tilde{J}^{3\top}_{n-1,m} \\ &+ B_{n-1,m-1}J^2_{n-1,m} - A^\top_{n-1,m-1}I_{n-2,m-1}\tilde{A}_{n-2,m} \\ &+ A_{n,m-1}J^3_{n,m}I^\top_{n-1,m-1}\mathcal{K}_{n-1,m})(U_{n-1}\tilde{\Gamma}^\top_{n-1,m})^{-1} \end{aligned}$$

allows us to compute everything except the last column of $J^2_{n,m}$.

Finally, we put $(J^2_{n,m})_{m,n} = s_{2n-1,2m}$.

Computing $J^1_{n,m}$ We put $(J^1_{n,m})_{m,m} = s_{2n,2m-1}$ and $(J^1_{n,m})_{m,m+1} = s_{2n,2m}$ using the last two parameters. Everything else can be recovered from the previous levels. Rewriting (4.13) as

$$U_{m-1}J^1_{n,m} = (\Gamma_{n,m-1}U^\top_{m-1})^{-1}J^1_{n,m-1}\Gamma_{n,m},$$

we get the matrix obtained from $J^1_{n,m}$ by deleting the last row.

Consider now the $(m+1)\times(m-1)$ matrix $\Gamma^\top_{n,m}\Gamma^\top_{n,m-1}$. It is easy to see that the last two rows of this matrix are zeros and deleting these two rows we obtain an $(m-1)\times(m-1)$ upper-triangular matrix with positive entries on the main diagonal. Therefore the matrix $U_{m-1}U_m\Gamma^\top_{n,m}\Gamma^\top_{n,m-1}$ is invertible. Since $U^\top_m U^\top_{m-1}U_{m-1}U_m = I_{m+1} - E_{m,m} - E_{m+1,m+1}$, we can write

$$\Gamma^\top_{n,m}\Gamma^\top_{n,m-1} = U^\top_m U^\top_{m-1}(U_{m-1}U_m\Gamma^\top_{n,m}\Gamma^\top_{n,m-1}).$$

Combining this with formula (4.14) we see that

$$J^1_{n,m}U^\top_m U^\top_{m-1} = (J^{1\top}_{n,m-1} + J^3_{n,m}\mathcal{K}^\top_{n,m-1} + J^2_{n,m}\mathcal{K}^\top_{n,m}\Gamma^\top_{n,m-1})(U_{m-1}U_m\Gamma^\top_{n,m}\Gamma^\top_{n,m-1})^{-1}.$$

The matrix $J^1_{n,m}U^\top_m U^\top_{m-1}$ is obtained from $J^1_{n,m}$ by deleting the last two columns. This completes the computation of $J^1_{n,m}$.

Computing $A_{n,m}$ Let us denote by $\mathcal{M}_{n-1,m}$ the $(m+1)\times(m+1)$ matrix obtained by adding the last row of $J^1_{n-1,m}$ to the bottom of $\Gamma_{n-1,m}$. This is an upper-triangular invertible matrix. Using (4.15) and the last row of (4.16) we obtain a formula for $\mathcal{M}_{n-1,m}A_{n,m}$ in terms of known matrices, which allows us to compute $A_{n,m}$.

Computing $I_{n,m}$ Writing (4.17) as

$$U_m I_{n,m} = -(\Gamma_{n,m}U^\top_m)^{-1}\mathcal{K}_{n,m}\tilde{\Gamma}_{n,m},$$

we can compute all entries of $I_{n,m}$ except the last row. But the last row is simply $(0,0,\ldots,0,1)$, which completes the computation of $I_{n,m}$.

Computing $B_{n-1,m}$ Similarly to $A_{n,m}$, we can combine (4.19) and the last of (4.20) to obtain a formula for $\mathcal{M}_{n-1,m}B_{n-1,m}$ in terms of known matrices.

Computing $\tilde{A}_{n,m}$, $\tilde{J}^1_{n,m}$, $\tilde{J}^2_{n,m}$, and $\tilde{B}_{n,m-1}$ from (4.26), (4.24), (4.25), and (4.27).

Computing $\mathbb{P}_{n,m}(x,y)$ using (3.13) and (3.12). Similar to the computation of $A_{n,m}$ we obtain a formula for $\mathcal{M}_{n,m}\mathbb{P}_{n,m}$ in terms of known expressions, where $\mathcal{M}_{n,m}$ denotes the $(m+1)\times(m+1)$ matrix obtained by adding the last row of $J^1_{n,m}$ to the bottom of $\Gamma_{n,m}$. $\tilde{\mathbb{P}}_{n,m}(x,y)$ can be computed from the relation (3.14) analogous to (3.14) for $\tilde{\mathbb{P}}_{n,m}$.

7. Construction of the linear functional. The above algorithm allows us to find a linear functional given the coefficients in the recurrence formulas. More precisely, we have the next theorem.

THEOREM 7.1. $s_{0,0}, \dots, s_{2n,2m} \in \mathbb{R}$
- $A_{i+1,0}$ $B_{i,0}$ $i = 0, \dots, n-1$ $\tilde{A}_{0,j+1}, \tilde{B}_{0,j}$ $j = 0, \dots, m-1$
- $j \times i$ $\mathcal{K}_{i,j}$ $J^2_{i,j}$ $i = 1, \dots, n$ $j = 1, \dots, m$
- $j \times (j+1)$ $J^1_{i,j}$ $i = 1, \dots, n$ $j = 1, 2 \dots, m$

(7.1) $$s_{2i,2j} > 0 \quad ||\mathcal{K}_{i,j}|| < 1,$$

$\mathcal{L}$

(7.2) $$\mathcal{L}(\mathbb{P}_{i,m}, \mathbb{P}_{j,m}) = \delta_{i,j} I_{m+1} \quad \mathcal{L}(\tilde{\mathbb{P}}_{n,i}, \tilde{\mathbb{P}}_{n,j}) = \delta_{i,j} I_{n+1}.$$

(7.1)

7.2. The condition $||\mathcal{K}_{i,j}|| < 1$ imposes restrictions on the parameters $s_{i,j}$. In particular, it forces $|s_{2i-1,2j-1}| < 1$ for $i = 1, \dots, n$, and $j = 1, \dots, m$.

. We construct the linear functional by induction. First, if $n = m = 0$ we set

$$\mathcal{L}(1) = \frac{1}{s^2_{0,0}} \text{ and } p_{0,0} = \tilde{p}_{0,0} = s_{0,0},$$

and thus $\mathcal{L}(p_{0,0}, p_{0,0}) = \mathcal{L}(\tilde{p}_{0,0}, \tilde{p}_{0,0}) = 1$.

If $m = 0$, we construct $\mathbb{P}_{n,0} = p^0_{n,0}$ using (6.3) and then we define

$$\mathcal{L}(\mathbb{P}_{i,0}, \mathbb{P}_{j,0}) = \delta_{i,j}.$$

This gives a well-defined positive linear functional on x^j for $j = 0, 1 \dots, n$.

Likewise, if $n = 0$, we construct $\tilde{\mathbb{P}}_{0,k} = \tilde{p}^0_{0,k}$ using (6.5) and define

$$\mathcal{L}(\tilde{\mathbb{P}}_{0,i}, \tilde{\mathbb{P}}_{0,j}) = \delta_{i,j},$$

which defines the linear functional on y^j for $j = 0, 1 \dots, m$. Thus formula (7.2) will hold if $m = 0$ or $n = 0$.

Assume now that the functional $\mathcal{L}$ is defined for all levels before (n, m). We first extend $\mathcal{L}$ so that

(7.3) $$\mathcal{L}(\mathbb{P}_{n,m-1}, \tilde{\mathbb{P}}_{n-1,m}) = \mathcal{K}_{n,m}.$$

To check that the above equation is consistent with how $\mathcal{L}$ is defined on the previous levels, note that

(7.4) $$\mathcal{L}(\Gamma_{n,m-1}\mathbb{P}_{n,m-1}, \tilde{\mathbb{P}}_{n-1,m}) = \Gamma_{n,m-1}\mathcal{K}_{n,m},$$

which follows from the construction of $\mathcal{K}_{n,m}$ and the definition of $\mathcal{L}$ on the previous levels (see Lemma 4.1). Similarly, using the second defining relation of $\mathcal{K}_{n,m}$ (i.e., the last row of (4.2)) we see that

(7.5) $$\mathcal{L}(E_{m,m}\mathbb{P}_{n,m-1}, \tilde{\mathbb{P}}_{n-1,m}\tilde{\Gamma}^\top_{n-1,m}) = E_{m,m}\mathcal{K}_{n,m}\tilde{\Gamma}^\top_{n-1,m}.$$

Equations (7.4) and (7.5) show that (7.3) is automatically true except the equality of the entries at (m, n) place (i.e., the definition of the linear functional on the previous levels and the construction of $\mathcal{K}_{n,m}$ imply most of (7.3)). We use the (m, n) entry to extend the functional on $x^{2n-1}y^{2m-1}$, i.e., we define $\mathcal{L}(x^{2n-1}y^{2m-1})$ so that (7.3) holds.

Using the same arguments as in the proof of (3.25) we show that

$$J^3_{n,m} = -\mathcal{K}_{n,m}\tilde{A}^{\top}_{n-1,m} = -\mathcal{L}(y\mathbb{P}_{n,m-1}, \tilde{\mathbb{P}}_{n-1,m-1}). \tag{7.6}$$

Similar to $\mathcal{K}_{n,m}$ we can use the construction of $J^2_{n,m}$ to extend the functional on $x^{2n-1}y^{2m}$ so that

$$J^2_{n,m} = -\mathcal{L}(y\mathbb{P}_{n,m-1}, \tilde{\mathbb{P}}_{n-1,m}). \tag{7.7}$$

Finally we use $J^1_{n,m}$ to extend the functional on $x^{2n}y^{2m-1}$ and $x^{2n}y^{2m}$ in such a way that

$$J^1_{n,m} = \mathcal{L}(y\mathbb{P}_{n,m-1}, \mathbb{P}_{n,m}). \tag{7.8}$$

This completes the extension of the linear functional. It remains to show that the orthogonality relations (7.2) hold. Recall that $\mathbb{P}_{n,m}$ is constructed by using (3.12) and the last row of (3.13). The orthogonality relations in the previous levels and (7.3), (7.6), and (7.7) imply that

$$\mathcal{L}(\Gamma_{n,m}\mathbb{P}_{n,m}, \tilde{\mathbb{P}}_{n-1,k}) = \mathcal{L}(E_{m,m}J^1_{n,m}\mathbb{P}_{n,m}, \tilde{\mathbb{P}}_{n-1,k}) = 0 \text{ for } k = 0, 1, \ldots, m,$$

hence

$$\mathcal{L}(\mathbb{P}_{n,m}, \tilde{\mathbb{P}}_{n-1,k}) = 0 \text{ for } k = 0, 1, \ldots, m.$$

From the last equation it follows that

$$\mathcal{L}(\mathbb{P}_{n,m}, \mathbb{P}_{k,m}) = 0 \text{ for } k = 0, 1, \ldots, n-1. \tag{7.9}$$

It remains to show that

$$\mathcal{L}(\mathbb{P}_{n,m}, \mathbb{P}_{n,m}) = I_{m+1}. \tag{7.10}$$

This can be derived from the two equalities

$$\mathcal{L}(\Gamma_{n,m}\mathbb{P}_{n,m},\, \Gamma_{n,m}\mathbb{P}_{n,m}) = \Gamma_{n,m}\Gamma^{\top}_{n,m},$$
$$\mathcal{L}(E_{m,m}J^1_{n,m}\mathbb{P}_{n,m},\, \mathbb{P}_{n,m}) = E_{m,m}J^1_{n,m}.$$

Conversely, one can easily show that conditions (7.1) are necessary. Indeed, $s_{2i,2j} > 0$ follows from the normalization in (3.2) that the coefficient of the highest term is positive and (3.13). (3.24) shows that $\mathcal{K}_{i,j}$ must be a contraction, i.e., $||\mathcal{K}_{i,j}|| < 1$. □

7.3. The above construction gives simple criteria for the existence of a one-step extension of the functional. That is, given moments so that there exists a positive linear functional on $\prod^{2n-2,2m} \cup \prod^{2n,2m-2}$, any set

$$\{h_{2n-1,2m-1},\, h_{2n-1,2m},\, h_{2n,2m-1},\, h_{2n,2m}\}$$

that satisfies (7.1) can be used to extend the functional to $\prod^{2n,2m}$.

8. Interpretation of the condition $\mathcal{K}_{n,m} = 0$. In this section we classify two variable orthogonal polynomials, which can be obtained as a tensor product of two sets of (one variable) orthogonal polynomials. In other words, we want to see when $\mathbb{P}_{i,m}(x,y)$ can be written as

$$\mathbb{P}_{i,m}(x,y) = p_i(x) \begin{bmatrix} \tilde{p}_0(y) \\ \vdots \\ \tilde{p}_m(y) \end{bmatrix} \tag{8.1}$$

for some orthogonal polynomials $p_i(x)$ and $\tilde{p}_j(y)$. The next proposition lists simple implications of (8.1).

PROPOSITION 8.1. [illegible] (8.1) [illegible] $i = 1, 2, \dots, m$ [illegible] $i = 1, 2, \dots, m$ [illegible] $j = 1, 2, \dots, n$ [illegible]

$$\mathbb{P}_{i,j}(x,y) = p_i(x) \begin{bmatrix} \tilde{p}_0(y) \\ \\ \tilde{p}_j(y) \end{bmatrix}, \qquad \tilde{\mathbb{P}}_{i,j}(x,y) = \tilde{p}_j(y) \begin{bmatrix} p_0(x) \\ \\ p_i(x) \end{bmatrix}, \tag{8.2}$$

[illegible]

(8.3)

$$\mathcal{K}_{ij} = 0, \quad J^2_{i,j} = 0, \quad A_{i,j} = a_i I_{j+1}, \quad \tilde{A}_{i,j} = \tilde{a}_j I_{i+1},$$

$$J^1_{i,j} = \begin{bmatrix} \tilde{b}_0 & \tilde{a}_1 & & & \\ \tilde{a}_1 & \tilde{b}_1 & \tilde{b}_2 & & \\ & \tilde{a}_2 & \tilde{b}_2 & \tilde{a}_3 & \\ & & & & \\ & & & \tilde{a}_{j-1} & \tilde{b}_{j-1} & \tilde{a}_j \end{bmatrix}, \; \tilde{J}^1_{i,j} = \begin{bmatrix} b_0 & a_1 & & & \\ a_1 & b_1 & b_2 & & \\ & a_2 & b_2 & a_3 & \\ & & & & \\ & & & a_{i-1} & b_{i-1} & a_i \end{bmatrix}.$$

[illegible] a_i, b_{i-1} [illegible] $i = 1, 2, \dots, n$ [illegible] $\tilde{a}_j, \tilde{b}_{j-1}$ [illegible] $j = 1, 2, \dots, m$ [illegible] [illegible] $p_i(x)$ [illegible] $\tilde{p}_j(y)$ [illegible]

$$x p_i(x) = a_{i+1} p_{i+1}(x) + b_i p_i(x) + a_i p_{i-1}(x), \tag{8.4}$$

$$y \tilde{p}_j(x) = \tilde{a}_{j+1} \tilde{p}_{j+1}(y) + \tilde{b}_j \tilde{p}_j(y) + \tilde{a}_j \tilde{p}_{j-1}(y). \tag{8.5}$$

[illegible]. If (8.1) holds, then the orthogonality $\langle \mathbb{P}_{i_1,m}, \mathbb{P}_{i_2,m} \rangle = \delta_{i_1,i_2} I_{m+1}$ is equivalent to

$$\langle p_{i_1}(x)\tilde{p}_{j_1}(y),\, p_{i_2}(x)\tilde{p}_{j_2}(y) \rangle = \delta_{i_1,i_2}\delta_{j_1,j_2}. \tag{8.6}$$

From this relation one can easily obtain (8.2) and (8.3). □

[illegible] 8.2. Notice that if $p_i(x)$ and $\tilde{p}_j(y)$ satisfy (8.2) and c is a nonzero constant, then the polynomials

$$q_i(x) = c p_i(x) \text{ and } \tilde{q}_j(y) = \tilde{p}_j(y)/c \tag{8.7}$$

also satisfy (8.2). Conversely, if two pairs $p_i, \tilde{p}_j$ and $q_i, \tilde{q}_j$ of scalar orthogonal polynomials satisfy (8.2), then there is a nonzero constant such that (8.7) holds. Thus the polynomials $p_i, \tilde{p}_j$ are unique up to a multiplicative constant.

LEMMA 8.3. [illegible] $\mathcal{K}_{i,j} = 0$ [illegible] $i = 1, \ldots, (n+1)$ [illegible] $j = 1, \ldots, (m+1)$ [illegible] (8.2) [illegible] $i \leq n$ [illegible] $j \leq m$ [illegible]. The proof is by induction. This is obvious if $n = 0$ or $m = 0$. Assume now that this is true for levels $(n-1, m)$ and $(n, m-1)$. We need to show that it is also true for level (n, m). The induction hypothesis and Proposition 8.1 show that (8.2) and (8.3) hold

- if $i \leq n-1$ and $j \leq m$, or
- if $i \leq n$ and $j \leq m-1$.

In particular, $A_{n,m-1} = a_n I_m$. (3.24), (3.25), (4.1), and (4.2), combined with $\mathcal{K}_{i,j} = 0$ for $i \leq n+1$ and $j \leq m+1$, imply that

$$\Gamma_{i,j} = [I_j | 0], \ J^3_{i,j} = 0, \ \text{ and } \ J^2_{i,j} = 0 \ \text{ for } \ i \leq n+1, \ j \leq m. \tag{8.8}$$

Using (8.8) and the defining relations (4.13)–(4.14) for $J^1_{n,m}$ we see that $J^1_{n,m}$ has the form

$$J^1_{n,m} = \begin{bmatrix} \tilde{b}_0 & \tilde{a}_1 & & & & & \\ \tilde{a}_1 & \tilde{b}_1 & \tilde{b}_2 & & & & \\ & \tilde{a}_2 & \tilde{b}_2 & \tilde{a}_3 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \tilde{a}_{m-2} & \tilde{b}_{m-2} & \tilde{a}_{m-1} & \\ & & & & \tilde{a}_{m-1} & c_1 & c_2 \end{bmatrix} \tag{8.9}$$

for some scalars c_1, c_2. (4.15) combined with (8.3) and (8.8) shows that

$$A_{n,m} = \begin{bmatrix} a_n & & & & & \\ & a_n & & & & \\ & & a_n & & & \\ & & & \ddots & & \\ & & & & a_n & \\ d_1 & d_2 & \ldots & d_{m-1} & d_m & d \end{bmatrix} \tag{8.10}$$

for some scalars $d_1, d_2, \ldots, d_m, d$. Plugging (8.9), (8.10), and $A_{n,m-1} = a_n I_m$ into (4.16) and comparing the entries on the last row, we get the equalities

$$\begin{aligned} \tilde{a}_m d_1 &= 0, \\ \tilde{a}_m d_2 &= 0, \\ &\vdots \\ \tilde{a}_m d_{m-2} &= 0, \\ \tilde{a}_{m-1} a_n + \tilde{a}_m d_{m-1} &= a_n \tilde{a}_{m-1}, \\ \tilde{b}_{m-1} a_n + \tilde{a}_m d_m &= a_n c_1, \\ \tilde{a}_m d &= a_n c_2. \end{aligned} \tag{8.11}$$

The first $(m-1)$ equations simply imply that $d_1 = d_2 = \cdots = d_{m-1} = 0$.

Formula (4.4) (for $J^2_{n+1,m}$) becomes

$$0 = J^1_{n,m} A^\top_{n,m} I_{n-1,m} - A^\top_{n,m-1} I_{n-1,m-1} \tilde{A}_{n-1,m}.$$

Using (8.3) we can rewrite the last equation as

$$J^1_{n,m}A^\top_{n,m}I_{n-1,m} = a_n\tilde{a}_m I_{n-1,m-1}.$$

On the other hand, from (4.17) it follows that for $i \le n$ and $j \le m$ all entries, except $(j+1, i+1)$, of the matrix $I_{i,j}$ are equal to zero. Thus the last equality simply means that the last column of the matrix $J^1_{n,m}A^\top_{n,m}$ is equal to $(0, 0, \dots, 0, a_n\tilde{a}_m)^\top$. Using (8.9) and (8.10) we see that the bottom two entries of the last column of the matrix $J^1_{n,m}A^\top_{n,m}$ are $\tilde{a}_{m-1}d_m$ and $c_2 d$, i.e., we have

$$\begin{aligned} \tilde{a}_{m-1}d_m &= 0, \\ c_2 d &= a_n\tilde{a}_m. \end{aligned} \tag{8.12}$$

The first equation implies that $d_m = 0$, while the second one combined with (8.15) implies that $d^2 = a_n^2$ and $c_2^2 = \tilde{a}_m^2$. Since all the numbers d, c_2, a_n, and $\tilde{a}_m$ are positive, it follows that $d = a_n$ and $c_2 = \tilde{a}_m$. Finally notice that the $(m-1)$st equation in formula (8.11) gives $c_1 = \tilde{b}_{m-1}$. Thus we proved that

$$A_{n,m} = a_n I_{m+1} \text{ and } J^1_{n,m} = \begin{bmatrix} \tilde{b}_0 & \tilde{a}_1 & & & & \\ \tilde{a}_1 & \tilde{b}_1 & \tilde{b}_2 & & & \\ & \tilde{a}_2 & \tilde{b}_2 & \tilde{a}_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \tilde{a}_{m-1} & \tilde{b}_{m-1} & \tilde{a}_m \end{bmatrix}. \tag{8.13}$$

From the last formula and (3.12), (3.13), and (3.14), one can easily see that (8.2)–(8.3) hold for $i = n$ and $j = m$, which completes the proof. □

As a corollary we get the following theorem.

THEOREM 8.4. [illegible]

(i) [illegible] n [illegible] m $\mathbb{P}_{n,m}$ [illegible] $p_k(x)$ [illegible] $\tilde{p}_j(y)$ [illegible]

$$\mathbb{P}_{n,m}(x,y) = p_n(x)\begin{bmatrix} \tilde{p}_0(y) \\ \\ \tilde{p}_m(y) \end{bmatrix}. \tag{8.14}$$

(ii) $\mathcal{K}_{n,m} = 0$ [illegible] $n, m = 1, 2, \dots$

Next we want to prove a finite analogue of Theorem 8.4, i.e., to give necessary and sufficient conditions for (8.4) to hold up to a given level (n, m). We need the following lemma.

LEMMA 8.5. [illegible] $\mathcal{K}_{i,j} = 0$ [illegible] $(J^2_{i,j})_{j,i} = 0$ [illegible] $i = 1, \dots, n$ [illegible] $j = 1, \dots, m$ [illegible] (8.2) [illegible] $i < n$ [illegible] $j \le m$ [illegible] $i \le n$ [illegible] $j < m$ [illegible]. We prove the statement by induction. If $n = 0$ or $m = 0$ this is a trivial statement. Assume now that this is true for levels $(n-1, m)$ and $(n, m-1)$. We will show that it also holds for level (n, m). The induction hypothesis means that (8.2) and (8.3) hold

- if $i \le n-2$ and $j \le m$, or
- if $i \le n-1$ and $j \le m-1$, or
- if $i \le n$ and $j \le m-2$.

As before, the condition $\mathcal{K}_{i,j} = 0$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$ implies that

$$J^3_{i,j} = 0 \text{ and } \Gamma_{i,j} = [I_j|0] \text{ for } i = 1, \ldots, n \text{ and } j = 1, \ldots, m. \tag{8.15}$$

Formula (4.2) gives

$$\tilde{J}^2_{n-1,m} = 0,$$

which combined with (4.25) for $\tilde{J}^2_{n-1,m}$ shows that

$$I^\top_{n-2,m} A_{n-1,m} \Gamma^\top_{n-1,m} = 0. \tag{8.16}$$

Using the induction hypothesis and (4.3)–(4.4) one can prove that $J^2_{n-1,m} = 0$. Indeed, the right-hand side of (4.3) is obviously zero, while on the right-hand side of (4.4) there are only two nonzero terms:

$$J^1_{n-2,m} A^\top_{n-2,m} I_{n-3,m} - A^\top_{n-2,m-1} I_{n-3,m-1} \tilde{A}_{n-3,m}.$$

Using (8.3) one can easily see that the above expression vanishes. This shows that all entries of $J^2_{n-1,m}$ are equal to 0, except probably $(J^2_{n-1,m})_{m,n-1}$, but we know that this entry is 0, because this is one of the assumptions in the lemma.

Formula (8.3) and the defining relations (4.13)–(4.14) for $J^1_{n-1,m}$ show that it has the form

$$J^1_{n-1,m} = \begin{bmatrix} \tilde{b}_0 & \tilde{a}_1 & & & & & \\ \tilde{a}_1 & \tilde{b}_1 & \tilde{b}_2 & & & & \\ & \tilde{a}_2 & \tilde{b}_2 & \tilde{a}_3 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \tilde{a}_{m-2} & \tilde{b}_{m-2} & \tilde{a}_{m-1} & \\ & & & & \tilde{a}_{m-1} & c_1 & c_2 \end{bmatrix} \tag{8.17}$$

for some scalars c_1, c_2. (4.15) for $A_{n-1,m}$ together with (8.3) imply

$$A_{n-1,m} = \begin{bmatrix} a_{n-1} & & & & & & \\ & a_{n-1} & & & & & \\ & & a_{n-1} & & & & \\ & & & & \ddots & & \\ & & & & & a_{n-1} & \\ d_1 & d_2 & \ldots & d_{m-1} & d_m & d \end{bmatrix} \tag{8.18}$$

for some scalars $d_1, d_2, \ldots, d_m, d$. Plugging (8.18) into (8.16) and using (8.15) we see that $d_1 = d_2 = \cdots = d_m = 0$ and $A_{n-1,m}$ is a diagonal matrix of the form

$$A_{n-1,m} = \begin{bmatrix} a_{n-1} & & & & & \\ & a_{n-1} & & & & \\ & & a_{n-1} & & & \\ & & & \ddots & & \\ & & & & a_{n-1} & \\ & & & & & d \end{bmatrix}. \tag{8.19}$$

Writing (4.16) for $n-1$, plugging (8.17) and (8.19), and comparing the last two entries on the last row we get the equalities

$$\tilde{b}_{m-1}a_{n-1} = a_{n-1}c_1, \tag{8.20}$$

$$\tilde{a}_m d = a_{n-1}c_2. \tag{8.21}$$

The first equality gives $c_1 = \tilde{b}_{m-1}$.

Finally, notice that the last column of the left-hand side of (4.4) must be 0, which, in particular, implies that the last entry in the last column on the right-hand side of (4.4) must be zero. Computing this entry we get

$$c_2 d - a_{n-1}\tilde{a}_m = 0. \tag{8.22}$$

(8.21) and (8.22) show that $c_2 = \tilde{a}_m$ and $d = a_{n-1}$. Thus

$$A_{n-1,m} = a_{n-1}I_{m+1} \quad \text{and} \quad J^1_{n-1,m} = \begin{bmatrix} \tilde{b}_0 & \tilde{a}_1 & & & & & \\ \tilde{a}_1 & \tilde{b}_1 & \tilde{b}_2 & & & & \\ & \tilde{a}_2 & \tilde{b}_2 & \tilde{a}_3 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \tilde{a}_{m-2} & \tilde{b}_{m-2} & \tilde{a}_{m-1} & \\ & & & & \tilde{a}_{m-1} & \tilde{b}_{m-1} & \tilde{a}_m \end{bmatrix},$$

which combined with (3.13) shows that (8.2)–(8.3) hold also for $i = n-1$ and $j = m$. Similarly one can show that (8.2)–(8.3) hold for $i = n$ and $j = m-1$, completing the induction. □

COROLLARY 8.6. [illegible] $n, m \geq 0$ [illegible]

(i) [illegible] $i = 0, 1, 2, \ldots, n$ $\mathbb{P}_{i,m}$ [illegible] $p_k(x)$ [illegible] $\tilde{p}_j(y)$ [illegible] (8.1) [illegible]

(ii)
- $\mathcal{K}_{i,j} = 0$ [illegible] $i \leq n$ $j \leq m$.
- $(J^2_{i,j})_{j,i} = 0$, $i = n$, $j = m$.
- $(J^1_{n,m})_{m,m+1} = \tilde{A}_{0,m}$ [illegible] $(J^1_{n,m})_{m,m} = \tilde{B}_{0,m-1}$

9. Numerical examples. In this last section we give several numerical examples which illustrate the algorithm.

[illegible] 9.1. Let us choose parameters $s_{i,j}$ as follows:

$s_{0,0} = 0.7543$	$s_{0,1} = 0$	$s_{0,2} = 0.4633$	$s_{0,3} = 0$	$s_{0,4} = 0.4997$
$s_{1,0} = 0$	$s_{1,1} = -0.1185$	$s_{1,2} = 0$	$s_{1,3} = -0.1128$	$s_{1,4} = 0$
$s_{2,0} = 0.4634$	$s_{2,1} = 0$	$s_{2,2} = 0.4681$	$s_{2,3} = 0$	$s_{2,4} = 0.4998$
$s_{3,0} = 0$	$s_{3,1} = -0.1128$	$s_{3,2} = 0$	$s_{3,3} = -0.1073$	$s_{3,4} = 0$
$s_{4,0} = 0.4997$	$s_{4,1} = 0$	$s_{4,2} = 0.4682$	$s_{4,3} = 0$	$s_{4,4} = 0.4997$

First we apply the algorithm at levels $(0,0)$, $(0,1)$, $(0,2)$, $(1,0)$, $(2,0)$. (There is no restriction on the parameters at these levels.) Next we compute $\mathcal{K}_{1,1} = [s_{1,1}]$:

$$\mathcal{K}_{1,1} = [-0.1185].$$

Since $|s_{1,1}| < 1$, $\mathcal{K}_{1,1}$ is a contraction and we can compute all other matrices and the polynomials at level $(1,1)$. Then we proceed to levels $(1,2)$ and $(2,1)$. We obtain

$$\mathcal{K}_{1,2} = \begin{bmatrix} 0 \\ -0.1128 \end{bmatrix}$$

and

$$\mathcal{K}_{2,1} = \begin{bmatrix} 0 & -0.1128 \end{bmatrix},$$

which are obviously contractions, hence we can compute everything else at these levels using the algorithm. Finally, at level $(2,2)$ we get

$$\mathcal{K}_{2,2} = \begin{bmatrix} 0.0174 & 0 \\ 0 & -0.1073 \end{bmatrix},$$

which again is a contraction and allows us to apply the algorithm.

The example described above corresponds to the following moment problem:

$$\begin{array}{lllll} h_{0,0} = 1.7575 & h_{0,1} = 0 & h_{0,2} = 0.3773 & h_{0,3} = 0 & h_{0,4} = 0.1752 \\ h_{1,0} = 0 & h_{1,1} = -0.0447 & h_{1,2} = 0 & h_{1,3} = -0.0196 & h_{1,4} = 0 \\ h_{2,0} = 0.3773 & h_{2,1} = 0 & h_{2,2} = 0.0838 & h_{2,3} = 0 & h_{2,4} = 0.0395 \\ h_{3,0} = 0 & h_{3,1} = -0.0196 & h_{3,2} = 0 & h_{3,3} = -0.0087 & h_{3,4} = 0 \\ h_{4,0} = 0.1752 & h_{4,1} = 0 & h_{4,2} = 0.0395 & h_{4,3} = 0 & h_{4,4} = 0.0188. \end{array}$$

Next we present an example of a moment problem which cannot be extended.

EXAMPLE 9.2. Let us fix the free parameters at levels $(0,0)$, $(0,1)$, $(0,2)$, $(1,0)$, $(2,0)$, and $(1,1)$ as follows:
Level $(0,0)$: $s_{0,0} = 1$;
Level $(0,1)$: $s_{0,1} = 1/7$, $s_{0,2} = 2$;
Level $(0,2)$: $s_{0,3} = 1/3$, $s_{0,4} = 7$;
Level $(1,0)$: $s_{1,0} = 1$, $s_{2,0} = 3$;
Level $(2,0)$: $s_{3,0} = 0$, $s_{4,0} = 0.33$;
Level $(1,1)$: $s_{1,1} = 0.5$, $s_{1,2} = 0$, $s_{2,1} = 0$, $s_{2,2} = 1$.

Applying the algorithm described in section 6, we obtain a functional $\mathcal{L}$ defined on the space $\{x^i y^j : i+j \le 2\}$ and the orthogonal polynomials $\mathbb{P}$ and $\tilde{\mathbb{P}}$ corresponding to this functional.

Computing $\mathcal{K}_{2,1}$ and $\mathcal{K}_{1,2}$ we obtain

$$\mathcal{K}_{2,1} = \begin{bmatrix} 0.9997407262 & s_{3,1} \end{bmatrix}, \qquad \mathcal{K}_{1,2} = \begin{bmatrix} -0.02749286996 \\ s_{1,3} \end{bmatrix},$$

which shows that if we pick $s_{3,1}$ and $s_{1,3}$ with absolute value less than 1 and such that $\mathcal{K}_{2,1}$ and $\mathcal{K}_{1,2}$ are contractions, we can extend the functional to the space $\{x^i y^j : i+j \le 3\}$. All other parameters at levels $(0,3)$, $(1,2)$, $(2,1)$, and $(3,0)$ can be chosen arbitrary. (Of course $s_{0,6}$, $s_{2,4}$, $s_{4,2}$ and $s_{6,0}$ must be positive.)

Finally let us compute $\mathcal{K}_{2,2}$. Entry $(1,1)$ of this matrix is

$$(\mathcal{K}_{2,2})_{1,1} = \frac{-59.47189 - 0.2717694 \times 10^{-10} s_{1,3} + 0.1087078 \times 10^{-11} s_{3,1} s_{1,3}^2 - 43.93372 s_{3,1} s_{1,3}}{\sqrt{1 - 1928.713 s_{3,1}^2}\sqrt{1 - 1.000756 s_{1,3}^2}(1 + 1.237179 \times 10^{-14} s_{1,3})}.$$

From the above formula it is clear that $|(\mathcal{K}_{2,2})_{1,1}| > 1$, which means that $\mathcal{K}_{2,2}$ is not a contraction. Thus, we see that the functional $\mathcal{L}$ cannot be extended to level

$(2,2)$ no matter how we choose the parameters at levels $(1,2)$ and $(2,1)$. In particular, it follows that $\mathcal{L}$ cannot be extended to the space of polynomials of (total) degree 3.

REMARK 9.3. The above example shows that not every functional defined on levels $(n, m-1)$ and $(n-1, m)$ can be extended to level (n, m) even if we modify the parameters entering one step back in each direction, that is, at levels $(n, m-1)$ and $(n-1, m)$. Several numerical experiments indicate that deforming the parameters two steps back in each direction is enough to extend the functional. Whether this is true or false in general is an interesting open problem.

REMARK 9.4. Example 9.2 shows the simplest possible case of a moment problem which cannot be extended to a level (n, m) even if we modify the parameters entering one step back in each direction. More precisely, one can easily show that if $n = 1$ or $m = 1$ the moment problem can always be extended by deforming just one parameter entering one step back. Indeed, if, for example, $m = 1$, then $\mathcal{K}_{n,1}$ is a $1 \times n$ matrix. The first $(n-1)$ entries are computed from (4.2). Notice that in this equation the only matrix coming from level $(n, 0)$ is $A_{n,0} = (s_{2n,0})$. Thus, if we make $s_{2n,0}$ large enough, $\mathcal{K}_{n,1}$ will be a contraction.

Acknowledgments. We would like to thank a referee for a careful reading of the manuscript and helpful suggestions. AMD and FM would like to thank the School of Mathematics at Georgia Tech for its hospitality and support.

REFERENCES

[1] Ju. M. Berezanskii, *Expansions in eigenfunctions of self-adjoint operators*, Trans. Math. Mono. Amer. Math. Soc., 17 (1968).

[2] Ju. M. Berezanskii, *Direct and inverse spectral problems for a Jacobi field*, Algebra i Analiz, 9 (1997), pp. 38–61; translation in St. Petersburg Math. J., 9 (1998), pp. 1053–1071.

[3] C. F. Dunkl, *Intertwining operators and polynomials associated with the symmetric group*, Monatsh. Math., 126 (1998), pp. 181–209.

[4] C. F. Dunkl and Y. Xu, *Orthogonal Polynomials of Several Variables*, Encyclopedia of Mathematics and Its Applications 81. Cambridge University Press, Cambridge, UK, 2001.

[5] L. Fernández, T. E. Pérez, and M. A. Piñar, *Weak classical orthogonal polynomials in two variables*, J. Comput. Appl. Math., 178 (2005), pp. 191–203.

[6] M. I. Gekhtman and A. A. Kalyuzhny, *On the orthogonal polynomials in several variables*, Integral Equations Operator Theory, 19 (1994), pp. 404–418.

[7] J. S. Geronimo, *Scattering theory and matrix orthogonal polynomials on the real line*, Circuits Systems Signals Process, 1 (1982), pp. 471–495.

[8] J. S. Geronimo and H. J. Woerdeman, *Positive extensions, Fejér-Riesz factorization and autoregressive filters in two variables*, Ann. of Math., 160 (2004), pp. 839–906.

[9] J. S. Geronimo and H. J. Woerdeman, *Two variable orthogonal polynomials on the bi-circle and structured matrices*.

[10] D. Jackson, *Formal properties of orthogonal polynomials in two variables*, Duke Math. J., 2 (1936), pp. 423–434.

[11] Y. J. Kim, K. H. Kwon, and J. K. Lee, *Multi-variate orthogonal polynomials and second order partial differential equations*, Commun. Appl. Anal., 6 (2002), pp. 479–504.

[12] T. H. Koornwinder, *Orthogonal polynomials in two variables which are eigenfunctions of two algebraically independent partial differential operators*, I, II, Indag. Math., 36 (1974), pp. 48–66.

[13] T. H. Koornwinder, *Two variable analogues of the classical orthogonal polynomials*, in Theory and Applications of Special Functions, R. Askey, ed., Academic Press, NY, 1975, pp. 435–495.

[14] T. H. Koornwinder, *Askey-Wilson polynomials for root systems of type BC*, in Hypergeometric Functions on Domains of Positivity, Jack Polynomials, and Applications, Contemp. Math., Amer. Math. Soc., 138 (1992), pp. 189–204.

[15] M. A. Kowalski, *The recursion formulas for orthogonal polynomials in n variables*, SIAM J. Math. Anal., 13 (1982), pp. 309–315.

[16] M. A. Kowalski, *Orthogonality and recursion formulas for polynomials in n variables*, SIAM J. Math. Anal., 13 (1982), pp. 316–323.
[17] H. L. Krall and I. M. Sheffer, *Orthogonal polynomials in two variables*, Ann. Mat. Pura Appl., 76 (1967), pp. 325–376.
[18] M. G. Krein, *Infinite J-matrices and the matrix moment problem*, Dokl. Akad. Nauk. SSSR, 69 (1949), pp. 125–128.
[19] K. H. Kwon, J. K. Lee, and L. L. Littlejohn, *Orthogonal polynomial eigenfunctions of second order partial differential equations*, Trans. Amer. Math. Soc., 353 (2001), pp. 3629–3647.
[20] I. G. Macdonald, *Symmetric Functions and Hall Polynomials*, Oxford University Press, New York, 1995.
[21] P. K. Suetin, *Orthogonal Polynomials in Two Variables*, Anal. Methods Spec. Funct., 3, Gordon and Breach Science Publishers, Amsterdam, 1999.
[22] Y. Xu, *On multivariate orthogonal polynomials*, SIAM J. Math. Anal., 24 (1993), pp. 783–794.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 1, pp. 148–169

LINEAR PERTURBATION THEORY FOR STRUCTURED MATRIX PENCILS ARISING IN CONTROL THEORY*

SHREEMAYEE BORA† AND VOLKER MEHRMANN‡

Abstract. We investigate the effect of linear perturbations on several structured matrix pencils arising in control theory. These include skew-symmetric/symmetric pencils arising in the computation of optimal H_∞ control and linear-quadratic control for continuous and discrete time systems.

Key words. H_∞ control, linear-quadratic control, Hamiltonian matrix, continuous-time control, discrete-time control, skew-Hamiltonian/Hamiltonian pencil, skew-Hermitian/Hermitian pencil

AMS subject classifications. 93B36, 93B40, 49N35, 65F15, 93B52, 93C05

DOI. 10.1137/040609355

1. Introduction.

In this paper we study the effects of linear perturbations on the spectra of structured matrix pencils arising in control theory. The results that we present complement and generalize general perturbation results for Hamiltonian matrices as they were recently studied in [14] and we also extend results in [21, 22, 23].

Our main motivation arises from the following classical problems in optimal and robust control. Consider a linear constant coefficient dynamical system of the form

$$E\dot{x} = Ax + Bu, \qquad x(\tau_0) = x^0, \tag{1.1}$$

where $x(\tau) \in \mathbb{C}^n$ is the state, x^0 is an initial vector, $u(\tau) \in \mathbb{C}^m$ is the control input of the system and the matrices $E, A \in \mathbb{C}^{n,n}$, $B \in \mathbb{C}^{n,m}$ are constant. Here we discuss only the case that the matrix E is nonsingular; thus we allow implicit systems but we do not discuss descriptor systems.

The objective in linear quadratic optimal control, see e.g., [12, 17] is to find a control law $u(\tau)$ such that the closed loop system is asymptotically stable and such that the performance criterion

$$\mathcal{S}(x,u) = \int_{\tau_0}^{\infty} \begin{bmatrix} x(\tau) \\ u(\tau) \end{bmatrix}^T \begin{bmatrix} Q & S \\ S^H & R \end{bmatrix} \begin{bmatrix} x(\tau) \\ u(\tau) \end{bmatrix} d\tau \tag{1.2}$$

is minimized, where $Q = Q^H \in \mathbb{C}^{n,n}$, $R = R^H \in \mathbb{C}^{m,m}$ is positive definite and $\left[\begin{smallmatrix} Q & S \\ S^H & R \end{smallmatrix}\right]$ is positive semidefinite. Here A^H denotes the transpose of the complex conjugate of $A \in \mathbb{C}^{n,n}$.

Application of the maximum principle [17, 20] leads to the problem of finding a stable solution to the two-point boundary value problem of Euler-Lagrange equations

$$N_c \begin{bmatrix} \dot{\mu} \\ \dot{x} \\ \dot{u} \end{bmatrix} = H_c \begin{bmatrix} \mu \\ x \\ u \end{bmatrix}, \qquad x(\tau_0) = x^0, \qquad \lim_{\tau\to\infty} \mu(\tau) = 0, \tag{1.3}$$

*Received by the editors June 3, 2004; accepted for publication (in revised form) by P. Van Dooren October 19, 2005; published electronically March 17, 2006. This work was completed during the stay of the first author at the Institut für Mathematik, TU Berlin. The authors were partially supported by Deutsche Forschungsgemeinschaft research grant Me 790/11-3.

http://www.siam.org/journals/simax/28-1/60935.html

†Department of Mathematics, Indian Institute of Technology Guwahati, Guwahati 781039, Assam, India (shbora@iitg.ac.in, bora@math.tu-berlin.de).

‡Institut für Mathematik, Ma 4-5, TU Berlin, Straße des 17. Juni 136, D-10623 Berlin, FRG (mehrmann@math.tu-berlin.de).

with the matrix pencil

$$H_c - \lambda N_c := \begin{bmatrix} 0 & A & B \\ A^H & Q & S \\ B^H & S^H & R \end{bmatrix} - \lambda \begin{bmatrix} 0 & E & 0 \\ -E^H & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{1.4}$$

It is well known that the finite eigenvalues of $H_c - \lambda N_c$ are symmetric with respect to the imaginary axis (and if the problem is real then also with respect to the real axis). If E is invertible, then under the usual control theoretic assumptions [17, 26, 27], this pencil has exactly n eigenvalues in the left half plane and n eigenvalues in the right half plane plus m infinite eigenvalues. Clearly then the pencil has a unique deflating subspace associated with the eigenvalues in the open left half complex plane. If E or R are not invertible, then the situation is more complex and different approaches can be taken, [4, 6, 7, 17]. In this paper we discuss mainly the case that E and R are invertible.

The solution of the boundary value problem (1.3) can be obtained in many different ways. The approach in most computer aided control design packages is to decouple the boundary value problem via the computation of the solution of an associated algebraic Riccati equation. But one may also directly solve the boundary value problem (1.3) by computing the generalized Schur form of the pencil $H_c - \lambda N_c$, [2, 17, 27, 26], i.e., one determines unitary matrices $P, Q \in \mathbb{C}^{2n+m,2n+m}$ such that

$$PN_cQ = \begin{bmatrix} N_{11} & N_{12} & N_{13} \\ 0 & N_{22} & N_{23} \\ 0 & 0 & N_{33} \end{bmatrix}, \qquad PH_cQ = \begin{bmatrix} H_{11} & H_{12} & H_{13} \\ 0 & H_{22} & H_{23} \\ 0 & 0 & H_{33} \end{bmatrix},$$

where the subpencil $H_{11} - \lambda N_{11}$ has all its eigenvalues in the left half plane, to decouple the forward and backward integration in the boundary value problem.

In this paper we study the perturbation theory for the eigenvalue problem (1.4). For this, several different types of perturbations should be considered as separate cases. If one uses classical methods that do not preserve the structure, like the QZ-algorithm [9], to compute the generalized Schur form in finite precision arithmetic, then the special structure of the pencil is ignored and hence the whole matrices H_c, N_c are subject to perturbations. We do not discuss this case here, since it is well analyzed in the monograph [24].

If one studies perturbation theory in order to deal with uncertainties in the data of the system, then the blocks E, A, B, Q, S, R are subject to perturbations of only the blocks E, A, B, since typically the matrices of the cost function are free to be chosen under the constraints that $\left[\begin{smallmatrix} Q & S \\ S^H & R \end{smallmatrix}\right]$ is positive semidefinite and R positive definite. Also one may study the particular case that $E = I$ is not perturbed.

In all cases it is essential to analyze whether the perturbations can lead to eigenvalues on the imaginary axis, in which case the spectral symmetry and the uniqueness of the deflating subspace associated with the open left half plane may be lost; see [8, 17, 21, 22, 23].

It is well known, see [17, 18], that the discrete-time analogue to the linear quadratic control problem leads to slightly different matrix pencils of the form

$$H_d - \lambda N_d = \begin{bmatrix} 0 & A & B \\ -E^H & Q & S \\ 0 & S^H & R \end{bmatrix} - \lambda \begin{bmatrix} 0 & E & 0 \\ -A^H & 0 & 0 \\ -B^H & 0 & 0 \end{bmatrix}. \tag{1.5}$$

Here the spectral symmetry is with respect to the unit circle, i.e., the finite eigenvalues come in pairs $\lambda, \frac{1}{\lambda}$ or quadruples $\lambda, \bar{\lambda}, \frac{1}{\lambda}, \frac{1}{\bar{\lambda}}$ in the case of real matrices.

The perturbation problems can be discussed analogously and here the important question that arises is the study of perturbations which lead to eigenvalues on the unit circle, where again the spectral symmetry and the uniqueness of the deflating subspace associated with the eigenvalues in the open unit disk may be disturbed.

The second motivation comes from the optimal H_∞-control problem which arises in the context of robust control in frequency domain, see, e.g., the recent monographs [10, 28]. In the context of the so-called γ-iteration, in the newly developed approach suggested in [5], generalized Schur forms have to be computed for matrix pencils of the form

$$\hat{H}_c(t) - \lambda \hat{N}_c := \begin{bmatrix} 0 & A & B \\ A^H & 0 & S \\ B^H & S^H & R(t) \end{bmatrix} - \lambda \begin{bmatrix} 0 & E & 0 \\ -E^H & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{1.6}$$

with an indefinite Hermitian matrix

$$R(t) = \begin{bmatrix} R_{11} - tI & R_{12} \\ R_{12}^H & R_{22} \end{bmatrix}$$

which varies with the positive parameter t (playing the role of the parameter γ in the γ-iteration), while the other coefficients are constant in t. Here, besides the classical questions of perturbation theory as above, we are interested in the eigenvalues and deflating subspaces as functions of t and we want to study the size of perturbations that is needed to bring any of the finite eigenvalues to the imaginary axis.

Again there is a discrete-time H_∞ analogue to this case [10] which leads to matrix pencils

$$\hat{H}_d(t) - \lambda \hat{N}_d = \begin{bmatrix} 0 & A & B \\ -E^H & 0 & S \\ 0 & S^H & R(t) \end{bmatrix} - \lambda \begin{bmatrix} 0 & E & 0 \\ -A^H & 0 & 0 \\ -B^H & 0 & 0 \end{bmatrix}. \tag{1.7}$$

Here again we are interested in the eigenvalues and deflating subspaces as functions of t and we want to study the size of perturbations that is needed to bring any of the finite eigenvalues to the unit circle.

The paper is organized as follows. First we introduce the notation and give some preliminary results in section 2. In section 3 we formulate a framework for analyzing the effect of linear perturbations on general matrix pencils. We then study the special cases of perturbations for general skew-symmetric/symmetric pencils arising from continuous-time problems in section 4 and the corresponding discrete-time problems in section 5.

2. Notation and preliminaries. We denote the set of all complex (real) matrices of size n by $\mathbb{C}^{n,n}$ ($\mathbb{R}^{n,n}$). Given a matrix A, we denote its complex conjugate by $\overline{A}$, its transpose by A^T and the transpose of its complex conjugate by A^H. We denote the identity matrix of size n by I_n. Also we consider the "flip" permutation matrix

$$F_n := \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix} \in \mathbb{C}^{n,n}.$$

We denote the spectrum of a square matrix A and a pencil (A, B) by $\sigma(A)$ and $\sigma(A, B)$, respectively. Given a set S we denote its boundary by ∂S. For $A \in \mathbb{C}^{n,n}$, we define the [illegible] A as $r(A) := \max\{|\lambda| : \lambda \in \sigma(A)\}$.

Also, given $z \in \mathbb{C}$, we define

$$\operatorname{sep}_{\mathbb{R}}(z, \Delta H, \Delta N) := \min\{|t| : z \in \sigma(H + t\Delta H, N + t\Delta N), t \in \mathbb{R}\}.$$

It is well known (see, e.g., [25]) that for every matrix $A \in \mathbb{C}^{n,n}(\mathbb{R}^{n,n})$ there exist symmetric matrices $T = T^T$ and $S = S^T$, where S is also nonsingular, such that $A = TS^{-1}$. Note that if A is not real then these factors in general are complex symmetric but not Hermitian. Furthermore, if $A \in \mathbb{R}^{n,n}$, then T and S can be chosen to be real matrices. Since this result is due to Frobenius, we refer to T and S as [illegible] of A. In our work, we will need similar factorizations, however with Frobenius factors that are Hermitian. It is easy to see that if $A \in \mathbb{R}^{n,n}$, then A always has Hermitian Frobenius factors. However, if $A \in \mathbb{C}^{n,n}$, then Hermitian factors need not exist. This follows by observing the fact that if $A = TS^{-1}$ with $T^H = T$ and $S^H = S$, then we must have $AS = SA^H$, that is, $A = SA^HS^{-1}$. This implies that the matrices A and A^H must be similar and hence they must have the same eigenvalues. Thus, a necessary condition for the existence of Hermitian Frobenius factors T and S is that $\sigma(A) = \sigma(A^H)$.

We show that $\sigma(A) = \sigma(A^H)$ is also a sufficient condition for A to have Hermitian Frobenius factors. For this, we first observe that $\sigma(A) = \sigma(A^H)$ implies that for every nonreal eigenvalue of A its complex conjugate is also an eigenvalue with the same multiplicity.

PROPOSITION 2.1. [illegible] $A \in \mathbb{C}^{n,n}$ [illegible] $\sigma(A) = \sigma(A^H)$ [illegible] $\eta_1, \eta_2, \ldots, \eta_r$ [illegible] A [illegible] $\lambda_1, \lambda_2, \ldots, \lambda_p, \overline{\lambda_1}, \overline{\lambda_2}, \ldots, \overline{\lambda_p}$ [illegible] A [illegible] $m_1, m_2, \ldots, m_p$ [illegible] $\lambda_1, \lambda_2, \ldots, \lambda_p$ [illegible] $\sum_{i=1}^{p} m_i = m$ [illegible] $k_1, k_2, \ldots, k_r$ [illegible] $\eta_1, \eta_2, \ldots, \eta_r$ [illegible] $\sum_{j=1}^{r} k_j = k$ [illegible] $n = 2m + k$ [illegible]

$$U = \operatorname{diag}\left(F_{2m_1}, F_{2m_2}, \ldots, F_{2m_p}, F_{n_1}, F_{n_2}, \ldots, F_{n_s}\right),$$

[illegible] A [illegible] A^H [illegible]

$$A = PJP^{-1}, \qquad A^H = (P^{-H}U)\overline{J}(P^{-H}U)^{-1},$$

[illegible]

$$J := \operatorname{diag}\left(J_{m_1}(\lambda_1), J_{m_1}(\overline{\lambda_1}), \ldots, J_{m_p}(\lambda_p), J_{m_p}(\overline{\lambda_p}),\right.$$
$$\left. J_{n_1}(\eta_1), J_{n_2}(\eta_2), \ldots, J_{n_r}(\eta_r)\right) \tag{2.1}$$

[illegible] $i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, r$,

$$J_{m_i}(\lambda_i) := \begin{bmatrix} \lambda_i & \varphi & 0 & \cdots & 0 & 0 \\ 0 & \lambda_i & \varphi & \cdots & 0 & 0 \\ & & & & & \\ 0 & 0 & 0 & \cdots & \lambda_i & \varphi \\ 0 & 0 & 0 & \cdots & 0 & \lambda_i \end{bmatrix}$$

$$J_{n_j}(\eta_j) := \begin{bmatrix} \eta_j & \varphi & 0 & \cdots & 0 & 0 \\ 0 & \eta_j & \varphi & \cdots & 0 & 0 \\ & & & & & \\ 0 & 0 & 0 & \cdots & \eta_j & \varphi \\ 0 & 0 & 0 & \cdots & 0 & \eta_j \end{bmatrix}$$

$\varphi = 0$ $\varphi = 1$ U $\overline{J}U - UJ^H = 0$

Let $A = PJP^{-1}$ with J as in (2.1) be a Jordan decomposition of A. Then, $A^H = P^{-H}J^HP^H$ with

$$J^H = (\overline{J})^T = \operatorname{diag}\left(J_{m_1}(\overline{\lambda_1}), J_{m_1}(\lambda_1), \ldots, J_{m_p}(\overline{\lambda_p}), J_{m_p}(\lambda_p), J_{n_1}(\eta_1), J_{n_2}(\eta_2), \ldots, J_{n_r}(\eta_r)\right)^T,$$

and for $i = 1, 2, \ldots, p$,

$$\begin{aligned} \begin{bmatrix} J_{m_i}(\overline{\lambda_i}) & 0 \\ 0 & J_{m_i}(\lambda_i) \end{bmatrix} F_{2m_i} &= \begin{bmatrix} \overline{\lambda_i} & \varphi & & & \\ & \overline{\lambda_i} & \varphi & & \\ & & \ddots & \ddots & \\ & & & \lambda_i & \varphi \\ & & & & \lambda_i \end{bmatrix} \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & \cdots & 0 & \varphi & \overline{\lambda_i} \\ 0 & 0 & \cdots & \varphi & \overline{\lambda_i} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \varphi & \lambda_i & \cdots & 0 & 0 & 0 \\ \overline{\lambda_i} & 0 & \cdots & 0 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix} \begin{bmatrix} \overline{\lambda_i} & & & & \\ \varphi & \overline{\lambda_i} & & & \\ & \ddots & \ddots & & \\ & & \varphi & \lambda_i & \\ & & & \varphi & \lambda_i \end{bmatrix} \\ &= F_{2m_i} \begin{bmatrix} J_{m_i}(\overline{\lambda_i}) & 0 \\ 0 & J_{m_i}(\lambda_i) \end{bmatrix}^T. \end{aligned}$$

Similarly for $j = 1, 2, \ldots, s$, we have $J_{n_j}F_{n_j} = F_{n_j}J_{n_j}^T$ and thus it follows that $\overline{J}U - UJ^H = 0$. □

Using Proposition 2.1 together with Theorem 12.5.1 of [16] we then construct a nonsingular Hermitian solution S of the equation $AS - SA^H = 0$.

THEOREM 2.2. $A \in \mathbb{C}^{n,n}$ $\sigma(A) = \sigma(A^H)$ S $AS = SA^H$

Using the notation of Proposition 2.1, it follows by Theorem 12.5.1 in [16] that all solutions of the equation $AS - SA^H = 0$ are of the form $S = PY(P^{-H}U)^{-1} = PYUP^H$. Here Y satisfies $JY - Y\overline{J} = 0$ and has the block form $Y = \operatorname{diag}(Y_1, Y_2, \ldots, Y_p, I_k)$, where $Y_i = \left[\begin{smallmatrix} 0 & Y_{i,i} \\ Y_{i,i} & 0 \end{smallmatrix}\right] \in \mathbb{C}^{2m_i, 2m_i}$, each $Y_{i,i}$ being an arbitrary upper triangular Toeplitz matrix for $i = 1, 2, \ldots, p$. Since we want YU to

be Hermitian, a possible choice is $Y_{i,i} = I_{m_i}$. Then for $i = 1, 2, \ldots, p$, we have

$$Y_i F_{2m_i} = \operatorname{diag}(F_{m_i}, F_{m_i}) = F_{2m_i} Y_i,$$

which implies that $(YU)^H = U^H Y^H = UY = YU$. Thus, $S := PYUP^H$ is nonsingular and Hermitian. ∎

Theorem 2.2 immediately provides a necessary and sufficient condition for $A \in \mathbb{C}^{n,n}$ to have Hermitian Frobenius factors.

COROLLARY 2.3. *$A \in \mathbb{C}^{n,n}$ [illegible] T S [illegible] S [illegible] $A = TS^{-1}$ [illegible] $\sigma(A) = \sigma(A^H)$.*

Proof. Suppose that there exist Hermitian matrices T and S, where S is also nonsingular such that $A = TS^{-1}$. Then $T = AS$ and $T = T^H$ and hence $AS = SA^H$ or $A = SA^H S^{-1}$, i.e., $\sigma(A) = \sigma(A^H)$. For the converse, suppose that $\sigma(A) = \sigma(A^H)$. Then by Theorem 2.2, there exists a nonsingular Hermitian matrix S such that $AS = SA^H$. This implies that $A = (SA^H)S^{-1}$ and the proof follows by setting $T = SA^H$. ∎

Since the factorization $A = TS^{-1}$ with $S^H = S$ and $T^H = T$, if it exists, depends on the choice of S as a solution of $AX - XA^H = 0$, it is evident from Theorem 2.2 that this factorization is, in general, not unique.

In the following we will need Frobenius factorizations for matrices that depend on a complex parameter. Suppose that A depends smoothly upon a complex parameter z and $\sigma(A(z)) = \sigma(A(z)^H)$, and let $A(z) = T(z)S(z)^{-1}$ be a Frobenius factorization of $A(z)$ with $T(z) = T(z)^H$ and $S(z) = S(z)^H$. Using the spectral factorization of $T(z)$, there exists a unitary matrix $U(z)$ such that

$$T(z) = U(z) \begin{bmatrix} D_+(z) & & \\ & D_-(z) & \\ & & 0 \end{bmatrix} U(z)^H,$$

where $D_+(z) \in \mathbb{C}^{\pi,\pi}, -D_-(z) \in \mathbb{C}^{\nu,\nu}$ are diagonal matrices with positive diagonal elements and $\pi(z) \geq \nu(z)$, where $(\pi(z), \nu(z), \omega(z))$ with $\pi(z) + \nu(z) + \omega(z) = n$ is the [illegible] $T(z)$; see [16].

Setting

$$Q(z) := U(z) \begin{bmatrix} (D_+(z))^{\frac{1}{2}} & & \\ & (-D_-(z))^{\frac{1}{2}} & \\ & & I_\omega(z) \end{bmatrix}, \qquad \tilde{I}_T(z) := \begin{bmatrix} I_\pi(z) & & \\ & -I_\nu(z) & \\ & & 0 \end{bmatrix},$$

we have for given z a factorization

$$A(z) = Q(z)\tilde{I}_T(z)Q^H(z)S(z)^{-1}. \tag{2.2}$$

Note that the choice $\pi \geq \nu$ makes the matrix $\tilde{I}_T(z)$ unique, while there is still much freedom in the choice of the transformation matrix $Q(z)$. In an analogous way we can construct a factorization

$$A(z) = T(z)(V(z)(\tilde{I}_S(z))V(z)^H)^{-1} = T(z)V(z)^{-H}(\tilde{I}_S(z))V(z)^{-1} \tag{2.3}$$

by using the spectral factorization and the inertia index of $S(z)$.

An interesting open question that one may discuss in this context is how to obtain a smooth Frobenius factorization when the matrix depends smoothly on a parameter, as in our case.

3. Linear perturbation of general matrix pencils. In this section we consider the effect of perturbing a regular square matrix pencil (H, N) where $H, N \in \mathbb{C}^{n,n}$ $(\mathbb{R}^{n,n})$ by linear perturbations $(H+t\Delta H, N+t\Delta N)$. Here $\Delta H, \Delta N \in \mathbb{C}^{n,n}(\mathbb{R}^{n,n})$ are fixed perturbation matrices and the parameter t varies over the real numbers.

LEMMA 3.1. [illegible] $z \in \mathbb{C}$ [illegible] $\text{sep}_{\mathbb{R}}(z, \Delta H, \Delta N) < \infty$ [illegible] $(\Delta H - z\Delta N)(H - zN)^{-1}$ [illegible] $\text{sep}_{\mathbb{R}}(z, \Delta H, \Delta N) < \infty$ [illegible]

$$\text{sep}_{\mathbb{R}}(z, \Delta H, \Delta N) = \left[\max_{\lambda \in \mathbb{R}} \{\lambda \in \sigma((\Delta H - z\Delta N)(H - zN)^{-1})\}\right]^{-1}.$$

[illegible] The proof follows immediately from the fact that, for $\lambda \notin \sigma(H, N)$, we have

$$H + t\Delta H - \lambda(N + t\Delta N) = \left[I + t(\Delta H - \lambda\Delta N)(H - \lambda N)^{-1}\right](H - \lambda N).$$

Therefore, $\lambda \in \sigma(H + t\Delta H, N + t\Delta N)$ if and only if $-1/t \in \sigma((\Delta H - \lambda\Delta N)(H - \lambda N)^{-1})$. □

As discussed in the introduction, we are interested in conditions which guarantee that all the eigenvalues of the perturbed pencils $(H + t\Delta H, N + t\Delta N)$, $t \in \mathbb{R}$, remain within a particular open subset, say, $\mathbb{C}_g$, of the complex plane. Since the eigenvalues of $(H + t\Delta H, N + t\Delta N)$ move continuously as t varies in $\mathbb{R}$, the smallest value of $|t|$ for which these eigenvalues move out of $\mathbb{C}_g$ is evidently equal to

$$r(\mathbb{C}_g, \Delta H, \Delta N) := \inf_{z \in \delta\mathbb{C}_g} \text{sep}(z, \Delta H, \Delta N).$$

Note that similar distances are very important in other contexts of control theory, where the smallest perturbation that makes a system unstable is called the stability radius [11] and the smallest perturbation that makes a system nonpassive is called the passivity radius [19].

Since a complex number z becomes an eigenvalue of the perturbed pencil $(H + t\Delta H, N + t\Delta N)$ for some $t \in \mathbb{R}$ if and only if the matrix $(\Delta H - z\Delta N)(H - zN)^{-1}$ has a nonzero real eigenvalue, it is possible that there exist pencils (H, N) with corresponding perturbations $(\Delta H, \Delta N)$ and sets $\mathbb{C}_g$ such that $r(\mathbb{C}_g, \Delta H, \Delta N) = \infty$, that is, the eigenvalues of the perturbed pencils $(H + t\Delta H, N + t\Delta N)$ always remain inside $\mathbb{C}_g$ as t varies over the real numbers. In such cases, $\text{sep}(z, \Delta H, \Delta N) = \infty$ for all $z \in \partial\mathbb{C}_g$. This is illustrated by the following example.

EXAMPLE 3.2. Consider the pencil (H, N), where

$$H := \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}, \text{ and } N := \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Its eigenvalues are $\sqrt{3}$ and $-\sqrt{3}$. Let

$$\Delta H := \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \qquad \Delta N := 0$$

be the perturbations to H and N, respectively.

Let $\mathbb{C}_g := \mathbb{C} \setminus \{z \in \mathbb{C} : Re(z) = 0\}$. Then the boundary $\partial\mathbb{C}_g$ of $\mathbb{C}_g$ is evidently the imaginary axis. Therefore, by Lemma 3.1 for all $t \in \mathbb{R}$, the eigenvalues of the pencils $(H + t\Delta H, N + t\Delta N)$ are always in $\mathbb{C}_g$, if and only if the matrix

$$(\Delta H - z\Delta N)(H - zN)^{-1} = \frac{1}{3 - z^2} \begin{bmatrix} -1 & 2 - z \\ -(2 + z) & 1 \end{bmatrix}$$

has no nonzero real eigenvalue for every $z \in \mathbb{C}$ lying on the imaginary axis. The eigenvalues of $(\Delta H - z\Delta N)(H - zN)^{-1}$ are $i/\sqrt{(3 - z^2)}$ and $-i/\sqrt{(3 - z^2)}$. Now for every z lying on the imaginary axis, there exists a real number γ, such that $z = i\gamma$. Therefore, for $z \in \partial\mathbb{C}_g$, the eigenvalues of $(\Delta H - z\Delta N)(H - zN)^{-1}$ are $i/\sqrt{(3 + \gamma^2)}$ and $-i/\sqrt{(3 + \gamma^2)}$. This shows that $(\Delta H - z\Delta N)(H - zN)^{-1}$ has no real eigenvalues for all $z \in \partial\mathbb{C}_g$. Hence, $\sigma(H + t\Delta H, N + t\Delta N) \subset \mathbb{C}_g$ for all $t \in \mathbb{R}$.

Since $z \in \sigma(H + t\Delta H, N + t\Delta N)$ for some $t \in \mathbb{R}$ if and only if the matrix $F(z) := (\Delta H - z\Delta N)(H - zN)^{-1}$ has a real eigenvalue, we identify conditions under which the latter matrix has a real eigenvalue. Under the assumption that $\sigma(F(z)) = \sigma(F(z)^H)$, let $F(z) = T(z)\{S(z)\}^{-1}$ be a Frobenius factorization of $F(z)$ where $T(z)^H = T(z)$ and $S(z)^H = S(z)$. The following result gives a necessary and sufficient condition for this matrix to have a real eigenvalue.

LEMMA 3.3. [illegible] $z \in \mathbb{C}$ [illegible] $T(z)$ [illegible] $S(z)$ [illegible] $F(z) := (\Delta H - z\Delta N)(H - zN)^{-1}$ [illegible] $T(z)^H = T(z)$ $S(z)^H = S(z)$ [illegible] $S(z)$ [illegible] $(\pi_T(z), \nu_T(z), \omega_T(z))$ [illegible] $\pi_T(z) \geq \nu_T(z)$ [illegible] $T(z)$ [illegible] $\tilde{I}_T(z)$ $Q(z)$ [illegible] (2.2) [illegible] $F(z)$ [illegible] $(\Delta H - z\Delta N)(H - zN)^{-1}$ [illegible] $\tilde{I}_T(z)Q(z)\{S(z)\}^{-1}Q(z)^H$ [illegible] The proof follows, since $F(z)$ and $\tilde{I}_T(z)Q(z)\{S(z)\}^{-1}Q(z)^H$ are similar. □

It is evident that the roles of the matrices $T(z)$ and $S(z)$ in Lemma 3.3 can be interchanged.

LEMMA 3.4. [illegible] $z \in \mathbb{C}$ [illegible] $T(z)$ [illegible] $S(z)$ [illegible] $F(z) := (\Delta H - z\Delta N)(H - zN)^{-1}$ [illegible] $T(z)^H = T(z)$, $S(z)^H = S(z)$ [illegible] $S(z)$ [illegible] $(\pi_S(z), \nu_S(z), \omega_S(z))$ [illegible] $\pi_S(z) \geq \nu_S(z)$ [illegible] $S(z)$ [illegible] $\tilde{I}_S(z)$ $V(z)$ [illegible] (2.3) [illegible] $F(z)$ [illegible] $(\Delta H - z\Delta N)(H - zN)^{-1}$ [illegible] $V(z)^{-1}T(z)V(z)^{-H}\tilde{I}_S(z)$ [illegible]

In general, the function $\text{sep}_{\mathbb{R}}(z, \Delta H, \Delta N)$ is discontinuous as a function of z, since it depends on the matrix $(\Delta H - z\Delta N)(H - zN)^{-1}$ having a real eigenvalue. However, it is possible that given a set $\mathbb{C}_g \subset \mathbb{C}$, the structure of the matrices $H, N, \Delta H$, and ΔN are such that the matrix $(\Delta H - z\Delta N)(H - zN)^{-1}$ always has one or more real eigenvalues for $z \in \partial\mathbb{C}_g$. Let these eigenvalues be $h_1(z), \ldots, h_p(z)$. Then $(\Delta H - z\Delta N)(H - zN)^{-1}$ is an analytic function of $z \in \mathbb{C} \setminus \sigma(H, N)$, and hence in particular of $z \in \partial\mathbb{C}_g$ (Theorem 1.5, pp. 66, [13]) and the eigenvalues $h_1(z), \ldots, h_p(z)$ are continuous (Corollary 3, pp. 105, [3]). Therefore, for such cases we have for $z \in \partial\mathbb{C}_g$,

$$\text{sep}_{\mathbb{R}}(z, \Delta H, \Delta N) = \{\max\{|h_k(z)| : k = 1, \ldots, p\}\}^{-1},$$

which implies that $\text{sep}_{\mathbb{R}}(z, \Delta H, \Delta N)$ is a continuous function of z. In the special situation that all the eigenvalues of $(\Delta H - z\Delta N)(H - zN)^{-1}$ are real, we have $\text{sep}_{\mathbb{R}}(z, \Delta H, \Delta N) = \{r((\Delta H - z\Delta N)(H - zN)^{-1})\}^{-1}$ for all $z \in \partial\mathbb{C}_g$. In such cases, the distribution of the eigenvalues of $(H + t\Delta H, N + t\Delta N)$ on $\partial\mathbb{C}_g$ may be analyzed by plotting the level curves of the spectral radius function $r((\Delta H - z\Delta N)(H - zN)^{-1})$ in neighborhoods of $\partial\mathbb{C}_g$. Then the smallest value of $|t|$ for which some $z \in \partial\mathbb{C}_g$ is an eigenvalue of $(H + t\Delta H, N + t\Delta N)$ is evidently given by the smallest value of ϵ for which the level set

$$L(\epsilon, \Delta H, \Delta N) := \{z \in \mathbb{C} \setminus \sigma(H, N) : r((\Delta H - z\Delta N)(H - zN)^{-1}) = \epsilon^{-1}\}$$

intersects $\partial\mathbb{C}_g$. In other words, for such problems, the distance to the boundary of

$\mathbb{C}_g$ is given by

$$r(\mathbb{C}_g, \Delta H, \Delta N) := \min\{\epsilon \in \mathbb{R} : L(\epsilon, \Delta H, \Delta N) \cap \partial\mathbb{C}_g \neq 0\}. \tag{3.1}$$

By Proposition 2.1 of [1] the spectral radius function $r((\Delta H - z\Delta N)(H - zN)^{-1})$ is nonconstant on open subsets of $\mathbb{C} \setminus \sigma(H, N)$. This together with the fact that it is also continuous on $\mathbb{C} \setminus \sigma(H, N)$ implies that the level sets $L(\epsilon, \Delta H, \Delta N)$ are closed sets which have no interior points. In other words, they are curves on the complex plane. Furthermore, the curve $L(\epsilon, \Delta H, \Delta N)$ intersects $\partial\mathbb{C}_g$ at only a finite number of points, since at each such point we must either have $z \in \sigma(H + \epsilon\Delta H, N + \epsilon\Delta N)$ or $z \in \sigma(H - \epsilon\Delta H, N - \epsilon\Delta N)$. This justifies the use of "minimum" instead of "infimum" in (3.1).

In the following theorem, we give sufficient conditions for all the eigenvalues of the matrix $(\Delta H - z\Delta N)(H - zN)^{-1}$ to be real.

THEOREM 3.5. [illegible] $F(z) = (\Delta H - z\Delta N)(H - zN)^{-1}$, $z \in \mathbb{C}$ [illegible] $F(z) = T(z)S(z)^{-1}$ [illegible] $T(z)^H = T(z)$ [illegible] $S(z)^H = S(z)$ [illegible] $T(z)$ [illegible] $S(z)$ [illegible]

(i) $T(z)$ [illegible] $S(z)^{-1}$ [illegible]
(ii) $T(z)$ [illegible]
(iii) $S(z)$ [illegible]

[illegible]

(i) Since $(\Delta H - z\Delta N)(H - zN)^{-1} = T(z)S(z)^{-1}$, where $T(z)$ and $S(z)$ are Hermitian, the matrix $(\Delta H - z\Delta N)(H - zN)^{-1}$ is Hermitian if $T(z)S(z)^{-1} = S(z)^{-1}T(z)$ and therefore all its eigenvalues are real. This proves (i).

(ii) If $T(z)$ and $S(z)$ do not commute but $T(z)$ is positive semidefinite with $\pi(z)$ nonzero eigenvalues, then we obtain the Frobenius factorization (2.2) as $F(z) = Q(z)\tilde{I}_T(z)Q(z)^H S(z)^{-1}$ with $\tilde{I}_T(z) = \left[\begin{smallmatrix} I_\pi & 0 \\ 0 & 0 \end{smallmatrix}\right]$. If we partition

$$Q(z)^H S(z)^{-1} Q(z) = \begin{bmatrix} S_{11}(z) & S_{12}(z) \\ S_{21}(z) & S_{22}(z) \end{bmatrix}$$

conformally with $\tilde{I}_T(z)$, then $S_{11}(z)$ is Hermitian and

$$Q(z)^H T(z)\{S(z)\}^{-1} Q(z) = \begin{bmatrix} S_{11}(z) & S_{12}(z) \\ 0 & 0 \end{bmatrix}.$$

Therefore, $\sigma((\Delta H - z\Delta N)(H - zN)^{-1}) = \sigma(S_{11}(z)) \cup \{0\}$, which is real.

(iii) The proof follows as in (ii) by exchanging the roles of S and T and using the factorization (2.3). ∎

Note that Theorem 3.5 also holds if the condition of positive semidefiniteness in (ii) and positive definiteness in (iii) are replaced by negative semidefiniteness and negative definiteness, respectively.

4. Linear perturbation of structured matrix pencils arising from continuous-time control problems. In this section we apply the results from section 3 to the specific pencils from control theory that we introduced in section 1. The matrices H and N then have special structure and, in order not to destroy the properties of the pencils, it should be guaranteed that the perturbations preserve this structure.

This means that we study the effect of perturbations $(H + t\Delta H, N + t\Delta N)$, $t \in \mathbb{R}$, where the matrices ΔH and ΔN have the same structure as H and N, respectively. Although we consider complex pencils, the results of this section also hold for real pencils.

4.1. Perturbation of pencils arising in continuous-time control. The first application that we discuss are matrix pencils of the form (1.4), where we perturb only the blocks E, A, B, Q, S, R but such that Q and R stay Hermitian, i.e., we consider the case

$$H = \begin{bmatrix} 0 & A & B \\ A^H & Q & S \\ B^H & S^H & R \end{bmatrix}, \qquad N = \begin{bmatrix} 0 & E & 0 \\ -E^H & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{4.1}$$

with $A, Q, E \in \mathbb{C}^{n,n}$, $B, S^H \in \mathbb{C}^{n,m}$, $R \in \mathbb{C}^{m,m}$, $Q = Q^H$, $R = R^H$ and we assume that E is invertible. The perturbation matrices are

$$\Delta H = \begin{bmatrix} 0 & \Delta A & \Delta B \\ (\Delta A)^H & \Delta Q & \Delta S \\ (\Delta B)^H & (\Delta S)^H & \Delta R \end{bmatrix}, \qquad \Delta N := \begin{bmatrix} 0 & \Delta E & 0 \\ -(\Delta E)^H & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{4.2}$$

where the dimensions are analogous and where we assume that $(\Delta Q)^H = \Delta Q$, $(\Delta R)^H = \Delta R$ and that $E + \Delta E$ is still invertible. The pencils (H, N) and $(H + \Delta H, N + \Delta N)$ are then both Hermitian/skew-Hermitian pencils and we are interested in the set $\mathbb{C}_g := \mathbb{C} \setminus \{z \in \mathbb{C} : Re(z) = 0\}$. Hence, the quantity of interest is the smallest value $|t|$, $t \in \mathbb{R}$, such that $(H + t\Delta H, N + t\Delta N)$ has a purely imaginary eigenvalue. In view of Lemma 3.1 (with $z = i\gamma$, $\gamma \in \mathbb{R}$) this is equivalent to finding the smallest $|\gamma|$ such that the matrix $(\Delta H - i\gamma\Delta N)(H - i\gamma N)^{-1}$ has a nonzero real eigenvalue.

Evidently, we have the following expressions for Hermitian Frobenius factors $T(i\gamma)$ and $S(i\gamma)$ of the matrix $(\Delta H - i\gamma\Delta N)(H - i\gamma N)^{-1}$.

$$T(i\gamma) = \begin{bmatrix} 0 & \Delta A - i\gamma\Delta E & \Delta B \\ (\Delta A - i\gamma\Delta E)^H & \Delta Q & \Delta S \\ (\Delta B)^H & (\Delta S)^H & \Delta R \end{bmatrix},$$

$$S(i\gamma) = \begin{bmatrix} 0 & A - i\gamma E & B \\ (A - i\gamma E)^H & Q & S \\ B^H & S^H & R \end{bmatrix}.$$

We may directly use Lemmas 3.3 and 3.4 to obtain conditions for $(H+t\Delta H, N+t\Delta N)$ to have a purely imaginary eigenvalue, as t varies in $\mathbb{R}$. But the special structure of the Frobenius factors provides another condition that is more specific to the problem at hand. To obtain it, we assume without loss of generality that the matrix $[A\ B]$ is not a square matrix, i.e., that the matrix B has at least one column.

THEOREM 4.1. [illegible] (H, N) [illegible] ΔH [illegible] ΔN [illegible] (4.1) [illegible] (4.2) [illegible]

$$P(t, \gamma) := [A - i\gamma E + t(\Delta A - i\gamma\Delta E)\ \ B + t\Delta B],$$
$$Z(t) := \begin{bmatrix} Q + t\Delta Q & S + t\Delta S \\ (S + t\Delta S)^H & R + t\Delta R \end{bmatrix}.$$

[illegible] $V(t, \gamma)$ [illegible] $P(t, \gamma)$ [illegible] 0 [illegible] $W(t, \gamma)$ [illegible] $P(t, \gamma)^H$
[illegible] $t \neq 0$ [illegible] γ [illegible] $i\gamma$ [illegible] $(H + t\Delta H, N + t\Delta N)$ [illegible]

$$Z(t)(V(t, \gamma)) \cap W(t, \gamma) \neq \emptyset.$$

. We make use of the fact $(H + t\Delta H, N + t\Delta N)$, $t \in \mathbb{R}$ has a purely imaginary eigenvalue $i\gamma$, $\gamma \in \mathbb{R}$ if and only if $-1/t$ is an eigenvalue of the matrix $(\Delta H - i\gamma\Delta N)(H - i\gamma N)^{-1}$. Considering a Frobenius factorization

$$(\Delta H - i\gamma\Delta N)(H - i\gamma N)^{-1} = T(i\gamma)(S(i\gamma))^{-1},$$

it follows that $i\gamma$ is an eigenvalue of $(H + t\Delta H, N + t\Delta N)$ if and only $-1/t$ is an eigenvalue of $T(i\gamma)S(i\gamma)^{-1}$, i.e., if and only if there exists a vector $x \neq 0$ such that $T(i\gamma)S(i\gamma)^{-1}x = -\frac{1}{t}x$. Setting $y := S(i\gamma)^{-1}x$, this, in turn, implies that $i\gamma$ is an eigenvalue of $(H + t\Delta H, N + t\Delta N)$, if and only if there exists a vector $y \neq 0$ such that $S(i\gamma)y = -tT(i\gamma)y$. Writing down the expressions for $T(i\gamma)$ and $S(i\gamma)$, we have

$$\left[\begin{array}{c|cc} 0 & A - i\gamma E + t(\Delta A - i\gamma\Delta E) & B + t\Delta B \\ \hline (A - i\gamma E + t(\Delta A - i\gamma\Delta E)^H & Q + t\Delta Q & S + t\Delta S \\ (B + t\Delta B)^H & (S + t\Delta S)^H & R + t\Delta R \end{array}\right] y = 0.$$

This in turn can be written as

$$\left[\begin{array}{cc} 0 & P(t,\gamma) \\ P(t,\gamma)^H & Z(t) \end{array}\right] \left[\begin{array}{c} y_1 \\ y_2 \end{array}\right] = 0.$$

Hence, we have the following system of equations:

$$\begin{aligned} P(t,\gamma)y_2 &= 0 \\ P(t,\gamma)^H y_1 + Z(t)y_2 &= 0. \end{aligned}$$

From the first equation we have that either $y_2 = 0$ or 0 is a singular value of $P(t,\gamma)$ and y_2 a corresponding singular vector. But as $[A\ B]$ is not a square matrix, neither is $P(t,\gamma) = [A - i\gamma E + t(\Delta A - i\gamma\Delta E)\ \ B + t\Delta B]$. As a consequence a nonzero vector y_2 satisfying the first equation always exists. Therefore, a necessary and sufficient condition for $i\gamma$ to be an eigenvalue of $(H + t\Delta H, N + t\Delta N)$ is that, for every right singular vector y_2 of $P(t,\gamma)$ corresponding to the singular value 0, there exists some vector y_1 such that $-P(t,\gamma)^H y_1 = Z(t)y_2$. This implies that the matrix $Z(t)$ maps at least one right singular vector of $P(t,\gamma)$ corresponding to the singular value 0, to the range of $P(t,\gamma)^H$. Since $V(t,\gamma)$ is the set of all these right singular vectors of $P(t,\gamma)$, it follows that $i\gamma$ is an eigenvalue of $(H + t\Delta H, N + t\Delta N)$ if and only if $Z(t)(V(t,\gamma)) \cap W(t,\gamma) \neq \emptyset$. □

In the applications from control theory, the matrices Q, R and S are associated with the cost function and often these cost functions can be chosen. If this is the case, then we may assume that the corresponding perturbations $\Delta Q, \Delta R$ and ΔS are all equal to zero. Under this assumption, we have the following immediate corollary of Theorem 4.1.

COROLLARY 4.2. $W(t,\gamma)$ $V(t,\gamma)$ 4.1 $Z_0 := \left[\begin{smallmatrix} Q & S \\ S^H & R \end{smallmatrix}\right]$ $(H + t\Delta H, N + t\Delta N)$ $i\gamma$ $Z_0(V(t,\gamma)) \cap W(t,\gamma) \neq \emptyset$

. The proof follows immediately from Theorem 4.1 by noticing the fact that $Z(0) = Z_0$. □

Corollary 4.2 implies that for a given fixed real number t, the matrices Q, S and R of the cost functional can be chosen in such a way that the pencil $(H+t\Delta H, N+t\Delta N)$ does not have any purely imaginary eigenvalues, cp. [15].

COROLLARY 4.3. $P(t,\gamma)$ $V(t,\gamma)$ 4.1 $Z_0 := \left[\begin{smallmatrix} Q & S \\ S^H & R \end{smallmatrix}\right]$ $t \in \mathbb{R}$ $(H+t\Delta H, N+t\Delta N)$

$$Z_0(\cup_{\gamma\in\mathbb{R}} V(t,\gamma)) \cap (\cup_{\gamma\in\mathbb{R}} W(t,\gamma)) = \emptyset.$$

The condition of Corollary 4.3 is necessary and sufficient for a pencil $(\tilde{H}, \tilde{N}) := (H+t\Delta H, N+t\Delta N)$ to have no imaginary eigenvalues. Thus, this condition generalizes well-known classical conditions that guarantee that the considered pencil has no purely imaginary eigenvalue, see, e.g., [15, 17].

For instance, it is well known that a matrix pencil (H, N), with H and N as in (4.1), has no purely imaginary eigenvalues if its blocks satisfy the following conditions:

(i) The matrix R is positive definite and the matrix $Q - SR^{-1}S^H$ is positive semidefinite.
(ii) The triple (E, A, B) where A has size n, is , i.e., for all complex numbers λ in the closed right half plane the rank of $[A - \lambda E, B]$ is n.
(iii) If $Q - SR^{-1}S^H = C^H C$ is a full rank factorization of $Q - SR^{-1}S^H$, then (E, A, C) is , i.e., (E^H, A^H, C^H) is stabilizable.

The following example shows that there exist pencils (H, N) which arise from systems that are not stabilizable and detectable and yet they do not have any purely imaginary eigenvalues. This is due to the fact that they satisfy the condition given in Corollary 4.3.

EXAMPLE 4.4. Let

$$H := \left[\begin{array}{cc|ccc} 0 & 0 & 2 & 3 & 2 \\ 0 & 0 & 0 & 5 & 2 \\ \hline 2 & 0 & 1 & -1 & 1 \\ 3 & 5 & -1 & 1 & -1 \\ \hline 2 & 2 & 1 & -1 & 5 \end{array}\right], \qquad N := \left[\begin{array}{cc|ccc} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ \hline -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

The eigenvalues of the pencil (H, N) are $2, -2, 5, -3$ and ∞. For this pencil we have,

$$A := \begin{bmatrix} 2 & 3 \\ 0 & 5 \end{bmatrix}, \qquad E := \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad B := \begin{bmatrix} 2 \\ 2 \end{bmatrix},$$
$$Q := \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \qquad R := [5], \qquad S := \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

It is easy to see that (E, A, B) is not stabilizable, as the matrix $[A-2I, B] = \left[\begin{smallmatrix} 0 & 3 & 2 \\ 0 & 3 & 2 \end{smallmatrix}\right]$, evidently has rank 1. We also note that

$$\begin{bmatrix} -\frac{2}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{bmatrix}^H \begin{bmatrix} -\frac{2}{\sqrt{5}} & -\frac{2}{\sqrt{5}} \end{bmatrix} = \begin{bmatrix} \frac{4}{5} & -\frac{4}{5} \\ -\frac{4}{5} & \frac{4}{5} \end{bmatrix} = Q - SR^{-1}S^H.$$

Setting $C := \begin{bmatrix} -\frac{2}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{bmatrix}$ we observe that (E, A, C) is not detectable, since

$$[A^H - 5I, C^H] = \begin{bmatrix} -3 & 0 & -\frac{2}{\sqrt{5}} \\ 3 & 0 & \frac{2}{\sqrt{5}} \end{bmatrix}$$

has rank 1. Hence the triples (E, A, B) and (E, A, C) are also not completely controllable and completely observable, respectively; see [15]. In this case, 0 is a simple

singular value of

$$[A - i\gamma E \ B] := \begin{bmatrix} 2 - i\gamma & 3 & 2 \\ 0 & 5 - i\gamma & 2 \end{bmatrix}$$

with corresponding singular vector $v := [-2, \ -2, \ 5+i\gamma]^T$. The range of $[A-i\gamma E \ B]^H$ is spanned by the vectors $u_1 := [2 + i\gamma, \ 3, \ 2]^T$ and $u_2 := [0, \ 5 + i\gamma, \ 2]^T$. Therefore, (H, N) has a purely imaginary eigenvalue if and only if some linear combination of u_1 and u_2 is equal to $\left[\begin{smallmatrix} Q & S \\ S^H & R \end{smallmatrix}\right] v$. This gives rise to the following equations:

$$\begin{aligned} (2 + i\gamma)x_1 &= 5 + i\gamma, \\ 3x_1 + (5 + i\gamma)x_2 &= -(5 + i\gamma), \\ 2x_1 + 2x_2 &= 5(5 + i\gamma). \end{aligned}$$

Eliminating x_1 and x_2 from these equations, we get the relation $\gamma = 5i \notin \mathbb{R}$.

We note that Theorem 4.1 may be generalized to the case when the zero blocks of the perturbation matrix ΔN are filled in such a way that the resulting matrix remains skew-Hermitian, that is, ΔN is replaced by $\Delta\hat{N}$ where

$$\Delta\hat{N} := \begin{bmatrix} 0 & \Delta E & \Delta F \\ -(\Delta E)^H & 0 & \Delta G \\ -(\Delta F)^H & -(\Delta G)^H & 0 \end{bmatrix}. \tag{4.3}$$

In this case, the matrix $(\Delta H - i\gamma\Delta\hat{N})(H - i\gamma N)^{-1}$ has a Frobenius factorization

$$(\Delta H - i\gamma\Delta\hat{N})(H - i\gamma N)^{-1} = \hat{T}(i\gamma)\{S(i\gamma)\}^{-1},$$

where

$$\hat{T}(i\gamma) = \begin{bmatrix} 0 & \Delta A - i\gamma\Delta E & \Delta B - i\gamma\Delta F \\ (\Delta A - i\gamma\Delta E)^H & 0 & \Delta S - i\gamma\Delta G \\ (\Delta B - i\gamma\Delta F)^H & (\Delta S - i\gamma\Delta G)^H & 0 \end{bmatrix}.$$

THEOREM 4.5. [illegible] (H, N) [illegible] ΔH [illegible] $\Delta\hat{N}$ [illegible] (4.1) (4.2) [illegible] (4.3) [illegible]

$$\hat{P}(t, \gamma) := [A - i\gamma E + t(\Delta A - i\gamma\Delta E) \ \ B + t(\Delta B - i\gamma\Delta F)],$$
$$\hat{Z}(t, \gamma) := \begin{bmatrix} Q + t\Delta Q & S + t(\Delta S - i\gamma\Delta G) \\ (S + t(\Delta S - i\gamma\Delta G))^H & R + t\Delta R \end{bmatrix}.$$

[illegible] $\hat{V}(t, \gamma)$ [illegible] $\hat{P}(t, \gamma)$ [illegible] 0 [illegible] $\hat{P}(t, \gamma)^H$ [illegible] $\hat{W}(t, \gamma)$ [illegible] $t \neq 0$ [illegible] γ [illegible] $i\gamma$ [illegible] $(H + t\Delta H, N + t\Delta\hat{N})$ [illegible] $\hat{Z}(t, \gamma)(\hat{V}(t, \gamma)) \cap \hat{W}(t, \gamma) \neq \emptyset$ [illegible]. The proof follows by replacing the set $T(i\gamma)$ by $\hat{T}(i\gamma)$ in the proof of Theorem 4.1. □

It follows trivially that all the results of this section also hold for those special cases when one or more of the blocks $\Delta A, \Delta B, \Delta Q, \Delta R$ and ΔS in the perturbation matrix ΔH or the block ΔE in the perturbation matrix ΔN are equal to 0.

4.2. The continuous-time H_∞ problem. As mentioned in the introduction, in the case of the continuous-time, optimal H_∞-control problem, from (1.6) we have $\Delta N = 0$. Furthermore, the perturbation ΔH of H has a very special structure. All its entries are zero, except for the entries of the block ΔR which is itself a special diagonal matrix, the first few entries on the main diagonal being each equal to -1 and the remaining all being equal to zero. Due to this special structure, all the eigenvalues of $(\Delta H - z\Delta N)(H - zN)^{-1}$ are real and its nonzero eigenvalues are precisely the nonzero eigenvalues of a leading principal submatrix of a Hermitian matrix whose size is the same as that of the block R of H.

THEOREM 4.6. [illegible] H [illegible] N [illegible] (4.1) [illegible] $R := \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$ [illegible] ΔH [illegible] ΔN [illegible] (4.2) [illegible] $\Delta A = \Delta B = \Delta Q = \Delta S = \Delta E = 0$ [illegible] $\Delta R := \begin{bmatrix} -I_j & 0 \\ 0 & 0 \end{bmatrix}$, [illegible] I_j [illegible] j, [illegible] R [illegible] $\gamma \in \mathbb{R}$ [illegible] $(\Delta H - i\gamma\Delta N)(H - i\gamma N)^{-1}$ [illegible] j [illegible]

$$W(\gamma) := R^{-1}\left(\begin{bmatrix} B \\ S \end{bmatrix}^H \begin{bmatrix} -BR^{-1}B^H & A - i\gamma E - BR^{-1}S^H \\ \left(A - i\gamma E - BR^{-1}S^H\right)^H & Q - SR^{-1}S^H \end{bmatrix}^{-1} \times \begin{bmatrix} B \\ S \end{bmatrix} + R\right) R^{-1}.$$

[illegible]. For $\gamma \in \mathbb{R}$,

$$\begin{aligned} H - i\gamma N &= \begin{bmatrix} 0 & A - i\gamma E & B \\ A^H + i\gamma E^H & Q & S \\ B^H & S^H & R \end{bmatrix} \\ &= \begin{bmatrix} I & 0 & BR^{-1} \\ 0 & I & SR^{-1} \\ 0 & 0 & I \end{bmatrix} \\ &\quad \times \begin{bmatrix} -BR^{-1}B^H & A - i\gamma E - BR^{-1}S^H & 0 \\ A^H + i\gamma E^H - SR^{-1}B^H & Q - SR^{-1}S^H & 0 \\ B^H & S^H & R \end{bmatrix}. \end{aligned}$$

Therefore,

$$\begin{aligned} (H - i\gamma N)^{-1} &= \left[\begin{array}{cc|c} -BR^{-1}B^H & A - i\gamma E - BR^{-1}S^H & 0 \\ A^H + i\gamma E^H - SR^{-1}B^H & Q - SR^{-1}S^H & 0 \\ \hline B^H & S^H & R \end{array}\right]^{-1} \\ &\quad \times \begin{bmatrix} I & 0 & -BR^{-1} \\ 0 & I & -SR^{-1} \\ 0 & 0 & I \end{bmatrix}. \end{aligned}$$

Let

$$\tilde{H} := \begin{bmatrix} -BR^{-1}B^H & A - i\gamma E - BR^{-1}S^H \\ A^H + i\gamma E^H - SR^{-1}B^H & Q - SR^{-1}S^H \end{bmatrix}, \qquad M := \begin{bmatrix} -BR^{-1} \\ -SR^{-1} \end{bmatrix},$$

and $Z := \begin{bmatrix} B^H & S^H \end{bmatrix}$. Then

$$\begin{aligned}(H - i\gamma N)^{-1} &= \begin{bmatrix} \tilde{H} & 0 \\ Z & R \end{bmatrix}^{-1} \begin{bmatrix} I & M \\ 0 & I \end{bmatrix} = \begin{bmatrix} \tilde{H}^{-1} & 0 \\ -R^{-1}Z\tilde{H}^{-1} & R^{-1} \end{bmatrix} \begin{bmatrix} I & M \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} \tilde{H}^{-1} & \tilde{H}^{-1}M \\ -R^{-1}Z\tilde{H}^{-1} & -R^{-1}Z\tilde{H}^{-1}M + R^{-1} \end{bmatrix} \\ &= \begin{bmatrix} \tilde{H}^{-1} & \tilde{H}^{-1}M \\ M^H\tilde{H}^{-1} & -M^H\tilde{H}^{-1}M + R^{-1} \end{bmatrix},\end{aligned}$$

which is Hermitian, since $\tilde{H}$ is Hermitian. Then, we obtain

$$\begin{aligned}(\Delta H - i\gamma\Delta N)(H - i\gamma N)^{-1} &= \begin{bmatrix} 0 & 0 \\ 0 & \Delta R \end{bmatrix} \begin{bmatrix} \tilde{H}^{-1} & \tilde{H}^{-1}M \\ M^H(\tilde{H})^{-1} & M^H\tilde{H}^{-1}M + R^{-1} \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 \\ (\Delta R)M^H\tilde{H}^{-1} & (\Delta R)(M^H\tilde{H}^{-1}M + R^{-1}) \end{bmatrix}.\end{aligned}$$

Therefore, the matrix $(\Delta H - i\gamma\Delta N)(H - i\gamma N)^{-1}$ has a nonzero real eigenvalue if and only if the matrix $(\Delta R)(M^H\tilde{H}^{-1}M + R^{-1})$ has a nonzero real eigenvalue. Replacing M by $\left[\begin{smallmatrix} -BR^{-1} \\ -SR^{-1} \end{smallmatrix}\right]$ we have

$$\begin{aligned}(\Delta R)(M^H\tilde{H}^{-1}M + R^{-1}) &= (\Delta R)R^{-1}\left(\begin{bmatrix} B \\ S \end{bmatrix}^H \tilde{H}^{-1} \begin{bmatrix} B \\ S \end{bmatrix} + R\right) R^{-1} \\ &= (\Delta R)W(\gamma).\end{aligned}$$

Note that since the matrices R and Q are Hermitian, $W(\gamma)$ is also Hermitian and hence all its eigenvalues are real. Let

$$W(\gamma) := \begin{bmatrix} W_{11}(\gamma) & W_{12}(\gamma) \\ W_{21}(\gamma) & W_{22}(\gamma) \end{bmatrix} \tag{4.4}$$

be a partition of $W(\gamma)$, conformal with that of ΔR. In view of the structure of ΔR, it follows that

$$(\Delta R)W(\gamma) = \begin{bmatrix} -W_{11}(\gamma) & -W_{12}(\gamma) \\ 0 & 0 \end{bmatrix}.$$

Hence $(\Delta H - i\gamma\Delta N)(H - i\gamma N)^{-1}$ has a nonzero real eigenvalue if and only if the block $W_{11}(\gamma)$ has a nonzero real eigenvalue. The proof follows from the fact that $W_{11}(\gamma)$ is Hermitian. ∎

From Theorem 4.6 it follows that for the continuous time H_∞-control problem we have $\mathrm{sep}_{\mathbb{R}}(z, \Delta H, \Delta N) = \{r((\Delta H - z\Delta N)(H - zN)^{-1})\}^{-1}$ for all purely imaginary complex numbers z. However, for these problems we are interested only in positive values of the parameter t for which $(H + t\Delta H, N + t\Delta N)$ has a purely imaginary eigenvalue. The following immediate corollary of Theorem 4.6 suggests a procedure for obtaining the exact value of the smallest positive parameter t for which the pencil $(H + t\Delta H, N + t\Delta N)$ has a purely imaginary eigenvalue or an upper or lower bound of this value.

COROLLARY 4.7. [illegible] $H, N, \Delta H$ [illegible] ΔN [illegible] 4.6 [illegible] t [illegible] $(H+t\Delta H, N+t\Delta N)$ [illegible] $t = t_0$ [illegible]

(i) $1/t_0$ [illegible] ϵ [illegible]

$$L(\epsilon, \Delta H, \Delta N) := \{z \in \mathbb{C} \setminus \sigma(H, N) : r((\Delta H - z\Delta N)(H - zN)^{-1}) = \epsilon^{-1}\}$$

[illegible]

(ii) $1/t_0$ [illegible] $W_{11}(\gamma_0)$, [illegible] (4.4)
$i\gamma_0$ [illegible] $L(1/t_0, \Delta H, \Delta N)$ [illegible]
[illegible] $W_{11}(\gamma_0)$ [illegible] t [illegible] $(H+t\Delta H, N+t\Delta N)$ [illegible]
[illegible] $\{r(W_{11}(\gamma_0))\}^{-1}$.
[illegible] $W_{11}(\gamma_0)$ [illegible] $r(W_{11}(\gamma_0))$,
α [illegible] t [illegible] $(H + t\Delta H, N + t\Delta N)$ [illegible]
[illegible] $1/\alpha$.

[illegible]. The proof follows immediately from Theorem 4.6 in view of the fact that given a positive real number t_0, $i\gamma_0$ is a purely imaginary eigenvalue of $(H+t_0\Delta H, N+t_0\Delta N)$ if and only if $1/t_0$ is an eigenvalue of $W_{11}(\gamma_0)$. □

In this section we have discussed linear perturbation theory for structured pencils arising in continuous-time control theory. In the next section we discuss analogous results for discrete-time control problems.

5. Perturbation of structured pencils arising from discrete-time control. There is a well-known analogy between continuous- and discrete-time linear-quadratic optimal control problems, given by the Cayley transformation; see [17, 18]. Thus, we expect similar results for the discrete-time case. For these problems the pencils have the following structures:

$$H := \begin{bmatrix} 0 & A & B \\ -E^H & Q & S \\ 0 & S^H & R \end{bmatrix}, \qquad N := \begin{bmatrix} 0 & E & 0 \\ -A^H & 0 & 0 \\ -B^H & 0 & 0 \end{bmatrix}, \tag{5.1}$$

$$\Delta H := \begin{bmatrix} 0 & \Delta A & \Delta B \\ -(\Delta E)^H & \Delta Q & \Delta S \\ 0 & (\Delta S)^H & \Delta R \end{bmatrix}, \qquad \Delta N := \begin{bmatrix} 0 & \Delta E & 0 \\ -(\Delta A)^H & 0 & 0 \\ -(\Delta B)^H & 0 & 0 \end{bmatrix}, \tag{5.2}$$

where again $Q^H = Q$, $R^H = R, (\Delta Q)^H = \Delta Q$, and $(\Delta R)^H = \Delta R$ have the same dimensions as in (4.1). Although we consider complex matrices, the results of this section are true for real matrices as well.

In this case $\mathbb{C}_g := \{z \in \mathbb{C} : |z| \neq 1\}$, and the smallest $|t|$, $t \in \mathbb{R}$ such that the perturbed pencil $(H+t\Delta H, N+t\Delta N)$ has an eigenvalue $z \in \mathbb{C}$, $|z| = 1$ is the quantity of interest for these problems. This is equivalent to the matrix $(\Delta H - z\Delta N)(H - zN)^{-1}$ having a real eigenvalue for some $z \in \mathbb{C}$ such that $|z| = 1$. We show first that for any $z \in \mathbb{C}$ on the unit circle, Hermitian Frobenius factors $T(z)$ and $S(z)$ of the matrix $(\Delta H - z\Delta N)(H - zN)^{-1}$ may be obtained from the matrices $\Delta H - z\Delta N$ and $H - zN$ by a simple scaling.

THEOREM 5.1. [illegible] $H, N, \Delta H$ [illegible] ΔN [illegible] (5.1) [illegible] (5.2) [illegible] $z \in \mathbb{C}$, $|z| = 1$,

$$(\Delta H - z\Delta N)(H - zN)^{-1} = \begin{bmatrix} 0 & \Delta A - z\Delta E & \Delta B \\ (\Delta A - z\Delta E)^H & \Delta Q & \Delta S \\ (\Delta B)^H & (\Delta S)^H & \Delta R \end{bmatrix} \begin{bmatrix} 0 & A - zE & B \\ (A - zE)^H & Q & S \\ B^H & S^H & R \end{bmatrix}^{-1}.$$

. For $|z| = 1$, we have

$$H - zN = \begin{bmatrix} 0 & A - zE & B \\ z(A - zE)^H & Q & S \\ zB^H & S^H & R \end{bmatrix} = \begin{bmatrix} 0 & A - zE & B \\ (A - zE)^H & Q & S \\ B^H & S^H & R \end{bmatrix} \begin{bmatrix} zI & & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}.$$

Similarly,

$$\Delta H - z\Delta N = \begin{bmatrix} 0 & \Delta A - z\Delta E & \Delta B \\ (\Delta A - z\Delta E)^H & \Delta Q & \Delta S \\ (\Delta B)^H & (\Delta S)^H & \Delta R \end{bmatrix} \begin{bmatrix} zI & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}.$$

The proof follows from the fact that the matrix

$$\begin{bmatrix} zI & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}$$

is unitary, since $|z| = 1$. □

Hence, for all $z \in \mathbb{C}$ such that $|z| = 1$, we get Hermitian Frobenius factors $T(z)$ and $S(z)$ of $(\Delta H - z\Delta N)(H - zN)^{-1}$ as

$$T(z) := \begin{bmatrix} 0 & \Delta A - z\Delta E & \Delta B \\ (\Delta A - z\Delta E)^H & \Delta Q & \Delta S \\ (\Delta B)^H & (\Delta S)^H & \Delta R \end{bmatrix},$$

$$S(z) := \begin{bmatrix} 0 & A - zE & B \\ (A - zE)^H & Q & S \\ B^H & S^H & R \end{bmatrix}.$$

Thus, given a complex number z lying on the unit circle, the results of section 3 may be applied to these Frobenius factors to obtain necessary and sufficient conditions for the matrix $(\Delta H - z\Delta N)(H - zN)^{-1}$ to have a real eigenvalue. This in turn gives us necessary and sufficient conditions for the matrix pencil $(H + t\Delta H, N + t\Delta N)$ to have an eigenvalue on the unit circle. However, as in the continuous-time case, the special structure of the Frobenius factors leads to another necessary and sufficient condition for the pencil $(H + t\Delta H, N + t\Delta N)$ to have an eigenvalue on the unit circle on the lines of Theorem 4.1.

THEOREM 5.2. (H, N) ΔH ΔN (5.1) (5.2) $z \in \mathbb{C}$ $|z| = 1$,

$$P(t, z) := [A - zE + t(\Delta A - z\Delta E) \;\; B + t\Delta B],$$

$$Z(t) := \begin{bmatrix} Q + t\Delta Q & S + t\Delta S \\ (S + t\Delta S)^H & R + t\Delta R \end{bmatrix}.$$

$V(t, z)$ $P(t, z)$ 0 $W(t, z)$ $P(t, z)^H$ $t \neq 0$ $z \in \mathbb{C}$, $|z| = 1$ $(H + t\Delta H, N + t\Delta N)$ $Z(t)(V(t, z)) \cap W(t, z) \neq \emptyset$

. The proof follows by replacing $i\gamma$ by $z \in \mathbb{C}$, $|z| = 1$ in the proof of Theorem 4.1. ∎

As in the continuous-time case, if we assume that the matrices Q, S, and R which are associated with the cost function are unperturbed, that is, if $\Delta Q = \Delta S = \Delta R = 0$, then we have the following corollary to Theorem 5.2. It characterizes the choice of a cost function such that given $t \in \mathbb{R}$, and $z \in \mathbb{C}$ such that $|z| = 1$, z is not an eigenvalue of $(H + t\Delta H, N + t\Delta N)$.

COROLLARY 5.3. $W(t,z)$ $V(t,z)$ 5.2 $Z_0 := \begin{bmatrix} Q & S \\ S^H & R \end{bmatrix}$ $(H + t\Delta H, N + t\Delta N)$ z $|z| = 1$ $Z_0(V(t,z)) \cap W(t,z) \neq \emptyset$

. The proof follows immediately from Theorem 5.2 by the fact that $Z(0) = Z_0$. ∎

The next corollary provides a characterization of all cost functions such that for a fixed $t \in \mathbb{R}$, the pencil $(H + t\Delta H, N + t\Delta N)$ does not have any eigenvalues on the unit circle.

COROLLARY 5.4. $P(t,z)$ $V(t,z)$ 5.2 $Z_0 := \begin{bmatrix} Q & S \\ S^H & R \end{bmatrix}$ $t \in \mathbb{R}$ $(H + t\Delta H, N + t\Delta N)$

$$Z_0(\cup_{|z|=1} V(t,z)) \cap (\cup_{|z|=1} W(t,z)) = \emptyset.$$

The results of this section also hold if the perturbation matrices ΔH and ΔN are replaced by the matrices $\Delta\hat{H}$ and $\Delta\hat{N}$, respectively, which are given by

$$\Delta\hat{H} := \begin{bmatrix} 0 & \Delta A & \Delta B \\ -(\Delta E)^H & \Delta Q & \Delta S \\ -(\Delta F)^H & (\Delta S)^H & \Delta R \end{bmatrix}, \qquad \Delta\hat{N} := \begin{bmatrix} 0 & \Delta E & \Delta F \\ -(\Delta A)^H & 0 & 0 \\ -(\Delta B)^H & 0 & 0 \end{bmatrix}. \tag{5.3}$$

Then for $z \in \mathbb{C}$ such that $|z| = 1$, $(\Delta\hat{H} - z\Delta\hat{N})(H - zN)^{-1} = \hat{T}(z)S(z)^{-1}$ is a Frobenius factorization of $(\Delta\hat{H} - z\Delta\hat{N})(H - zN)^{-1}$, where

$$\hat{T}(z) := \begin{bmatrix} 0 & \Delta A - z\Delta E & \Delta B - z\Delta F \\ (\Delta A - z\Delta E)^H & \Delta Q & \Delta S \\ (\Delta B - z\Delta F)^H & (\Delta S)^H & \Delta R \end{bmatrix}.$$

With these new perturbation matrices, Theorem 5.2 takes the following form.

THEOREM 5.5. (H, N) $\Delta\hat{H}$ $\Delta\hat{N}$ (5.1) (5.3)

$$\hat{P}(t,z) := [A - zE + t(\Delta A - z\Delta E) \;\; B + t(\Delta B - z\Delta F)],$$
$$Z(t) := \begin{bmatrix} Q + t\Delta Q & S + t\Delta S \\ (S + t(\Delta S))^H & R + t\Delta R \end{bmatrix}.$$

$\hat{V}(t,z)$ $\hat{P}(t,z)$ 0 $\hat{W}(t,z)$ $\hat{P}(t,z)^H$ $t \neq 0$ $z \in \mathbb{C}$ $|z| = 1$ $(H + t\Delta\hat{H}, N + t\Delta\hat{N})$ $Z(t)(\hat{V}(t,z)) \cap \hat{W}(t,z) \neq \emptyset$

. The proof follows by using arguments similar to those of Theorem 4.1 with the matrix $P(t, i\gamma)$ being replaced by $\hat{P}(t, z)$. ∎

Given $t \in \mathbb{R}$ and $z \in \mathbb{C}$, $|z| = 1$, the following corollary provides a characterization of the cost function such that $z \notin \sigma(H + t\Delta\hat{H}, N + t\Delta\hat{N})$.

COROLLARY 5.6. $\hat{P}(t,z), \hat{V}(t,z)$ $\hat{W}(t,z)$ 5.5 $Z_0 := \left[\begin{smallmatrix} Q & S \\ S^H & R \end{smallmatrix}\right]$ t $z \in \mathbb{C}$ $|z| = 1$, $z \notin \sigma(H + t\Delta\hat{H}, N + t\Delta\hat{N})$ $Z_0(\hat{V}(t,z)) \cap \hat{W}(t,z) \neq \emptyset$

Finally we have a characterization of the cost function such that for a given $t \in \mathbb{R}$, the pencil $(H + t\Delta\hat{H}, N + t\Delta\hat{N})$ has no eigenvalues on the unit circle.

COROLLARY 5.7. $\hat{P}(t,z), \hat{V}(t,z)$ $\hat{W}(t,z)$ 5.5 $Z_0 := \left[\begin{smallmatrix} Q & S \\ S^H & R \end{smallmatrix}\right]$ t $(H + t\Delta\hat{H}, N + t\Delta\hat{N})$

$$Z_0(\cup_{|z|=1}\hat{V}(t,z)) \cap (\cup_{|z|=1}\hat{W}(t,z)) \neq \emptyset.$$

Note that all results hold if any one or more of the blocks in the perturbation matrices ΔH and ΔN are equal to 0.

6. The discrete-time H_∞ problem. In the discrete-time optimal H_∞-control problem, the matrices H and N of the pencil (H, N) are also given by (5.1). But as in the case of its continuous time analogue, the perturbations ΔN and ΔH are very special. From equation (1.7) we have $\Delta N = 0$ and all the blocks of ΔH are zero except for ΔR which is a special diagonal matrix. Only the first few diagonal entries of ΔR are nonzero and these are each equal to -1. We show that given $|z| = 1$, all the eigenvalues of the matrix $(\Delta H - z\Delta N)(H - zN)^{-1}$ are real and the nonzero eigenvalues are precisely the same as those of a leading principal submatrix of a Hermitian matrix which is of the same size as the block R of H.

THEOREM 6.1. H N (5.1) $R := \left[\begin{smallmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{smallmatrix}\right]$ ΔH ΔN (5.2) $\Delta A = \Delta B = \Delta Q = \Delta S = \Delta E = 0$ $\Delta R := \left[\begin{smallmatrix} -I_j & 0 \\ 0 & 0 \end{smallmatrix}\right]$, I_j j, R $z \in \mathbb{C}$ $|z| = 1$, $(\Delta H - z\Delta N)(H - zN)^{-1}$ j

$$W(z) := R^{-1}\left(\begin{bmatrix} B \\ S \end{bmatrix}^H \begin{bmatrix} -BR^{-1}B^H & A - zE - BR^{-1}S^H \\ \left(A - zE - BR^{-1}S^H\right)^H & Q - SR^{-1}S^H \end{bmatrix}^{-1} \times \begin{bmatrix} B \\ S \end{bmatrix} + R\right) R^{-1}.$$

. Since R and Q are Hermitian, it is clear that $W(z)$ is a Hermitian matrix. For $z \in \mathbb{C}, |z| = 1$, we have,

$$H - zN = \begin{bmatrix} 0 & A - zE & B \\ -E^H + zA^H & Q & S \\ zB^H & S^H & R \end{bmatrix} = \begin{bmatrix} I & 0 & BR^{-1} \\ 0 & I & SR^{-1} \\ 0 & 0 & I \end{bmatrix} \times \begin{bmatrix} -zBR^{-1}B^H & A - zE - BR^{-1}S^H & 0 \\ z(A - zE - BR^{-1}S^H)^H & Q - SR^{-1}S^H & 0 \\ zB^H & S^H & R \end{bmatrix}.$$

Therefore,

$$(H-zN)^{-1} = \begin{bmatrix} -zBR^{-1}B^H & A-zE-BR^{-1}S^H & 0 \\ -E^H+zA^H-zSR^{-1}B^H & Q-SR^{-1}S^H & 0 \\ zB^H & S^H & R \end{bmatrix}^{-1} \times \begin{bmatrix} I & 0 & -BR^{-1} \\ 0 & I & -SR^{-1} \\ 0 & 0 & I \end{bmatrix}.$$

Let

$$\tilde{H} = \begin{bmatrix} -zBR^{-1}B^H & A-zE-BR^{-1}S^H \\ z(A^H-\overline{z}E^H-SR^{-1}B^H) & Q-SR^{-1}S^H \end{bmatrix}, \qquad M = \begin{bmatrix} -BR^{-1} \\ -SR^{-1} \end{bmatrix}$$

and $Z = \begin{bmatrix} zB^H & S^H \end{bmatrix}$. Then,

$$(H-zN)^{-1} = \begin{bmatrix} \tilde{H} & 0 \\ Z & R \end{bmatrix}^{-1} \begin{bmatrix} I & M \\ 0 & I \end{bmatrix} = \begin{bmatrix} \tilde{H}^{-1} & 0 \\ -R^{-1}Z\tilde{H}^{-1} & R^{-1} \end{bmatrix} \begin{bmatrix} I & M \\ 0 & I \end{bmatrix} = \begin{bmatrix} \tilde{H}^{-1} & \tilde{H}^{-1}M \\ -R^{-1}Z\tilde{H}^{-1} & -R^{-1}Z\tilde{H}^{-1}M+R^{-1} \end{bmatrix}.$$

Let $\tilde{J} = \begin{bmatrix} \overline{z}I & 0 \\ 0 & I \end{bmatrix}$ where the partitioning is conformal with that of M. Since $|z|=1$, it is clear that $\tilde{J}$ is unitary and $R^{-1}Z = -M^H\tilde{J}^{-1}$. This gives $R^{-1}Z\tilde{H}^{-1} = -M^H(\tilde{H}\tilde{J})^{-1}$ and hence,

$$(H-zN)^{-1} = \begin{bmatrix} \tilde{H}^{-1} & \tilde{H}^{-1}M \\ M^H(\tilde{H}\tilde{J})^{-1} & M^H(\tilde{H}\tilde{J})^{-1}M+R^{-1} \end{bmatrix}.$$

Therefore,

$$(\Delta H-z\Delta N)(H-zN)^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & \Delta R \end{bmatrix} \begin{bmatrix} \tilde{H}^{-1} & \tilde{H}^{-1}M \\ M^H(\tilde{H}\tilde{J})^{-1} & M^H(\tilde{H}\tilde{J})^{-1}M+R^{-1} \end{bmatrix}.$$

This shows that the matrix $(\Delta H-z\Delta N)(H-zN)^{-1}$ has nonzero real eigenvalues if and only if the matrix

$$(\Delta R)(M^H(\tilde{H}\tilde{J})^{-1}M+R^{-1})$$

has nonzero real eigenvalues or equivalently

$$(\Delta R)R^{-1}\left(\begin{bmatrix} B \\ S \end{bmatrix}^H(\tilde{H}\tilde{J})^{-1}\begin{bmatrix} B \\ S \end{bmatrix}+R\right)R^{-1}$$

has nonzero real eigenvalues. Since,

$$\tilde{H}\tilde{J} = \begin{bmatrix} -BR^{-1}B^H & A-zE-BR^{-1}S^H \\ \left(A-zE-BR^{-1}S^H\right)^H & Q-SR^{-1}S^H \end{bmatrix},$$

we have

$$(\Delta R)R^{-1}\left(\begin{bmatrix} B \\ S \end{bmatrix}^H(\tilde{H}\tilde{J})^{-1}\begin{bmatrix} B \\ S \end{bmatrix}+R\right)R^{-1} = (\Delta R)W(z).$$

Let

$$W(z) := \begin{bmatrix} W_{11}(z) & W_{12}(z) \\ W_{21}(z) & W_{22}(z) \end{bmatrix} \tag{6.1}$$

be a partition of $W(z)$ conformal with that of ΔR. In view of the structure of ΔR, we have

$$(\Delta R)W(z) = \begin{bmatrix} -W_{11}(z) & W_{12}(z) \\ 0 & 0 \end{bmatrix}.$$

Hence the nonzero real eigenvalues of $(\Delta H - z\Delta N)(H - zN)^{-1}$ are the same as those of $-W_{11}(z)$ and the proof follows from the fact that $W_{11}(z)$ is Hermitian. ∎

By Theorem 6.1, it is clear that every point on the unit circle becomes an eigenvalue of $(H + t\Delta H, N + t\Delta N)$ for some real number t. However, as in the case of the continuous-time H_∞ problem, we are interested only in the positive values of the parameter t for which the pencil $(H + t\Delta H, N + t\Delta N)$ has eigenvalues on the unit circle. Since, for any $z \notin \sigma(H, N)$, we have $z \in \sigma(H + t\Delta H, N + t\Delta N)$ if and only if $-1/t \in \sigma(\Delta H - z\Delta N)(H - zN)^{-1}$, it follows from Theorem 6.1 that there exists $t > 0$ such that some $z \in \mathbb{C}$ with $|z| = 1$ is an eigenvalue of $(H + t\Delta H, N + t\Delta N)$ if and only if the matrix $W_{11}(z)$ has a positive eigenvalue. This suggests the following procedure for finding the smallest positive number t for which $(H + t\Delta H, N + t\Delta N)$ has an eigenvalue on the unit circle on the lines of Corollary 4.7.

COROLLARY 6.2. $H, N, \Delta H$ ΔN 6.1 t $(H+t\Delta H, N+t\Delta N)$ z $|z| = 1$ $t = t_0$

(i) $1/t_0$ ϵ

$$L(\epsilon, \Delta H, \Delta N) := \{z \in \mathbb{C} \setminus \sigma(H, N) : r((\Delta H - z\Delta N)(H - zN)^{-1}) = \epsilon^{-1}\}$$

(ii) $1/t_0$ $W_{11}(z_0)$, (6.1) z_0 $L(1/t_0, \Delta H, \Delta N)$ $W_{11}(z_0)$ t $(H + t\Delta H, N + t\Delta N)$ $\{r(W_{11}(z_0))\}^{-1}$.

$W_{11}(z_0)$ $r(W_{11}(z_0))$ α t $(H + t\Delta H, N + t\Delta N)$ $1/\alpha$.

. The proof is an immediate consequence of Theorem 6.1. ∎

7. Conclusion and future work. We have studied the effect of linear perturbations on several structured matrix pencils arising in control theory. These include skew-symmetric/symmetric pencils arising in the computation of optimal H_∞ control and linear-quadratic control for continuous- and discrete-time systems. We have given characterizations when these pencils have eigenvalues on the imaginary axis or the unit circle, respectively.

But several important questions remain open. Among these are a characterization of the Kronecker structure associated with eigenvalues on the imaginary axis and the development of numerical methods for the efficient computation of the smallest perturbations that move eigenvalues to the imaginary axis or unit circle, respectively. We will address these issues in our future work.

REFERENCES

[1] R. Alam and S. Bora, *Stability of eigenvalues and spectral decompositions under linear perturbation*, Linear Algebra Appl., 364 (2003), pp. 189–211.

[2] E. Anderson, Z. Bai, C. H. Bischof, J. M. Demmel, J. J. Dongarra, J. J. Du Croz, A. Greenbaum, S. J. Hammarling, A. McKenney, S. Ostrouchov, and D. C. Sorensen, *LAPACK Users' Guide*, 2nd ed., SIAM, Philadelphia, 1994.

[3] H. Baumgärtel, *Analytic Perturbation Theory for Matrices and Operators*, Birkhäuser, Philadelphia, 1985.

[4] P. Benner, R. Byers, V. Mehrmann, and H. Xu, *Numerical methods for linear quadratic and H_∞ control problems*, in Dynamical Systems, Control, Coding, Computer Vision. Prog. Syst. Control Theory 25, G. Picci and D. S. Gillian, eds., Birkhäuser-Verlag, Basel, Switzerland, 1999, pp. 203–222.

[5] P. Benner, R. Byers, V. Mehrmann, and H. Xu, *Robust method for robust control*, preprint 2004-6, Institut für Mathematik, TU Berlin, FRG, 2004.

[6] 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.

[7] R. Byers, T. Geerts, and V. Mehrmann, *Descriptor systems without controllability at infinity*, SIAM J. Control Optim., 35 (1997), pp. 462–479.

[8] G. Freiling, V. Mehrmann, and H. Xu, *Existence, uniqueness and parametrization of Lagrangian invariant subspaces*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 1045–1069.

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

[10] M. Green and D. J. N Limebeer, *Linear Robust Control*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

[11] D. Hinrichsen and A. J. Pritchard, *Mathematical Systems Theory I. Modelling, State Space Analysis, Stability and Robustness*, Springer-Verlag, Berlin, 2005.

[12] T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[13] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[14] M. M. Konstantinov, V. Mehrmann, and P. Hr. Petkov, *Perturbation analysis for the Hamiltonian Schur form*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 387–424.

[15] P. Lancaster and L. Rodman, *The Algebraic Riccati Equation*, Oxford University Press, Oxford, 1995.

[16] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, 2nd ed., Academic Press, Orlando, FL, 1985.

[17] V. Mehrmann, *The Autonomous Linear Quadratic Control Problem: Theory and Numerical Solution*, Lecture Notes in Control and Inform. Sci. 163, Springer-Verlag, New York, 1991.

[18] V. Mehrmann, *A step towards a unified treatment of continuous and discrete time control problems*, Linear Algebra Appl., 241–243 (1996), pp. 749–779.

[19] M. L. Overton and P. Van Dooren, *On computing the complex passivity radius*, in Proceedings of the IEEE Conference on Decision and Control, 2005, Sevilla, Spain, pp. 7960–7964.

[20] L. S. Pontryagin, V. Boltyanskii, R. Gamkrelidze, and E. Mishenko, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[21] A. C. M. Ran and L. Rodman, *Stability of invariant maximal semidefinite subspaces*, Linear Algebra Appl., 62 (1984), pp. 51–86.

[22] A. C. M. Ran and L. Rodman, *Stability of invariant Lagrangian subspaces i*, in Topics in Operator Theory, Oper. Theory Adv. Appl. 32, I. Gohberg, ed., Birkhäuser, Basel, Switzerland, 1988, pp. 181–218.

[23] A. C. M. Ran and L. Rodman, *Stability of invariant Lagrangian subspaces ii*, in the Gohberg Anniversary Collection, Oper. Theory Adv. Appl. 40, H. Dym, S. Goldberg, M. A. Kaashoek, and P. Lancaster, eds., Birkhäuser, Basel, Switzerland, 1989, pp. 391–425.

[24] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

[25] F. Uhlig, *Inertia and eigenvalue relations between symmetrized and symmetrizing matrices for the real and the general field case*, Linear Algebra Appl., 35 (1981), pp. 203–226.

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

[27] P. Van Dooren, *A generalized eigenvalue approach for solving Riccati equations*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 121–135.

[28] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice-Hall, Upper Saddle River, NJ, 1995.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 1, pp. 170–189

IMPLICIT-FACTORIZATION PRECONDITIONING AND ITERATIVE SOLVERS FOR REGULARIZED SADDLE-POINT SYSTEMS*

H. SUE DOLLAR†, NICHOLAS I. M. GOULD‡, WIL H. A. SCHILDERS§, AND ANDREW J. WATHEN†

Abstract. We consider conjugate-gradient like methods for solving block symmetric indefinite linear systems that arise from saddle-point problems or, in particular, regularizations thereof. Such methods require preconditioners that preserve certain sub-blocks from the original systems but allow considerable flexibility for the remaining blocks. We construct a number of families of implicit factorizations that are capable of reproducing the required sub-blocks and (some) of the remainder. These generalize known implicit factorizations for the unregularized case. Improved eigenvalue clustering is possible if additionally some of the noncrucial blocks are reproduced. Numerical experiments confirm that these implicit-factorization preconditioners can be very effective in practice.

Key words. regularized saddle-point systems, implicit-factorization preconditioners

AMS subject classifications. 15A23, 65F10, 90C20

DOI. 10.1137/05063427X

1. Introduction.

Given a symmetric n by n matrix H, a symmetric m by m ($m \leq n$) matrix C and a full-rank m ($\leq n$) by n matrix A, we are interested in solving structured linear systems of equations

$$\begin{pmatrix} H & A^T \\ A & -C \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = - \begin{pmatrix} g \\ 0 \end{pmatrix} \tag{1.1}$$

by iterative methods, in which preconditioners of the form

$$M_G = \begin{pmatrix} G & A^T \\ A & -C \end{pmatrix} \tag{1.2}$$

are used to accelerate the iteration for some suitable symmetric G. We denote the coefficient matrix in (1.1) by M_H. There is little loss of generality in assuming the right-hand side of (1.1) has the form given rather than with the more general

$$\begin{pmatrix} H & A^T \\ A & -C \end{pmatrix} \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} = \begin{pmatrix} b \\ c \end{pmatrix}. \tag{1.3}$$

For, so long as we have some mechanism for finding an initial (x_0, y_0) for which $Ax_0 - Cy_0 = c$, linearity of (1.1) implies that $(\bar{x}, \bar{y}) = (x_0 - x, y_0 - y)$ solves (1.3)

*Received by the editors June 22, 2005; accepted for publication (in revised form) by V. Simoncini October 31, 2005; published electronically March 17, 2006.
http://www.siam.org/journals/simax/28-1/63427.html

†Oxford University Computing Laboratory, Numerical Analysis Group, Wolfson Building, Parks Road, Oxford, OX1 3QD, England, UK (Sue.Dollar@comlab.ox.ac.uk, Andy.Wathen@comlab.ox.ac.uk).

‡Computational Science and Engineering Department, Rutherford Appleton Laboratory, Chilton, Oxfordshire, OX11 0QX, England, UK (n.i.m.gould@rl.ac.uk). This work was supported by the EPSRC grant GR/S42170.

§Philips Research Laboratories, Prof. Holstlaan 4, 5656 AA Eindhoven, The Netherlands (wil.schilders@philips.com). Also, Department of Mathematics and Computer Science, Technische Universiteit Eindhoven, PO Box 513, 5600 MB Eindhoven, The Netherlands.

when $b = g + Hx_0 + A^T y_0$. In particular, since we intend to use the preconditioner (1.2), solving

$$\begin{pmatrix} G & A^T \\ A & -C \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} 0 \\ c \end{pmatrix} \text{ or } = \begin{pmatrix} b \\ c \end{pmatrix} \tag{1.4}$$

to find suitable (x_0, y_0) are distinct possibilities.

When $C = 0$, (1.2) is commonly known as a constraint preconditioner [35] and in this case systems of the form (1.1) arise as stationarity (KKT) conditions for equality-constrained optimization [40, section 18.1], in mixed finite-element approximation of elliptic problems [6], including, in particular, problems of elasticity [41] and incompressible flow [23], as well as other areas. In practice C is often positive semi-definite (and frequently diagonal)—such systems frequently arise in interior-point and regularization methods in optimization, the simulation of electronic circuits [47] and other related areas; see [3] for an encyclopedic review of (regularized) saddle-point systems. Although such problems may involve m by n A with $m > n$, this is not a restriction, for in this case we might equally solve

$$\begin{pmatrix} C & A \\ A^T & -H \end{pmatrix} \begin{pmatrix} y \\ -x \end{pmatrix} = \begin{pmatrix} 0 \\ g \end{pmatrix},$$

for which A^T has more columns than rows. We place no restrictions on H, although we recognize that in some applications H may be positive (semi-) definite.

Let I be the (appropriately dimensioned) identity matrix. Given a symmetric matrix M with, respectively, m_+, m_- and m_0 positive, negative and zero eigenvalues, we denote its inertia by $\text{In}(M) = (m_+, m_-, m_0)$.

2. Suitable iterative methods. While it would be perfectly possible to apply general preconditioned iterative methods like GMRES [45] or the symmetric QMR method [25] to (1.1) with the indefinite preconditioner (1.2), the specific form of (1.2) allows the use of the more efficient preconditioned conjugate-gradient (PCG) method [12] instead. Use of GMRES would have the disadvantage that storage and orthogonalization of a set of k vectors would be required at the kth iteration, so that the work per iteration increases at each iteration. The symmetric QMR method does not suffer from this difficulty, though it does not minimize a quantity of interest (such as the residual) unlike GMRES and PCG. Because PCG has such a minimizing property and requires a fixed amount of work per iteration, we shall focus on this approach in this paper. We thus need to derive conditions for which PCG is an appropriate method.

Suppose that C is of rank l, and that we find a decomposition

$$C = EDE^T, \tag{2.1}$$

where E is m by l and D is l by l and invertible—either a spectral decomposition or an LDL^T factorization with pivoting are suitable—but the exact form is not relevant. In this case, on defining additional variables

$$z = -DE^T y,$$

we may rewrite (1.1) as

$$\begin{pmatrix} H & 0 & A^T \\ 0 & D^{-1} & E^T \\ A & E & 0 \end{pmatrix} \begin{pmatrix} x \\ z \\ y \end{pmatrix} = \begin{pmatrix} g \\ 0 \\ 0 \end{pmatrix}. \tag{2.2}$$

Noting the trailing zero block in the coefficient matrix of (2.2), we see that the required (x, z) components of the solution lie in the nullspace of $(A \;\; E)$.

Let the columns of the matrix

$$N = \begin{pmatrix} N_1 \\ N_2 \end{pmatrix}$$

form a basis for this null space. Then

$$\begin{pmatrix} x \\ z \end{pmatrix} = \begin{pmatrix} N_1 \\ N_2 \end{pmatrix} w \tag{2.3}$$

for some w, and (2.2) implies

$$H_N w = N_1^T g, \tag{2.4}$$

where

$$H_N \stackrel{\text{def}}{=} N_1^T H N_1 + N_2^T D^{-1} N_2. \tag{2.5}$$

Since we would like to apply PCG to solve (2.4), our fundamental assumption is then that

A1: the matrix H_N is positive definite.

Fortunately assumption **A1** is often easy to verify.

THEOREM 2.1. [illegible] M_H [illegible] m_{H-} [illegible] C [illegible] c_- [illegible] **A1** [illegible]

$$m_{H-} + c_- = m. \tag{2.6}$$

[illegible]. It is well known [29, Thm. 2.1] that under assumption **A1** the coefficient matrix E_H of (2.2) has inertia $(n + l, m, 0)$. The result then follows directly from Sylvester's law of inertia, since then $\text{In}(E_H) = \text{In}(D^{-1}) + \text{In}(M_H)$ and D^{-1} has as many negative eigenvalues as C has. □

Under assumption **A1**, we may apply the PCG method to find w, and hence recover (x, z) from (2.3). Notice that such an approach does not determine y, and additional calculations may need to be performed to recover it if it is required.

More importantly, it has been shown [8, 11, 31, 43] that rather than computing the iterates explicitly within the nullspace via (2.3), it is possible to perform the iteration in the original (x, z) space so long as the preconditioner is chosen carefully. Specifically, let G be any symmetric matrix for which

A2: the matrix

$$G_N \stackrel{\text{def}}{=} N_1^T G N_1 + N_2^T D^{-1} N_2 \tag{2.7}$$

is positive definite,

which we can check using Theorem 2.1. Then the appropriate projected preconditioned conjugate-gradient (PPCG) algorithm follows [31].

Projected Preconditioned Conjugate Gradients (variant 1):
Given $x = 0$, $z = 0$ and $h = 0$, solve

$$\begin{pmatrix} G & 0 & A^T \\ 0 & D^{-1} & E^T \\ A & E & 0 \end{pmatrix} \begin{pmatrix} r \\ d \\ u \end{pmatrix} = \begin{pmatrix} g \\ h \\ 0 \end{pmatrix}, \tag{2.8}$$

and set $(p, v) = -(r, d)$ and $\sigma = g^T r + h^T d$.
Iterate until convergence:

$$\begin{array}{lrcl} \text{Form } & Hp \text{ and } D^{-1}v. & & \\ \text{Set} & \alpha & = & \sigma/(p^T Hp + v^T D^{-1} v). \\ \text{Update} & x & \leftarrow & x + \alpha p, \\ & z & \leftarrow & z + \alpha v, \\ & g & \leftarrow & g + \alpha Hp \\ \text{and} & h & \leftarrow & h + \alpha D^{-1} v. \\ \text{Given } g \text{ and } h, & \text{solve (2.8) to find } r \text{ and } d. & & \\ \text{Set} & \sigma_{\text{new}} & = & g^T r + h^T d \\ \text{and} & \beta & = & \sigma_{\text{new}}/\sigma. \\ \text{Update} & \sigma & \leftarrow & \sigma_{\text{new}}, \\ & p & \leftarrow & -r + \beta p \\ \text{and} & v & \leftarrow & -d + \beta v. \end{array}$$

We note in passing that the algorithm above may be generalized by replacing D in the preconditioning step (2.8) by any nonsingular T for which $N_1^T G N_1 + N_2^T T^{-1} N_2$ is positive definite. The scalar σ gives an appropriate optimality measure [31], and a realistic termination rule is to stop when σ is small relative to its original value.

While this method is acceptable when a decomposition (2.1) of C is known, it is preferable to be able to work directly with C. To this end, suppose that at each iteration

$$h = -E^T a, \quad v = -DE^T q \text{ and } d = -DE^T t$$

for unknown vectors a, q and t—this is clearly the case at the start of the algorithm. Then, letting $w = Ca$, it is straightforward to show that $t = u + a$, and that we can replace our previous algorithm with the following equivalent one.

Projected Preconditioned Conjugate Gradients (variant 2):
Given $x = 0$, and $a = w = 0$, solve

$$\begin{pmatrix} G & A^T \\ A & -C \end{pmatrix} \begin{pmatrix} r \\ u \end{pmatrix} = \begin{pmatrix} g \\ w \end{pmatrix}, \tag{2.9}$$

and set $p = -r$, $q = -u$ and $\sigma = g^T r$.
Iterate until convergence:

Form Hp and Cq.
Set $\alpha = \sigma/(p^T Hp + q^T Cq)$.
Update $x \leftarrow x + \alpha p$,
$a \leftarrow a + \alpha q$,
$g \leftarrow g + \alpha Hp$
and $w \leftarrow w + \alpha Cq$.
Given g and w, solve (2.9) to find r and u.
Set $t = a + u$,
$\sigma_{\text{new}} = g^T r + t^T w$
and $\beta = \sigma_{\text{new}}/\sigma$.
Update $\sigma \leftarrow \sigma_{\text{new}}$,
$p \leftarrow -r + \beta p$
and $q \leftarrow -t + \beta q$.

Notice now that z no longer appears, and that the preconditioning is carried out using the matrix M_G mentioned in the introduction. Also note that although this variant involves two more vectors than its predecessor, t is simply used as temporary storage and may be omitted if necessary, while w may also be replaced by Ca if storage is tight.

When $C = 0$, this is essentially the algorithm given by [31], but for this case the updates for v and w are unnecessary and may be discarded. At the other extreme, when C is nonsingular the algorithm is precisely that proposed by [30, Alg. 2.3], and is equivalent to applying PCG to the system

$$(H + A^T C^{-1} A)x = g,$$

using a preconditioner of the form $G + A^T C^{-1} A$.

Which of the two variants is preferable depends on whether we have a decomposition (2.1) and whether l is small relative to m: the vectors h and v in the first variant are of length l, while the corresponding a and q in the second are of length m. Notice also that although the preconditioning steps in the first variant require that we solve (2.8) this is entirely equivalent to solving (2.9), where $w = -EDh$, and recovering

$$d = D(h - E^T v).$$

Thus our remaining task is to consider how to build suitable and effective preconditioners of the form (1.2). We recall that it is the distribution of the generalized eigenvalues λ for which

$$H_N \bar{v} = \lambda G_N \bar{v} \tag{2.10}$$

that determines the convergence of the preceding PPCG algorithms, and thus we will be particularly interested in preconditioners which cluster these eigenvalues. In particular, if we can efficiently compute G_N so that there are few distinct eigenvalues λ in (2.10), then PPCG convergence (termination) will be rapid.

3. Eigenvalue considerations. We first consider the spectral implications of preconditioning (1.1) by (1.2).

THEOREM 3.1 (see [16, Thm. 3.1] or, in special circumstances, [4, 44]). [illegible] M_H [illegible] (1.1) [illegible] $M_G^{-1} M_H$ [illegible] m [illegible] n [illegible]

$$(H - \lambda G)v = (\lambda - 1)A^T w, \quad \text{[illegible]} \quad Av - Cw = 0.$$

[illegible] C [illegible]

$$(H + A^T C^{-1} A)v = \lambda(G + A^T C^{-1} A)v. \tag{3.1}$$

Our goal in this section is to improve upon this result in the general case $C \neq 0$. For the special case in which $C = 0$, results are already known [18] concerning the eigenvalues of $K_G^{-1} K_H$ for the pair of matrices

$$K_H = \begin{pmatrix} H & A^T \\ A & 0 \end{pmatrix} \text{ and } K_G = \begin{pmatrix} G & A^T \\ A & 0 \end{pmatrix}.$$

These results refer to a partitioning of A as

$$A = (A_1 \;\; A_2), \tag{3.2}$$

so that its leading m by m submatrix

A3: A_1 is nonsingular;

and a similar partitioning of G and H as

$$G = \begin{pmatrix} G_{11} & G_{21}^T \\ G_{21} & G_{22} \end{pmatrix} \text{ and } H = \begin{pmatrix} H_{11} & H_{21}^T \\ H_{21} & H_{22} \end{pmatrix}, \tag{3.3}$$

where G_{11} and H_{11} are, respectively, the leading m by m submatrices of G and H.

In practice, the partitioning of A to ensure **A3** may involve column permutations, but without loss of generality we simply assume here that any required permutations have already been carried out. Given **A3**, we shall be particularly concerned with the [illegible] basis matrix

$$N = \begin{pmatrix} R \\ I \end{pmatrix}, \text{ where } R = -A_1^{-1} A_2. \tag{3.4}$$

Such basis matrices play vital roles in simplex (pivoting)-type methods for linear programming [2, 24], and more generally in active-set methods for nonlinear optimization [27, 38, 39].

THEOREM 3.2 (see [18, Thm. 2.3]). [illegible] G [illegible] H [illegible] (3.3) [illegible] **A3** [illegible]

$$G_{22} = H_{22}, \text{ [illegible] } G_{11} = 0 \text{ [illegible] } G_{21} = 0. \tag{3.5}$$

[illegible] H_{22} [illegible]

$$\begin{aligned} \rho \stackrel{\text{def}}{=} \; & \min\left[\operatorname{rank}(A_2), \operatorname{rank}(H_{21})\right] \\ & + \min\left[\operatorname{rank}(A_2), \operatorname{rank}(H_{21}) + \min[\operatorname{rank}(A_2), \operatorname{rank}(H_{11})]\right]. \end{aligned}$$

[illegible] $K_G^{-1} K_H$ [illegible]

$$\operatorname{rank}(R^T H_{21}^T + H_{21} R + R^T H_{11} R) + 1 \leq \min(\rho, n - m) + 1 \leq \min(2m, n - m) + 1$$

[illegible]

The restriction that H_{22} be positive definite is not as severe as it might first seem because the problem may be reformulated to use $H_{22} + A_2^T \Delta A_2$ for any symmetric

positive definite weight matrix Δ instead [18, Thm. 2.2]—this corresponds to the so-called augmented Lagrangian approach [33].

THEOREM 3.3 (see [18, Thm. 2.4]). ... G ... H ... (3.3) ... **A3** ...

$$G_{22} = H_{22} \quad \text{...} \quad G_{11} = H_{11} \quad \text{...} \quad G_{21} = 0. \tag{3.6}$$

... $H_{22} + R^T H_{11}^T R$ ...

$$\nu \stackrel{\text{def}}{=} 2\min\big[\operatorname{rank}(A_2), \operatorname{rank}(H_{21})\big].$$

... $K_G^{-1} K_H$ ...

$$\operatorname{rank}(R^T H_{21}^T + H_{21} R) + 1 \le \nu + 1 \le \min(2m, n-m) + 1$$

...

THEOREM 3.4 (see [18, Thm. 2.5]). ... G ... H ... (3.3) ... **A3** ...

$$G_{22} = H_{22} \quad \text{...} \quad G_{21} = H_{21} \quad \text{...} \quad G_{11} = 0. \tag{3.7}$$

... $H_{22} + R^T H_{21}^T + H_{21} R$ ...

$$\mu \stackrel{\text{def}}{=} \min\big[\operatorname{rank}(A_2), \operatorname{rank}(H_{11})\big].$$

... $K_G^{-1} K_H$ ...

$$\operatorname{rank}(R^T H_{11} R) + 1 \le \mu + 1 \le \min(m, n-m) + 1$$

...

Turning to the general case of $C \neq 0$, denote the coefficient matrices of the systems (2.2) and (2.8) by

$$\bar{K}_H \stackrel{\text{def}}{=} \begin{pmatrix} H & 0 & A^T \\ 0 & D^{-1} & E^T \\ A & E & 0 \end{pmatrix} \quad \text{and} \quad \bar{K}_G \stackrel{\text{def}}{=} \begin{pmatrix} G & 0 & A^T \\ 0 & D^{-1} & E^T \\ A & E & 0 \end{pmatrix},$$

respectively. Recalling the definitions (2.5) and (2.7) of H_N and G_N, the following result is a direct consequence of [35, Thm. 2.1].

THEOREM 3.5. ... N ... n ... $n+l-m$ ... $(A \;\; E)$ ... $\bar{K}_G^{-1} \bar{K}_H$ ... $2m$ ... $n+l-m$ ... (2.10)

We may improve on Theorem 3.5 by applying Theorems 3.2–3.4 in our more general setting. To do so, let

$$\bar{R} = -A_1^{-1}(A_2 \;\; E),$$

and note that (3.2) implies the partitioning

$$\bar{K}_H = \left(\begin{array}{c|cc|c} H_{11} & H_{21}^T & 0 & A_1^T \\ \hline H_{21} & H_{22}^T & 0 & A_2^T \\ 0 & 0 & D^{-1} & E^T \\ \hline A_1 & A_2 & E & 0 \end{array}\right) \text{ and } \bar{K}_G = \left(\begin{array}{c|cc|c} G_{11} & G_{21}^T & 0 & A_1^T \\ \hline G_{21} & G_{22}^T & 0 & A_2^T \\ 0 & 0 & D^{-1} & E^T \\ \hline A_1 & A_2 & E & 0 \end{array}\right).$$

We then have the following immediate consequences.

COROLLARY 3.6. ... G ... H ... (3.3) ... (3.5) ... **A3** ...

$$\begin{pmatrix} H_{22} & 0 \\ 0 & D^{-1} \end{pmatrix} \tag{3.8}$$

...

$$\bar{\rho} = \min\left[\eta, \operatorname{rank}(H_{21})\right] + \min\left[\eta, \operatorname{rank}(H_{21}) + \min[\eta, \operatorname{rank}(H_{11})]\right],$$

... $\eta = \operatorname{rank}(A_2 \;\; E)$ ... $\bar{K}_G^{-1}\bar{K}_H$ ...

$$\operatorname{rank}(\bar{R}^T H_{21}^T + H_{21}\bar{R} + \bar{R}^T H_{11}\bar{R}) + 1 \leq \min(\bar{\rho}, n + l - m) + 1 \leq \min(2m, n + l - m) + 1$$

...

COROLLARY 3.7. ... G ... H ... (3.3) ... (3.6) ... **A3** ...

$$\begin{pmatrix} H_{22} & 0 \\ 0 & D^{-1} \end{pmatrix} + \bar{R}^T H_{11}^T \bar{R} \tag{3.9}$$

...

$$\bar{\nu} = 2\min\left[\eta, \operatorname{rank}(H_{21})\right],$$

... $\eta = \operatorname{rank}(A_2 \;\; E)$ ... $\bar{K}_G^{-1}\bar{K}_H$ ...

$$\operatorname{rank}(\bar{R}^T H_{21}^T + H_{21}\bar{R}) + 1 \leq \bar{\nu} + 1 \leq \min(2m, n + l - m) + 1$$

...

COROLLARY 3.8. ... G ... H ... (3.3) ... (3.7) ... **A3** ...

$$\begin{pmatrix} H_{22} & 0 \\ 0 & D^{-1} \end{pmatrix} + \bar{R}^T H_{21}^T + H_{21}\bar{R} \tag{3.10}$$

...

$$\bar{\mu} = \min\left[\eta, \operatorname{rank}(H_{11})\right],$$

... $\eta = \operatorname{rank}(A_2 \;\; E)$ ... $\bar{K}_G^{-1}\bar{K}_H$ ...

$$\operatorname{rank}(\bar{R}^T H_{11}\bar{R}) + 1 \leq \bar{\mu} + 1 \leq \min(m, n + l - m) + 1$$

...

While the requirements that (3.8)–(3.10) be positive definite may at first seem strong assumptions, as in the case $C = 0$ we can also apply a so-called augmented Lagrangian approach for the general case $C \neq 0$.

THEOREM 3.9. ... (2.6) ... H ... $\bar{\Delta}$ ...

$$\begin{pmatrix} H & 0 \\ 0 & D^{-1} \end{pmatrix} + \begin{pmatrix} A^T \\ E^T \end{pmatrix} \Delta (A \;\; E)$$

[illegible] Δ [illegible] $\Delta - \bar{\Delta}$ [illegible]

(2.6) [illegible] $H + A^T\Delta A$ [illegible] $E^T\Delta E + D^{-1}$ [illegible] Δ [illegible]. This follows immediately by applying [18, Thm. 2.2] to $\bar{K}_H$. □

Because $Ax + Ez = 0$ we may rewrite (2.2) as the equivalent system

$$\begin{pmatrix} H + A^T\Delta A & A^T\Delta E & A^T \\ E^T\Delta A & E^T\Delta E + D^{-1} & E^T \\ A & E & 0 \end{pmatrix} \begin{pmatrix} x \\ z \\ y \end{pmatrix} = \begin{pmatrix} g \\ 0 \\ 0 \end{pmatrix}.$$

Eliminating the variable z, we find that

$$\begin{pmatrix} H + A^T\Delta A & A^T P^T \\ PA & -W \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = -\begin{pmatrix} g \\ 0 \end{pmatrix},$$

where

$$P = I - \Delta W \ \text{ and } \ W = E(E^T\Delta E + D^{-1})^{-1}E^T.$$

Hence

$$\begin{pmatrix} H + A^T\Delta A & A^T \\ A & -\bar{C} \end{pmatrix} \begin{pmatrix} x \\ \bar{y} \end{pmatrix} = -\begin{pmatrix} g \\ 0 \end{pmatrix}, \tag{3.11}$$

where

$$\bar{C} = P^{-1}WP^{-T} = (I - \Delta W)^{-1}W(I - W\Delta)^{-1} \ \text{ and } \ \bar{y} = P^T y. \tag{3.12}$$

Thus it follows from Theorem 3.9 that we may rewrite (2.2) so that its trailing and leading diagonal blocks are, respectively, negative semi- and positive definite. In doing so, any underlying structure (such as sparsity) may be compromised. For the sparse case, if we are prepared to tolerate fill-in in these blocks, requirements (3.8)–(3.10) then seem more reasonable.

Although (3.12) may appear complicated for general C, $\bar{C}$ is diagonal whenever C is. More generally, if $E = I$, $\bar{C} = D + D\Delta D$ and we may recover $y = (I + \Delta D)\bar{y}$.

4. Suitable preconditioners. It has long been common practice (at least in optimization circles) [4, 7, 13, 26, 36, 49] to use [illegible] preconditioners of the form (1.2) by specifying G and factorizing M_G using a suitable symmetric, indefinite package such as `MA27` [21] or `MA57` [20]. Given G, an alternative used commonly by the PDE community (see, for example, [1, 22, 36, 42, 44, 14, 48] and the many references in [3]) is to use the explicit block decomposition

$$M_G = \begin{pmatrix} I & 0 \\ AG^{-1} & I \end{pmatrix} \begin{pmatrix} G & 0 \\ 0 & -C - AG^{-1}A^T \end{pmatrix} \begin{pmatrix} I & G^{-1}A^T \\ 0 & I \end{pmatrix} \tag{4.1}$$

to solve (1.2) via factorizations of G and the Schur complement $C + AG^{-1}A^T$ (or an approximation to the latter) if these are viable. While such techniques for choosing G have often been successful, they have usually been rather ad hoc, with little attempt to improve upon the eigenvalue distributions beyond those suggested by Theorem 3.1. In this section we investigate an implicit-factorization alternative.

4.1. Implicit-factorization preconditioners. Recently Dollar and Wathen [19] proposed a class of incomplete factorizations for saddle-point problems ($C = 0$), based upon earlier work by Schilders [46]. They consider preconditioners of the form

$$M_G = PBP^T, \tag{4.2}$$

where solutions with each of the matrices P, B and P^T are easily obtained. In particular, rather than obtaining P and B from a given M_G, M_G [illegible] P [illegible] B In this section, we examine a broad class of methods of this form.

In order for the methods we propose to be effective, we shall require that **A3** holds. Since there is considerable flexibility in choosing the "basis" A_1 from the rectangular matrix A by suitable column interchanges, **A3** is often easily, and sometimes trivially, satisfied. Even though, theoretically, there is a lot of choice, the actual A_1 that is used for practical computation can have a significant effect on the overall effectiveness of the preconditioning strategies described in this paper. The problem of determining the "sparsest" A_1 is NP hard, [9, 10], while numerical considerations must be given to ensure that A_1 is not badly conditioned if at all possible [27]. More generally, we do not necessarily assume that A_1 is sparse or structured nor that it has a sparse (or other) factorization, merely that there are effective ways to solve systems involving A_1 and A_1^T. For example, for many problems involving constraints arising from the discretization of partial differential equations, there are highly effective [illegible] methods for such systems [5].

Suppose that

$$P = \begin{pmatrix} P_{11} & P_{12} & A_1^T \\ P_{21} & P_{22} & A_2^T \\ P_{31} & P_{32} & P_{33} \end{pmatrix} \text{ and } B = \begin{pmatrix} B_{11} & B_{21}^T & B_{31}^T \\ B_{21} & B_{22} & B_{32}^T \\ B_{31} & B_{32} & B_{33} \end{pmatrix}. \tag{4.3}$$

Our goal is to ensure that

$$(M_G)_{31} = A_1, \tag{4.4a}$$

$$(M_G)_{32} = A_2 \tag{4.4b}$$

$$\text{and} \quad (M_G)_{33} = -C, \tag{4.4c}$$

whenever $M_G = PBP^T$. Pragmatically, though, we are only interested in the case where one of the three possibilities

$$P_{11} = 0, \ P_{12} = 0 \text{ and } P_{32} = 0, \tag{4.5a}$$

$$\text{or } P_{11} = 0, \ P_{12} = 0 \text{ and } P_{21} = 0, \tag{4.5b}$$

$$\text{or } P_{12} = 0, \ P_{32} = 0 \text{ and } P_{33} = 0 \tag{4.5c}$$

(as well as nonsingular P_{31} and P_{22}) hold, since only then will P be easily block-invertible. Likewise, we restrict ourselves to the three general cases

(4.6a) $B_{21} = 0, \ B_{31} = 0$ and $B_{32} = 0$ with easily invertible B_{11}, B_{22} and B_{33},

(4.6b) $B_{32} = 0$ and $B_{33} = 0$ with easily invertible B_{31} and B_{22}, or

(4.6c) $B_{11} = 0$ and $B_{21} = 0$ with easily invertible B_{31} and B_{22},

so that B is block-invertible. B is also easily block-invertible if

$$B_{21} = 0 \text{ and } B_{32} = 0 \text{ with easily invertible } \begin{pmatrix} B_{11} & B_{31}^T \\ B_{31} & B_{33} \end{pmatrix} \text{ and } B_{22}, \tag{4.7}$$

and we will also consider this possibility.

TABLE 4.1

Possible implicit factors for the preconditioner (1.2). *We give the P and B factors and any necessary restrictions on their entries. We also associate a family number with each class of implicit factors. Full derivations are given in* [17, Appendix A].

Family/ reference	P	B	Conditions
1.	$\begin{pmatrix} 0 & 0 & A_1^T \\ 0 & P_{22} & A_2^T \\ P_{31} & 0 & P_{33} \end{pmatrix}$	$\begin{pmatrix} B_{11} & 0 & 0 \\ 0 & B_{22} & 0 \\ 0 & 0 & B_{33} \end{pmatrix}$	$B_{11} = -P_{31}^{-1}(C + P_{33})P_{31}^{-T}$ $B_{33} = P_{33}^{-1}$
2.	$\begin{pmatrix} 0 & 0 & A_1^T \\ 0 & P_{22} & A_2^T \\ P_{31} & 0 & P_{33} \end{pmatrix}$	$\begin{pmatrix} B_{11} & 0 & B_{31}^T \\ 0 & B_{22} & 0 \\ B_{31} & 0 & 0 \end{pmatrix}$	$P_{31} = B_{31}^{-T}$ $P_{33} + P_{33}^T + P_{31}B_{11}P_{31}^T = -C$
3.	$\begin{pmatrix} 0 & 0 & A_1^T \\ P_{21} & P_{22} & A_2^T \\ P_{31} & 0 & -C \end{pmatrix}$	$\begin{pmatrix} B_{11} & 0 & B_{31}^T \\ 0 & B_{22} & 0 \\ B_{31} & 0 & 0 \end{pmatrix}$	$B_{31} = P_{31}^{-T}$ $B_{11} = P_{31}^{-1}CP_{31}^{-T}$
4.	$\begin{pmatrix} 0 & 0 & A_1^T \\ P_{21} & P_{22} & A_2^T \\ P_{31} & 0 & P_{33} \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & B_{31}^T \\ 0 & B_{22} & B_{32}^T \\ B_{31} & B_{32} & 0 \end{pmatrix}$	$P_{21} = -P_{22}B_{32}^TB_{31}^{-T}$ $P_{31} = B_{31}^{-T}$ $P_{33} + P_{33}^T = -C$
5.	$\begin{pmatrix} 0 & 0 & A_1^T \\ P_{21} & P_{22} & A_2^T \\ P_{31} & 0 & P_{33} \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & B_{31}^T \\ 0 & B_{22} & B_{32}^T \\ B_{31} & B_{32} & B_{33} \end{pmatrix}$	$-C = P_{33} + P_{33}^T - P_{33}B_{33}P_{33}^T$ $B_{31} = (I - B_{33}P_{33}^T)P_{31}^{-T}$ $B_{32} = -B_{31}P_{21}^TP_{22}^{-T}$
6.	$\begin{pmatrix} 0 & 0 & A_1^T \\ 0 & P_{22} & A_2^T \\ P_{31} & P_{32} & P_{33} \end{pmatrix}$	$\begin{pmatrix} B_{11} & B_{21}^T & B_{31}^T \\ B_{21} & B_{22} & 0 \\ B_{31} & 0 & 0 \end{pmatrix}$	$P_{31} = B_{31}^{-T}$ $P_{32} = -P_{31}B_{21}^TB_{22}^{-1}$ $P_{33} + P_{33}^T$ $= -C - P_{31}(B_{11} - B_{21}^TB_{22}^{-1}B_{21})P_{31}^T$
7.	$\begin{pmatrix} 0 & 0 & A_1^T \\ 0 & P_{22} & A_2^T \\ P_{31} & P_{32} & P_{33} \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & B_{31}^T \\ 0 & B_{22} & B_{32}^T \\ B_{31} & B_{32} & B_{33} \end{pmatrix}$	$P_{33} + P_{33}^T + P_{33}(B_{33} - B_{32}B_{22}^{-1}B_{32}^T)P_{33}^T$ $= -C$ $P_{32} = -P_{33}B_{32}B_{22}^{-1}$ $P_{31} = (I - P_{32}B_{32}^T - P_{33}B_{33}^T)B_{31}^{-T}$

We consider all of these possibilities in detail in [17, Appendix A], and summarize our findings in Tables 4.1 and 4.2. We have identified eleven possible classes of easily invertible factors that are capable of reproducing the A and C blocks of M_G, a further two which may be useful when C is diagonal, and one that is only applicable if $C = 0$.

Notice that aside from invertibility, there are , , , restrictions on P_{22} and B_{22}.

4.2. Reproducing H. Having described families of preconditioners which are capable of reproducing the required components A and C of M_G, we now examine what form the resulting G takes. In particular, we consider which submatrices of G can be defined to completely reproduce the associated submatrix of H; we say that a

TABLE 4.2
Possible implicit factors for the preconditioner (1.2) (*cont.*). *We give the P and B factors and any necessary restrictions on their entries. We also associate a family number with each class of implicit factors. Full derivations are given in* [17, Appendix A].

Family/ reference	P	B	Conditions
8.	$\begin{pmatrix} A_1^T & 0 & A_1^T \\ A_2^T & P_{22} & A_2^T \\ -C & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} -C^{-1} & 0 & 0 \\ 0 & B_{22} & 0 \\ 0 & 0 & B_{33} \end{pmatrix}$	C invertible
9.	$\begin{pmatrix} P_{11} & 0 & A_1^T \\ P_{21} & P_{22} & A_2^T \\ P_{31} & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} B_{11} & B_{21}^T & B_{31}^T \\ B_{21} & B_{22} & 0 \\ B_{31} & 0 & 0 \end{pmatrix}$	$B_{11} = -P_{31}^{-1} C P_{31}^{-T}$ $B_{31} = P_{31}^{-T} - M B_{11}$ $B_{21} = P_{22}^{-1}(P_{21} - A_2^T M) B_{11}$ $P_{11} = A_1^T M$ for some invertible M
10.	$\begin{pmatrix} P_{11} & 0 & A_1^T \\ P_{21} & P_{22} & A_2^T \\ P_{31} & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & B_{31}^T \\ 0 & B_{22} & B_{32}^T \\ B_{31} & B_{32} & B_{33} \end{pmatrix}$	$C = 0$ $P_{31} = B_{31}^{-T}$
11.	$\begin{pmatrix} 0 & 0 & A_1^T \\ P_{21} & P_{22} & A_2^T \\ P_{31} & 0 & -C \end{pmatrix}$	$\begin{pmatrix} B_{11} & 0 & B_{31}^T \\ 0 & B_{22} & 0 \\ B_{31} & 0 & B_{33} \end{pmatrix}$	C invertible $P_{31}^T = B_{11}^{-1} B_{31}^T C$ $B_{33} = (B_{31} P_{31}^T - I) C^{-1}$
12.	$\begin{pmatrix} 0 & 0 & A_1^T \\ P_{21} & P_{22} & A_2^T \\ P_{31} & 0 & -C \end{pmatrix}$	$\begin{pmatrix} B_{11} & 0 & B_{31}^T \\ 0 & B_{22} & 0 \\ B_{31} & 0 & B_{33} \end{pmatrix}$	$B_{11} = P_{31}^{-1} C P_{31}^{-T}$ $B_{31} = P_{31}^{-T}$, where $B_{33} C = 0$
13.	$\begin{pmatrix} 0 & 0 & A_1^T \\ 0 & P_{22} & A_2^T \\ P_{31} & 0 & P_{33} \end{pmatrix}$	$\begin{pmatrix} B_{11} & 0 & B_{31}^T \\ 0 & B_{22} & 0 \\ B_{31} & 0 & B_{33} \end{pmatrix}$	$P_{31} = (I - P_{33} B_{33}) B_{31}^{-T}$ $B_{11} = P_{31}^{-1} \left(P_{33} B_{33} P_{33}^T - C - P_{33} - P_{33}^T \right) P_{31}^{-T}$
14.	$\begin{pmatrix} P_{11} & 0 & A_1^T \\ P_{21} & P_{22} & A_2^T \\ P_{31} & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} B_{11} & 0 & B_{31}^T \\ 0 & B_{22} & 0 \\ B_{31} & 0 & B_{33} \end{pmatrix}$	$B_{11} = -P_{31}^{-1} C P_{31}^{-T}$ $B_{31} = P_{31}^{-T} - M B_{11}$ $P_{11} = A_1^T M$ $P_{21} = A_2^T M$ for some invertible M

component G_{ij}, $i, j \in \{1, 2\}$, is [illegible] if it is possible to choose it so that $G_{ij} = H_{ij}$. We give the details in [17, Appendix B], and summarize our findings for each of the 14 families from section 4.1 in Table 4.3.

Some of the submatrices in the factors P and B can be arbitrarily chosen without changing the completeness of the family. We shall call these "free blocks." For example, consider family 2 from Table 4.1. The matrix G produced by this family always satisfies $G_{11} = 0$, $G_{21} = 0$, and $G_{22} = P_{22} B_{22} P_{22}^T$. Hence, P_{22} can be defined as any nonsingular matrix of suitable dimension, and B_{22}^T can be subsequently chosen so that $G_{22} = H_{22}$. The simplest choice for P_{22} is the identity matrix. We observe that

TABLE 4.3

Blocks of G for the families of preconditioners given in Tables 4.1 and 4.2. The superscript 1 indicates that the value of G_{21} is dependent on the choice of G_{11}. If G_{ij}, $i,j \in \{1,2\}$, is a zero matrix, then a superscript 2 is used. The superscript 3 means that G_{21} is dependent on the choice of G_{11} when $C = 0$, but complete otherwise, while the superscript 4 indicates that G_{11} is only guaranteed to be complete when $C = 0$.

Family	Completeness G_{11}	G_{21}	G_{22}	Conditions on C	Feasible to use	Comments
1.		$\times^1$		any C		
2.	$\times^2$	$\times^2$		any C		
3.	$\times^2$			any C		
4.	$\times^2$	$\times^2$		any C		Simplest choice of free blocks is the same as that for family 2.
5.		$\times^1$		any C	$C = 0$	
6.	$\times^2$	$\times^2$		any C		Simplest choice of free blocks is the same as that for family 2.
7.		^{3}		any C	$C = 0$	If $C = 0$ using simplest choice of free blocks, then same as that for family 5 with $C = 0$.
8.		$\times^1$		nonsingular		
9.				any C	$C = 0$	
10.				$C = 0$		Generalization of factorization suggested by Schilders [19, 46]; See also [37].
11.				nonsingular		
12.	^{4}			any C	diagonal C	$C = 0$ gives example of family 10. C nonsingular gives family 3.
13.		$\times^1$		any C		
14.		$\times^1$		any C		$C = 0$ gives example of family 10.

the choice of the remaining submatrices in P and B will not affect the completeness of the factorization, and are only required to satisfy the conditions given in Table 4.1. The simplest choices for these submatrices will be $P_{31} = I$, and $B_{11} = 0$, giving $P_{33} = -\frac{1}{2}C$, and $B_{31} = I$. Using these simple choices we obtain:

$$P = \begin{pmatrix} 0 & 0 & A_1^T \\ 0 & I & A_2^T \\ I & 0 & -\frac{1}{2}C \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 & 0 & I \\ 0 & B_{22} & 0 \\ I & 0 & 0 \end{pmatrix}.$$

The simplest choice of the free blocks may result in some of the families having the same factors as other families. This is indicated in the "comments" column of the table. Table 4.3 also gives the conditions that C must satisfy to use the family, and whether the family is feasible to use, i.e., are the conditions on the blocks given in Tables 4.1 and 4.2 easily satisfied?

Table 4.4 gives some guidance towards which families from Tables 4.1 and 4.2 should be used in the various cases of G given in section 3. We also suggest simple choices for the free blocks. In our view, although Table 4.3 indicates that it is theoret-

TABLE 4.4
Guidance towards which family to use to generate the various choices of G given in section 3.

Sub-blocks of G	Conditions on C	Family	Free block choices
$G_{22} = H_{22}$, $G_{11} = 0$, $G_{21} = 0$	any C	2	$P_{22} = I$, $P_{31} = I$, $B_{11} = 0$
$G_{22} = H_{22}$, $G_{11} = H_{11}$, $G_{21} = 0$	$C = 0$	10	$B_{21} = 0$, $P_{22} = I$, $P_{31} = I$
$G_{22} = H_{22}$, $G_{11} = H_{11}$, $G_{21} = 0$	C nonsingular	11	$P_{22} = I$, $P_{31} = I$
$G_{22} = H_{22}$, $G_{21} = H_{21}$, $G_{11} = 0$	any C	3	$P_{22} = I$, $P_{31} = I$

ically possible to reproduce all of H using, e.g., family 9, in practice this is unviable because structure, such as sparsity, could be severely compromised.

5. Numerical examples. In this section we examine how effective implicit-factorization preconditioners might be when compared with explicit-factorization ones. We consider problems generated using the complete set of quadratic programming examples from the CUTEr [32] test set used in our previous experiments for the $C = 0$ case [18]. All inequality constraints are converted to equations by adding slack variables, and a suitable "barrier" penalty term is added to the diagonal of the Hessian for each bounded or slack variable to simulate systems that might arise during an iteration of an interior-point method for such problems; in each of the test problems the value 1.1 is used—this sort of value would correspond to an intermediate stage of the outer (optimization) iteration. The resulting equality-constrained quadratic programs are then of the form

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad g^T x + \tfrac{1}{2} x^T H x \ \text{ subject to } \ Ax = 0. \tag{5.1}$$

Given this data H and A, two illustrative choices of diagonal C are considered, namely,

$$c_{ii} = 1 \ \text{ for } \ 1 \le i \le m, \tag{5.2}$$

and

$$c_{ii} = \begin{cases} 0 & \text{for } \ 1 \le i \le \lceil \frac{m}{2} \rceil \\ 1 & \text{for } \ \lceil \frac{m}{2} \rceil + 1 \le i \le m; \end{cases} \tag{5.3}$$

in practice such C may be thought of as regularization terms for some or all on the constraints in (5.1). Our aim is thus to solve for the primal variables x in the system (1.1) using a suitably preconditioned PPCG iteration.

Rather than present large tables of data (these can be found in [17, Appendix C]), here we use performance profiles [15] to illustrate our results. To explain the idea, let $\mathcal{P}$ represent the set of preconditioners that we wish to compare. Suppose that the run of PPCG using a given preconditioner $i \in \mathcal{P}$ reports the total CPU time $t_{ij} \ge 0$ when executed on example j from the test set $\mathcal{T}$. For all problems $j \in \mathcal{T}$, we want to compare the performance of algorithm i with the performance of the fastest algorithm in the set $\mathcal{P}$. For $j \in \mathcal{T}$, let $t_j^{\text{MIN}} = \min\{t_{ij}; i \in \mathcal{P}\}$. Then for $\alpha \ge 1$ and each $i \in \mathcal{P}$ we define

$$k(t_{ij}, t_j^{\text{MIN}}, \alpha) = \begin{cases} 1 & \text{if } \ t_{ij} \le \alpha t_j^{\text{MIN}} \\ 0 & \text{otherwise.} \end{cases}$$

The performance profile [15] of algorithm i is then given by the function

$$p_i(\alpha) = \frac{\sum_{j \in \mathcal{T}} k(t_{ij}, t_j^{\text{MIN}}, \alpha)}{|\mathcal{T}|}, \quad \alpha \geq 1.$$

Thus $p_i(1)$ gives the fraction of the examples for which algorithm i is the most effective (according to the statistic t_{ij}), $p_i(2)$ gives the fraction for which algorithm i is within a factor of 2 of the best, and $\lim_{\alpha \to \infty} p_i(\alpha)$ gives the fraction for which the algorithm succeeded.

We consider two explicit factorization preconditioners, one using exact factors ($G = H$), and the other using a simple projection ($G = I$). A Matlab interface to the HSL [34] package `MA57` [20] (version 2.2.1) is used to factorize M_G and subsequently solve (1.4); as we have already mentioned, in some cases it would have been both possible and preferable to use instead the explicit block decomposition (4.1) when $G = I$ (or for easily invertible H), and interpretation of the results we present should keep this in mind. Three implicit factorizations of the form (4.2) with factors (4.3) are also considered. The first is from family 1 (Table 4.1), and aims for simplicity by choosing $P_{31} = I$, $P_{33} = I = B_{33}$ and $B_{22} = I = P_{22}$, and this leads $B_{11} = -(C + I)$; such a choice does not necessarily reproduce any of H, but is inexpensive to use. The remaining implicit factorizations are from family 2 (Table 4.1). The former (marked (a) in the following figures) selects $G_{22} = H_{22}$ while the latter (marked (b) in the figures) chooses $G_{22} = I$; for simplicity we chose $P_{31} = I = B_{31}$, $B_{11} = 0$, $P_{22} = I$ and $P_{33} = -\frac{1}{2}C$ (see section 4.2), and thus we merely require that $B_{22} = H_{22}$ for case (a) and $B_{22} = I$ for case (b)—we use `MA57` to factorize H_{22} in the former case.

Given A, a suitable basis matrix A_1 is found by finding a sparse LU factorization of A^T using the built-in Matlab function `lu`. An attempt to correctly identify the rank is controlled by tight threshold column pivoting, in which any pivot may not be smaller than a factor $\tau = 2$ of the largest entry in its (uneliminated) column [27, 28]. The rank is estimated as the number of pivots, $\rho(A)$, completed before the remaining uneliminated submatrix is judged to be numerically zero, and the indices of the $\rho(A)$ pivotal rows and columns of A define A_1—if $\rho(A) < m$, the remaining rows of A are judged to be dependent, and are discarded. Although such a strategy may not be as robust as, say, a singular-value decomposition or a QR factorization with pivoting, both our and others' experience [27] indicate it to be remarkably reliable and successful in practice. Having found A_1, the factors are discarded, and a fresh LU decomposition of A_1, with a looser threshold column pivoting factor $\tau = 100$, is computed using `lu` in order to try to encourage sparse factors.

All of our experiments were performed using a dual processor Intel Xeon 3.2GHz Workstation with hyperthreading and 2 Gbytes of RAM. Our codes were written and executed in Matlab 7.0 Service Pack 1.

In Figures 5.1–5.2, (see the tables in [17, Appendix C] for the raw data), we compare our five preconditioning strategies for (approximately) solving the problem (1.1) when C is given by (5.2) using the PPCG scheme (variant 2) described in section 2. We consider both low and high(er) accuracy solutions. For the former, we terminate as soon as the residual σ has been reduced more than 10^{-2} from its original value, while the latter requires a 10^{-8} reduction; these are intended to simulate the levels of accuracy that might be required within a nonlinear equation or optimization solver in early (global) and later (asymptotic) phases of the solution process.

We see that if low accuracy solutions suffice, the implicit factorizations appear to be significantly more effective at reducing the residual than their explicit counter-

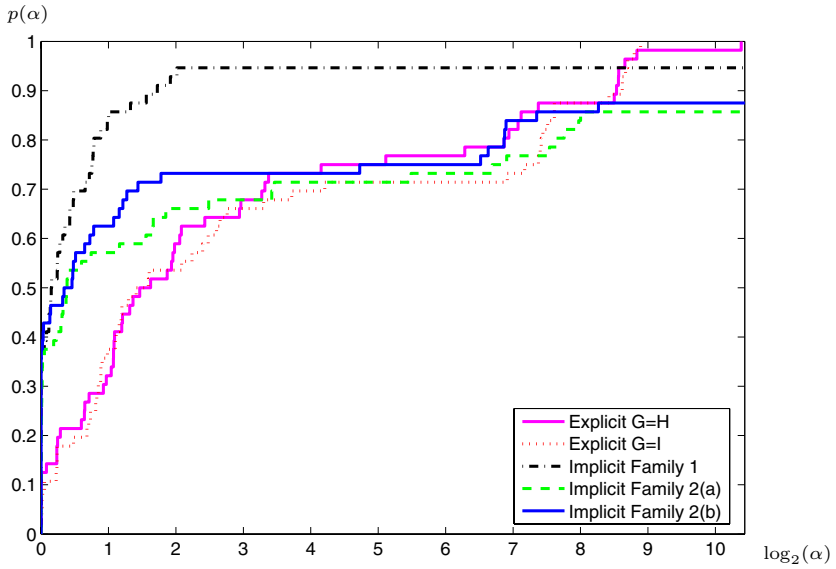


FIG. 5.1. *Performance profile, $p(\alpha)$: CPU time (seconds) to reduce relative residual by 10^{-2}, when C is given by* (5.2).

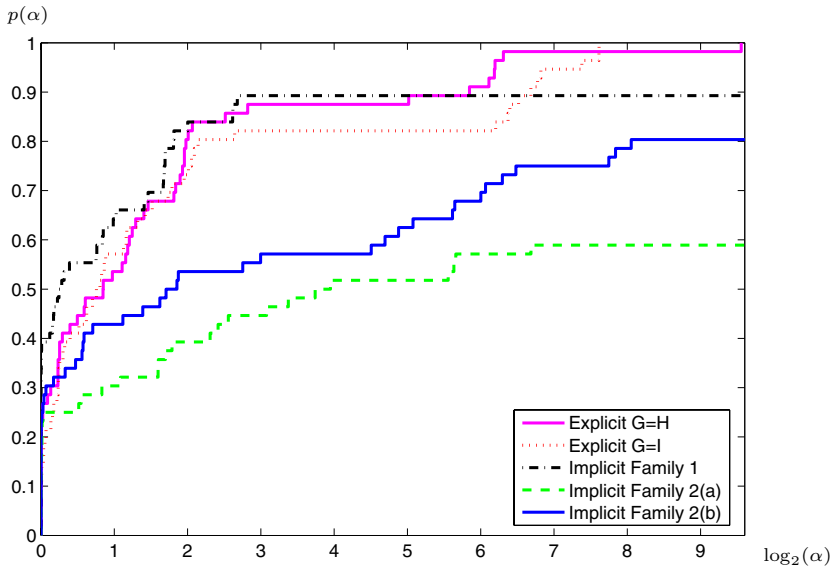


FIG. 5.2. *Performance profile, $p(\alpha)$: CPU time (seconds) to reduce relative residual by 10^{-8}, when C is given by* (5.2).

parts. In particular, the implicit factorization from family 1 seems to be the most effective. Of interest is that for family 2, the cost of applying the more accurate implicit factorization that reproduces H_{22} generally does not pay off relative to the cost of the cheaper implicit factorizations. For higher accuracy solutions, the leading implicit factorization still slightly outperforms the explicit factors, although now the remaining implicit factorizations are less effective.

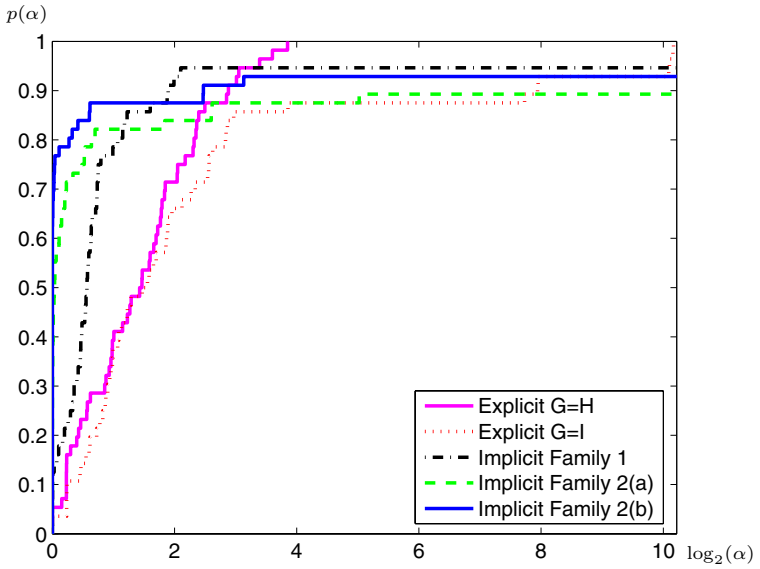


FIG. 5.3. *Performance profile, $p(\alpha)$: CPU time (seconds) to reduce relative residual by 10^{-2}, when C is given by* (5.3).

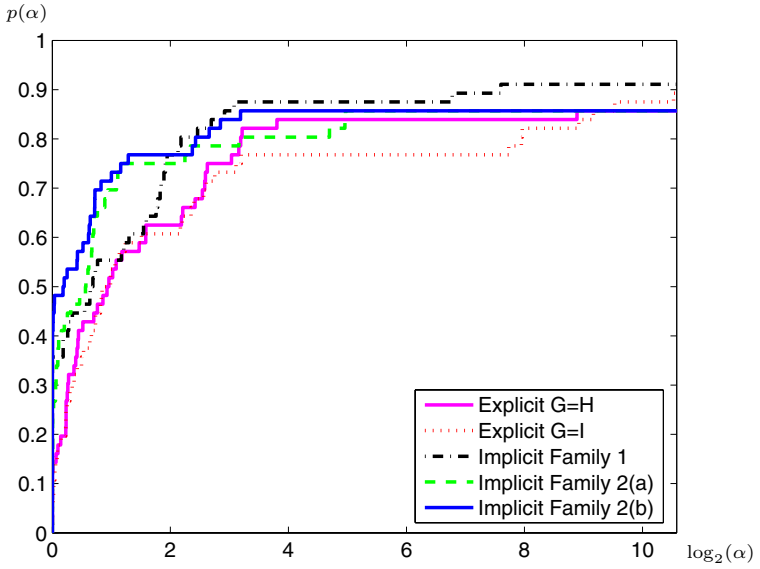


FIG. 5.4. *Performance profile, $p(\alpha)$: CPU time (seconds) to reduce relative residual by 10^{-8}, when C is given by* (5.3).

Figures 5.3–5.4 (see [17] for tables of the raw data) repeat the experiments when C is given by (5.3). Once again the implicit factorizations seem very effective, with a shift now to favor those from family 2, most especially the less sophisticated of these.

6. Comments and conclusions. In this paper we have considered conjugate-gradient like methods for block symmetric indefinite linear systems that arise from regularized saddle-point problems. Such methods require preconditioners that pre-

serve certain sub-blocks from the original systems but allow considerable flexibility for the remaining "noncrucial" blocks. To this end, we have constructed fourteen families of implicit factorizations that are capable of reproducing the required sub-blocks and (some) of the remainder. These generalize known implicit factorizations [18, 19] for the $C = 0$ case. Improved eigenvalue clustering is possible if additionally some of the "noncrucial" blocks are reproduced. We have shown numerically that these implicit-factorization preconditioners can be effective. However, further work is needed to see how these preconditioners compare against special-purpose ones based on (4.1) rather than generic ones using factors of (1.2).

A number of important issues remain. Firstly, we have made no effort to find the best preconditioner(s) from amongst our families, and indeed in most cases have not even tried them in practice. As always with preconditioning, there is a delicate balance between improving clustering of eigenvalues and the cost of doing so, especially since in many applications low accuracy estimates of the solution suffice. We expect promising candidates to emerge in due course, but feel it is beyond the scope of this paper to indicate more than that this is a promising approach.

Secondly, and as we pointed out in [18], the choice of the matrix A_1 is crucial, and considerations of both its stability and sparsity (or other structure), and of its effect on which of the "noncrucial" blocks may be reproduced, are vital. We have precisely defined the algorithm that we have used to select A_1 in the computations presented in this paper, but though this strategy seems to work reasonably across the wide range of test set problems we have computed, we make no claim to its relative quality. The most stringent practical requirement for computation with the preconditioners described in this paper is that there is an effective way to solve linear systems involving A_1.

Thirdly (and possibly related to the point above), when experimenting with family 3 (Table 4.1), we found that some very badly conditioned preconditioners were generated. Specifically, our aim had been to reproduce $G_{21} = H_{21}$, and for simplicity we had chosen $P_{31} = I = B_{31}$ and $B_{22} = I = P_{22}$, and this leads to $P_{21} = H_{21}A_1^{-1}$. Note that we did not try to impose additionally that $G_{22} = H_{22}$ as this would have lead to nontrivial B_{22}. Also notice that we did not need to form P_{21}, merely to operate with it (and its transpose) on given vectors. On examining the relevant spectrum for some small badly conditioned examples, the preconditioner appeared to have worsened rather than improved the range of the eigenvalues for these computations. Whether this is a consequence of requiring two solves with A_1 (and its transpose) when applying the preconditioner rather than the single solve required when not trying to reproduce H_{21}, and whether the same would be true for other families trying to do the same is simply conjecture at this stage. However, it is certainly a cautionary warning.

Acknowledgment. Thanks are due to Mario Arioli both for fruitful discussions on various aspects of this work and for providing us with a Matlab interface to `MA57`. We also thank two referees and the associate editor for their helpful comments.

REFERENCES

[1] O. Axelsson and M. Neytcheva, *Preconditioning methods for linear systems arising in constrained optimization problems*, Numer. Linear Algebra Appl., 10 (2003), pp. 3–31.

[2] R. H. Bartels and G. H. Golub, *The simplex method of linear programming using lu decompositions*, Comm. of the ACM, 12 (1969), pp. 266–268.

[3] M. Benzi, G. H. Golub, and J. Liesen, *Numerical solution of saddle point problems*, Acta Numer., 14 (2005), pp. 1–137.

[4] L. Bergamaschi, J. Gondzio, and G. Zilli, *Preconditioning indefinite systems in interior point methods for optimization*, Comput. Optim. Appl., 28 (2004), pp. 149–171.
[5] G. Biros and O. Ghattas, *A Lagrange-Newton-Krylov-Schur method for PDE-constrained optimization*, SIAG/OPT Views-and-News, 11 (2000), pp. 12–18.
[6] J. H. Bramble and J. E. Pasciak, *A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems*, Math. Comput., 50 (1988), pp. 1–17.
[7] R. H. Byrd, M. E. Hribar, and J. Nocedal, *An interior point algorithm for large scale nonlinear programming*, SIAM J. Optim., 9 (2000), pp. 877–900.
[8] T. F. Coleman, *Linearly constrained optimization and projected preconditioned conjugate gradients*, in Proceedings of the Fifth SIAM Conference on Applied Linear Algebra, J. Lewis, ed., SIAM Philadelphia, 1994, pp. 118–122.
[9] T. F. Coleman and A. Pothen, *The null space problem* I: *Complexity*, SIAM J. Algebraic Discrete Methods, 7 (1986), pp. 527–537.
[10] T. F. Coleman and A. Pothen, *The null space problem* II: *Algorithms*, SIAM J. Algebraic Discrete Methods, 8 (1987), pp. 544–563.
[11] T. F. Coleman and A. Verma, *A preconditioned conjugate gradient approach to linear equality constrained minimization*, Technical report, Department of Computer Sciences, Cornell University, Ithaca, New York, 1998.
[12] P. Concus, G. H. Golub, and D. P. O'Leary, *Numerical solution of nonlinear elliptic partial differential equations by a generalized conjugate gradient method*, in Sparse Matrix Computations, J. Bunch and D. Rose, eds., Academic Press, London, 1976, pp. 309–332,
[13] A. R. Conn, N. I. M. Gould, D. Orban, and Ph. L. Toint, *A primal-dual trust-region algorithm for non-convex nonlinear programming*, Math. Prog., 87 (2000), pp. 215–249.
[14] E. de Sturler and J. Liesen, *Block-diagonal and constraint preconditioners for nonsymmetric indefinite linear systems Part* I: *Theory*, SIAM J. Sci. Comput., 26 (2005), pp. 1598–1619.
[15] E. D. Dolan and J. J. Moré, *Benchmarking optimization software with performance profiles*, Math. Prog., 91 (2002), pp. 201–213.
[16] H. S. Dollar, *Extending constraint preconditioners for saddle-point problems*, Technical report NA-05-02, Oxford University Computing Laboratory, Oxford, England, 2005. (Submitted to SIAM J. Matrix Anal. Appl.)
[17] H. S. Dollar, N. I. M. Gould, W. H. A. Schilders, and A. J. Wathen, *On iterative methods and implicit-factorization preconditioners for regularized saddle-point systems*, Technical report RAL-TR-2005-011, Rutherford Appleton Laboratory, Chilton, Oxfordshire, England, 2005.
[18] H. S. Dollar, N. I. M. Gould, and A. J. Wathen, *On implicit-factorization constraint preconditioners*, Nonconvex Optimization and Its Applications 83, Springer Verlag, 2006.
[19] H. S. Dollar and A. J. Wathen, *Incomplete factorization constraint preconditioners for saddle point problems*, Technical report 04-01, Oxford University Computing Laboratory, Oxford, England, 2004. (To appear in SIAM J. Sci. Comput.)
[20] I. S. Duff, *MA57 - a code for the solution of sparse symmetric definite and indefinite systems*, ACM Trans. Math. Software, 30 (2004), pp. 118–144.
[21] I. S. Duff and J. K. Reid, *The multifrontal solution of indefinite sparse symmetric linear equations*, ACM Trans. Math. Software, 9 (1983), pp. 302–325.
[22] C. Durazzi and V. Ruggiero, *Indefinitely constrained conjugate gradient method for large sparse equality and inequality constrained quadratic problems*, Numer. Linear Algebra Appl., 10 (2002), pp. 673–688.
[23] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Finite-Elements and Fast Iterative Solvers: With Applications in Incompressible Fluid Dynamics*, Numerical Mathematics and Scientific Computation, Oxford University Press, London, 2005.
[24] J. J. H. Forrest and J. A. Tomlin, *Updating triangular factors of the basis to maintain sparsity in the product form simplex method*, Math. Prog., 2 (1972), pp. 263–278.
[25] R. W. Freund and N. M. Nachtigal, *A new Krylov-subspace method for symmetric indefinite linear systems*, in Proceedings of the 14th IMACS World Congress on Computational and Applied Mathematics, W. F. Ames, ed., IMACS, 1994, pp. 1253–1256.
[26] P. E. Gill, W. Murray, D. B. Ponceleón, and M. A. Saunders, *Preconditioners for indefinite systems arising in optimization*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 292–311.
[27] P. E. Gill, W. Murray, and M. A. Saunders, *SNOPT: An SQP algorithm for large-scale constrained optimization*, SIAM J. Optim., 12 (2002), pp. 979–1006.
[28] P. E. Gill and M. A. Saunders, *private communication*, 1999.
[29] N. I. M. Gould, *On practical conditions for the existence and uniqueness of solutions to the general equality quadratic-programming problem*, Math. Prog., 32 (1985), pp. 90–99.
[30] N. I. M. Gould, *Iterative methods for ill-conditioned linear systems from optimization*, in *Nonlinear Optimization and Related Topics*, G. Di Pillo and F. Giannessi, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999, pp. 123–142.

[31] N. I. M. Gould, M. E. Hribar, and J. Nocedal, *On the solution of equality constrained quadratic problems arising in optimization*, SIAM J. Sci. Comput., 23 (2001), pp. 1375–1394.

[32] N. I. M. Gould, D. Orban, and Ph. L. Toint, *CUTEr (and SifDec), a constrained and unconstrained testing environment, revisited*, ACM Trans. Math. Software, 29 (2003), pp. 373–394.

[33] C. Greif, G. H. Golub, and J. M. Varah, *Augmented Lagrangian techniques for solving saddle point linear systems*, Technical report, Computer Science Department, University of British Columbia, Vancouver, Canada, 2004.

[34] HSL, *A collection of Fortran codes for large-scale scientific computation*, see http://hsl.rl.ac.uk, 2004.

[35] C. Keller, N. I. M. Gould, and A. J. Wathen, *Constraint preconditioning for indefinite linear systems*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1300–1317.

[36] L. Lukšan and J. Vlček, *Indefinitely preconditioned inexact Newton method for large sparse equality constrained nonlinear programming problems*, Numer. Linear Algebra Appl., 5 (1998), pp. 219–247.

[37] M. Mihajlovic and D. Silvester, *A black-box multigrid preconditioner for the biharmonic equation*, BIT, 44 (2004), pp. 151–163.

[38] B. A. Murtagh and M. A. Saunders, *Large-scale linearly constrained optimization*, Math. Prog., 14 (1978), pp. 41–72.

[39] B. A. Murtagh and M. A. Saunders, *A projected Lagrangian algorithm and its implementation for sparse non-linear constraints*, Math. Programming Stud., 16 (1982), pp. 84–117.

[40] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Series in Operations Research, Springer, New York, 1999.

[41] L. A. Pavarino, *Preconditioned mixed spectral finite-element methods for elasticity and Stokes problems*, SIAM J. Sci. Comput., 19 (1998), pp. 1941–1957.

[42] I. Perugia and V. Simoncini, *Block-diagonal and indefinite symmetric preconditioners for mixed finite element formulations*, Numer. Linear Algebra Appl., 7 (2000), pp. 585–616.

[43] B. T. Polyak, *The conjugate gradient method in extremal problems*, U.S.S.R. Comput. Math. Math. Phys., 9 (1969), pp. 94–112.

[44] M. Rozložník and V. Simoncini, *Krylov subspace methods for saddle point problems with indefinite preconditioning*, SIAM J. Matrix Anal. Appl., 24 (2002), pp. 368–391.

[45] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Comput., 7 (1986), pp. 856–869.

[46] W. Schilders, *A preconditioning technique for indefinite systems arising in electronic circuit simulation*, Talk at the one-day meeting on preconditioning methods for indefinite linear systems, Technische Universiteit, Eindhoven, The Netherlands, December 9, 2002.

[47] W. H. A. Schilders and E. J. W. ter Maten, *Numerical methods in electromagnetics*, in Handbook of Numerical Analysis Vol XIII, P. G. Ciarlet, ed., Elsevier, North Holland, Amsterdam, 2005.

[48] K. Toh, K. Phoon, and S. Chan, *Block preconditioners for symmetric indefinite linear systems*, Internat. J. Numer. Methods Engrg., 60 (2004), pp. 1361–1381.

[49] R. J. Vanderbei and D. F. Shanno, *An interior point algorithm for nonconvex nonlinear programming*, Comput. Optim. Appl., 13 (1999), pp. 231–252.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 1, pp. 190–212

NUMERICAL METHODS FOR SOLVING INVERSE EIGENVALUE PROBLEMS FOR NONNEGATIVE MATRICES*

ROBERT ORSI†

Abstract. Presented here are two related numerical methods, one for the inverse eigenvalue problem for nonnegative or stochastic matrices and another for the inverse eigenvalue problem for symmetric nonnegative matrices. The methods are iterative in nature and utilize alternating projection ideas. For the algorithm for the symmetric problem, the main computational component of each iteration is an eigenvalue-eigenvector decomposition, while for the other algorithm, it is a Schur matrix decomposition. Convergence properties of the algorithms are investigated and numerical results are also presented. While the paper deals with two specific types of inverse eigenvalue problems, the ideas presented here should be applicable to many other inverse eigenvalue problems, including those involving nonsymmetric matrices.

Key words. inverse eigenvalue problem, nonnegative matrices, stochastic matrices, alternating projections, Schur's decomposition

AMS subject classifications. 15A51, 65F18

DOI. 10.1137/050634529

1. Introduction.

A real $n \times n$ matrix is said to be nonnegative if each of its entries is nonnegative.

The [illegible] (NIEP) is the following: given a list of n complex numbers $\lambda = \{\lambda_1, \ldots, \lambda_n\}$, find a nonnegative $n \times n$ matrix with eigenvalues λ (if such a matrix exists).

A related problem is the [illegible] (SNIEP): given a list of n real numbers $\lambda = \{\lambda_1, \ldots, \lambda_n\}$, find a symmetric nonnegative $n \times n$ matrix with eigenvalues λ (if such a matrix exists)[1].

Finding necessary and sufficient conditions for a list λ to be realizable as the eigenvalues of a nonnegative matrix has been a challenging area of research for over fifty years, and this problem is still unsolved [12]. As noted in [6, section 6], while various necessary or sufficient conditions exist, the necessary conditions are usually too general while the sufficient conditions are too specific. Under a few special sufficient conditions, a nonnegative matrix with the desired spectrum can be constructed; however, in general, proofs of sufficient conditions are nonconstructive. Two sufficient conditions that are constructive and not restricted to small n are, respectively, given in [20], for the SNIEP, and [21], for the NIEP with real λ. (See also [19] for an extension of the results of the latter paper.) A good overview of known results relating to necessary or sufficient conditions can be found in the recent survey paper [12] and general background material on nonnegative matrices, including inverse eigenvalue problems and applications, can be found in the texts [2] and [18]. We also mention the recent paper [9], which can be used to help determine whether a given list λ may be realizable as the eigenvalues of a nonnegative matrix.

*Received by the editors June 27, 2005; accepted for publication (in revised form) by M. Chu November 7, 2005; published electronically March 17, 2006. This work was supported by the Australian Research Council through grant DP0450539.

http://www.siam.org/journals/simax/28-1/63452.html

†Research School of Information Sciences and Engineering, Australian National University, Canberra ACT 0200, Australia (robert.orsi@anu.edu.au).

[1]The NIEP and SNIEP are different problems even if λ is restricted to contain only real entries; there exist lists of n real numbers λ for which the NIEP is solvable but the SNIEP is not [16].

In this paper we are interested in generally applicable numerical methods for solving NIEPs and SNIEPs. To the best of our knowledge, the only algorithms that have appeared up to now in the literature consist of [5] for the SNIEP and [8] for the NIEP. In the case of [5], the following constrained optimization problem is considered:

$$\min_{Q^TQ=I,\ R=R^T} \frac{1}{2}\|Q^T\Lambda Q - R\circ R\|^2. \tag{1.1}$$

Here Λ is a constant diagonal matrix with the desired spectrum and $\circ$ stands for the Hadamard product, i.e., componentwise product. Note that the symmetric matrices with the desired spectrum are exactly the elements of $\{Q^T\Lambda Q \mid Q \in \mathbb{R}^{n\times n} \text{ orthogonal}\}$ and that the symmetric nonnegative matrices are exactly the elements of $\{R\circ R \mid R \in \mathbb{R}^{n\times n} \text{ symmetric}\}$. In [5], a gradient flow based on (1.1) is constructed. A solution to the SNIEP is found if the gradient flow converges to a Q and an R that zero the objective function. The approach taken in [8] for the NIEP is similar but is complicated by the fact that the set of all matrices, both symmetric and nonsymmetric, with a particular desired spectrum is not nicely parameterizable. In particular, these matrices can no longer be parameterized by the orthogonal matrices.

In this paper we present a numerical algorithm for the NIEP and another for the SNIEP. In both cases, the problems are posed as problems of finding a point in the intersection of two particular sets. Unlike the approaches in [5] and [8] which are based on gradient flows, our algorithms are iterative in nature. For the SNIEP, the solution methodology is based on a alternating projection scheme between the two sets in question. The solution methodology for the NIEP is also based on an alternating projection-like scheme but is more involved, as we will shortly explain.

While alternating projections can often be a very effective means of finding a point in the intersection of two or more convex sets, for both the SNIEP and NIEP formulations, one set is nonconvex. Nonconvexity of one of the sets means that alternating projections may not converge to a solution. This is in contrast to the case where all sets are convex and convergence to a solution is guaranteed.

In addition to problem formulations, the development of the corresponding algorithms, and their convergence analysis, another contribution of the paper is as follows. As mentioned above, for each problem, one set in the problem formulation is nonconvex. For the NIEP, this set is particularly complicated; it consists of all matrices with the desired spectrum. At least some of the members of this set will be nonsymmetric matrices and it is this that causes complications. In particular, though the set is closed and hence projections are well defined theoretically, how to calculate projections onto such sets is an unsolved difficult problem. We formulate an alternate method for mapping onto this set. Though the resulting points are not necessarily projected points, they are members of the set and share a number of other desirable properties. As will be shown, this alternate "projection" is very effective in our context. Furthermore, we believe that it may also be quite effective for other inverse eigenvalue problems involving nonsymmetric matrices[2]. For more on other inverse eigenvalue problems, see the survey papers [4] and [6], and the recent text [7].

Before concluding this introductory section we would like to point out how the NIEP is related to another problem involving stochastic matrices. A $n \times n$ matrix is said to be stochastic if it is nonnegative and the sum of the entries in each row equals one. Another variation of the NIEP is the [illegible]

[2]Preliminary indications of this are given in [24] and [23], where this idea is applied to inverse eigenvalue type problems arising in control theory.

(StIEP): given a list of n complex numbers $\lambda = \{\lambda_1, \ldots, \lambda_n\}$, find a stochastic $n \times n$ matrix with eigenvalues λ (if such a matrix exists). It turns out that the NIEP and the StIEP are almost exactly the same problem, as we now show. (See also [8].)

The vector of all 1's is always an eigenvector for a stochastic matrix, implying each stochastic matrix must have 1 as an eigenvalue. Also, the maximum row sum matrix norm of a stochastic matrix equals 1 and hence the spectral radius cannot be greater than 1, and as a result, must actually equal 1. Suppose λ satisfies the above mentioned necessary conditions to be the spectrum of a stochastic matrix and that a nonnegative matrix A with this spectrum can be found. Then if an eigenvector x of A corresponding to the eigenvalue 1 can be chosen to have positive entries (by the Perron–Frobenius theorem this is certainly possible if A is irreducible), then, if we define $D = \mathrm{diag}(x)$, it is straightforward to verify that

$$D^{-1}AD$$

is a stochastic matrix with the desired spectrum. (In fact it can be shown that if λ satisfies the above mentioned necessary conditions, then it is the spectrum of a stochastic matrix if and only if it is the spectrum of a nonnegative matrix [22, Lemma 5.3.2].)

The rest of the paper is structured as follows. The last part of this section contains some notation. Projections play a key part in the algorithms and section 2 contains general properties of projections that are used throughout the paper. The SNIEP algorithm is presented first, in section 3, and then insights from this algorithm are used to address the more difficult NIEP in section 4. Section 5 contains convergence results. This includes a detailed analysis of fixed points of the SNIEP algorithm for the $n = 2$ case. This is the easiest case though we believe the analysis presented is still quite interesting and also gives insight into higher-dimensional problems. Numerical results for both algorithms are presented in section 6, and an appendix contains some supplementary projection results.

Notation. $\mathbb{R}$ is the set of real numbers. $\mathbb{C}$ is the set of complex numbers. $\mathcal{S}^n$ is the set of real symmetric $n \times n$ matrices. A^T denotes the transpose of a matrix A. A^* denotes the complex conjugate transpose of a matrix A. $\mathrm{tr}(A)$ denotes the sum of the diagonal elements of a square matrix A. For two $n \times n$ symmetric matrices A and B, $[A, B]$ denotes $AB - BA$. $\mathrm{diag}(v)$ for $v \in \mathbb{C}^n$ denotes the $n \times n$ diagonal matrix whose ith diagonal term is v_i. $\mathrm{Re}(z)$ denotes the real part of $z \in \mathbb{C}$.

2. Projections. Projections play a key part in the algorithms. This section contains general properties of projections that will be used throughout the paper.

Let x be an element in a Hilbert space H and let C be a closed (possibly non-convex) subset of H. Any $c_0 \in C$ such that $\|x - c_0\| \leq \|x - c\|$ for all $c \in C$ will be called a [illegible] of x onto C. In the cases of interest here, namely where H is a finite dimensional Hilbert space, there is always at least one such point for each x. If C is convex as well as closed, then each x has exactly one such minimum distance point [17]. Where convenient, we will use $y = P_C(x)$ to denote that y is a projection of x onto C. We emphasize that $y = P_C(x)$ only says y is [illegible] projection of x onto C and does [illegible] make any statement regarding uniqueness.

All problems of interest in this paper are feasibility problems of the following abstract form.

PROBLEM 2.1. [illegible] $C_1, \ldots, C_N$ [illegible]

H

$$\bigcap_{i=1}^{N} C_i$$

(In fact, we will solely be interested in the case $N = 2$.)

If all the C_i's in Problem 2.1 are convex, a classical method of solving Problem 2.1 is to alternatively project onto the C_i's. This method is often referred to as the method of alternating projections (MAP). If the C_i's have a nonempty intersection, the successive projections are guaranteed to asymptotically converge to an intersection point [3].

THEOREM 2.2 (MAP). $C_1, \ldots, C_N$ H $\bigcap_{i=1}^{N} C_i$ x_0

$$x_{i+1} = P_{C_{\phi(i)}}(x_i), \quad \phi(i) = (i \bmod N) + 1,$$

$\bigcap_{i=1}^{N} C_i$

We remark that the usefulness of MAP for finding a point in the intersection of a number of sets is dependent on being able to compute projections onto each of the C_i's.

While MAP is not guaranteed to converge to a solution if one or more of the C_i's is nonconvex, for alternating projections between two sets, the following distance reduction property always holds.

THEOREM 2.3. C_1 C_2 H $y_0 \in C_2$

$$x_1 = P_{C_1}(y_0), \quad y_1 = P_{C_2}(x_1), \quad x_2 = P_{C_1}(y_1),$$

$$\|x_2 - y_1\| \leq \|x_1 - y_1\| \leq \|x_1 - y_0\|.$$

. The second inequality holds as y_1 is a projection of x_1 onto C_2 and hence its distance to x_1 is less than or equal to the distance of x_1 to any other point in C_2 such as y_0. The first inequality holds by similar reasoning. □

COROLLARY 2.4. $i = 0, 1, \ldots,$

$$x_{i+1} = P_{C_1}(y_i), \quad y_{i+1} = P_{C_2}(x_{i+1}),$$

x_i y_i $\|x_i - y_i\|$ i

Suppose one is interested in solving Problem 2.1 in the case of two sets, C_1 and C_2, when one or both sets are nonconvex. If projections onto these sets are computable, a solution method is to alternately project onto C_1 and C_2. Corollary 2.4 ensures that the distance $\|x_i - y_i\|$ is nonincreasing with i. While this is promising, there is, however, no guarantee that this distance goes to zero and hence that a solution to the problem will be found.

Most of the literature on alternating projection methods deals with the case of convex subsets of a (possibly infinite dimensional) Hilbert space; a survey of these results is contained in [1]. The text [11] is also recommended. There is much less available for the case of one or more nonconvex sets; see in particular [10].

3. The symmetric problem. Our algorithm for solving the SNIEP consists of alternately projecting onto two particular sets. The details are given in this section.

Given a list of real eigenvalues $\lambda = \{\lambda_1, \dots, \lambda_n\}$, renumbering if necessary, suppose $\lambda_1 \geq \cdots \geq \lambda_n$. Let

$$\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n), \tag{3.1}$$

and let $\mathcal{M}$ denote the set of all real symmetric matrices with eigenvalues λ,

$$\mathcal{M} = \{A \in \mathcal{S}^n \mid A = V \Lambda V^T \text{ for some orthogonal } V\}. \tag{3.2}$$

Let $\mathcal{N}$ denote the set of symmetric nonnegative matrices,

$$\mathcal{N} = \{A \in \mathcal{S}^n \mid A_{ij} \geq 0 \text{ for all } i, j\}. \tag{3.3}$$

The SNIEP can now be stated as the following particular case of Problem 2.1:

$$\text{Find } X \in \mathcal{M} \cap \mathcal{N}. \tag{3.4}$$

Our solution approach is to alternatively project between $\mathcal{M}$ and $\mathcal{N}$, and we next show that it is indeed possible to calculate projections onto these sets. First, in order for the term "projection" to make sense, we need to define an appropriate Hilbert space and associated norm. From now on, $\mathcal{S}^n$ will be viewed as a Hilbert space with inner product

$$\langle A, B \rangle = \operatorname{tr}(AB) = \sum_{i,j} A_{ij} B_{ij}. \tag{3.5}$$

The associated norm is the Frobenius norm $\|A\| = \langle A, A \rangle^{\frac{1}{2}}$.

The projection of $A \in \mathcal{S}^n$ onto $\mathcal{M}$ is given by Theorem 3.2 below. More precisely, it gives [illegible] projection of A onto $\mathcal{M}$. The reason for this is that the set $\mathcal{M}$ is nonconvex[3] and hence projections onto this set are not guaranteed to be unique. We will need the following classical result [13, section 10.2].

Lemma 3.1. [illegible] $x, y \in \mathbb{R}^n$ [illegible] $x_1 \geq \cdots \geq x_n$ [illegible] $y_1 \geq \cdots \geq y_n$ [illegible] σ [illegible] $\{1, \dots, n\}$ [illegible]

$$\sum_i x_i y_i \geq \sum_i x_i y_{\sigma(i)}.$$

Theorem 3.2. [illegible] $A \in \mathcal{S}^n$ [illegible] $A = V \operatorname{diag}(\mu_1, \dots, \mu_n) V^T$ [illegible] V [illegible] $\mu_1 \geq \cdots \geq \mu_n$ [illegible] Λ [illegible] (3.1) [illegible] $V \Lambda V^T$ [illegible] $\mathcal{M}$ [illegible]

[illegible] For all $X \in \mathcal{M}$, $\operatorname{tr}(X^2) = \operatorname{tr}(\Lambda^2)$. As a result, finding $X \in \mathcal{M}$ that minimizes $\|X - A\|^2$ is the same as finding $X \in \mathcal{M}$ that maximizes $\operatorname{tr}(XA)$. Consider the function

$$f : \mathcal{M} \to \mathbb{R}, \; X \mapsto \operatorname{tr}(XA).$$

[3] $\mathcal{M}$ is nonconvex if its defining λ contains a pair of nonequal eigenvalues. For example, if $n = 2$, consider

$$A = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} \lambda_2 & 0 \\ 0 & \lambda_1 \end{bmatrix}.$$

If $\lambda_1 \neq \lambda_2$, then the convex combination $(A + B)/2$ does not have the same spectrum as A and B.

$\mathcal{M}$ is a smooth manifold and its tangent space at a point X is $\{[X,\Omega] \mid \Omega = -\Omega^T \in \mathbb{R}^{n\times n}\}$; see, for example, [14, Chapter 2]. The derivative of f at a point X in the tangent direction $[X,\Omega]$ is

$$Df(X)([X,\Omega]) = \operatorname{tr}([X,\Omega]A) = \operatorname{tr}((AX - XA)\Omega).$$

If X maximizes f, then this derivative must be zero in all tangent directions, or equivalently, $AX - XA$ must be symmetric. This in turn is equivalent to X and A commuting. X and A commute if and only if they are simultaneously diagonalizable; see [15, Theorem 2.5.15]. Hence if X maximizes f, then there must exist an orthogonal matrix U and a diagonal matrix Λ_σ with the same spectrum as Λ such that $A = U\operatorname{diag}(\mu_1,\dots,\mu_n)U^T$ and $X = U\Lambda_\sigma U^T$. This combined with Lemma 3.1 implies f has maximum value $\operatorname{tr}(\Lambda\operatorname{diag}(\mu_1,\dots,\mu_n))$ and implies the result. ∎

Projection onto $\mathcal{N}$ is straightforward and is given by Theorem 3.3 below.

THEOREM 3.3. [illegible] $A \in \mathcal{S}^n$ [illegible] $A_+ \in \mathcal{S}^n$ [illegible]

$$(A_+)_{ij} = \max\{A_{ij}, 0\}, \quad 1 \le i,j \le n. \tag{3.6}$$

A_+ [illegible] $\mathcal{N}$ [illegible]

[illegible]. The projection of $x \in \mathbb{R}$ onto the nonnegative real numbers equals $\max\{x,0\}$. The general result follows by noting that if $B \in \mathcal{S}^n$, and in particular if $B \in \mathcal{N}$, then

$$\|A - B\| = \left(\sum_{i,j} |A_{ij} - B_{ij}|^2\right)^{\frac{1}{2}}$$

and hence that the problem reduces to n^2 decoupled scalar problems. ∎

Our proposed algorithm for solving the SNIEP is the following.

SNIEP algorithm:

[illegible] List of desired real eigenvalues $\lambda = \{\lambda_1,\dots,\lambda_n\}$, $\lambda_1 \ge \cdots \ge \lambda_n$.

[illegible] Choose a randomly generated symmetric nonnegative matrix $Y \in \mathbb{R}^{n\times n}$.

repeat

1. Calculate an eigenvalue-eigenvector decomposition of Y: $Y = V\operatorname{diag}(\mu_1,\dots,\mu_n)V^T$, $\mu_1 \ge \cdots \ge \mu_n$.
2. $X := V\operatorname{diag}(\lambda_1,\dots,\lambda_n)V^T$.
3. $X := (X + X^T)/2$.
4. $Y := X_+$.

until $\|X - Y\| < \epsilon$.

In the above algorithm, X_+ is given by (3.6).

Note that at each iteration of the algorithm, X has the desired spectrum λ and Y is nonnegative. If ϵ is small, say $\epsilon = 10^{-14}$, termination of the loop ensures X equals Y (approximately) and hence that Y solves the SNIEP.

Due to small numerical inaccuracy, X from Step 2 of the algorithm may not be perfectly symmetric. Step 3 makes it so.

Of course, while Corollary 2.4 ensures $\|X - Y\|$ is nonincreasing from one iteration to the next, the set $\mathcal{M}$ is nonconvex and hence there is no guarantee that the algorithm will terminate. A detailed analysis of convergence is postponed to section 5.

4. The general problem. Throughout this section, $\mathbb{C}^{n\times n}$ will be viewed as a Hilbert space with inner product

$$\langle A, B\rangle = \operatorname{tr}(AB^*) = \sum_{i,j} A_{ij}\overline{B}_{ij}.$$

The associated norm is the Frobenius norm $\|A\| = \langle A, A\rangle^{\frac{1}{2}}$.

Recall Schur's result that any matrix $A \in \mathbb{C}^{n\times n}$ is unitarily equivalent to an upper triangular matrix.

THEOREM 4.1. [illegible] $A \in \mathbb{C}^{n\times n}$ [illegible] $\mu_1, \ldots, \mu_n$ [illegible] $U \in \mathbb{C}^{n\times n}$ [illegible] $T \in \mathbb{C}^{n\times n}$ [illegible]

$$A = UTU^* \tag{4.1}$$

[illegible] $T_{ii} = \mu_i \quad i = 1, \ldots, n$ [illegible]. See, for example, [15, Theorem 2.3.1]. ∎

We now redefine some terms from the prior section.

Let $\lambda = \{\lambda_1, \ldots, \lambda_n\}$ be a given list of complex eigenvalues. Define

$$\mathcal{T} = \{T \in \mathbb{C}^{n\times n} \mid T \text{ is upper triangular with spectrum } \lambda\}. \tag{4.2}$$

Theorem 4.1 implies that the set of all complex matrices with spectrum λ is given by the following set:

$$\mathcal{M} = \{A \in \mathbb{C}^{n\times n} \mid A = UTU^* \text{ for some unitary } U \text{ and some } T \in \mathcal{T}\}. \tag{4.3}$$

Let $\mathcal{N}$ denote the set of (not necessarily symmetric) nonnegative matrices,

$$\mathcal{N} = \{A \in \mathbb{R}^{n\times n} \mid A_{ij} \geq 0 \text{ for all } i, j\}. \tag{4.4}$$

Having redefined $\mathcal{M}$ and $\mathcal{N}$, the NIEP can now be stated as the following particular case of Problem 2.1:

$$\text{Find } X \in \mathcal{M} \cap \mathcal{N}. \tag{4.5}$$

A difficulty now occurs. We would like to use alternating projections to solve the NIEP. However, to the best of our knowledge, the way to calculate projections onto $\mathcal{M}$ is an unsolved problem. Suppose instead we could find a mapping that was in some sense a reasonable substitute for a projection map for $\mathcal{M}$. Using this substitute mapping and the projection map for $\mathcal{N}$ in an alternating projection-like scheme may still produce a viable algorithm. Indeed, we now propose the following function $P_{\mathcal{M}}$ as a substitute for a true projection map onto $\mathcal{M}$. (The notation $P_{\mathcal{M}}$ is used as it is suggestive; however, recall that we have already used $y = P_C(x)$ to denote that y is a projection of x onto a set C. The two different uses of the notation should be clear from their context and should not cause confusion.)

DEFINITION 4.2. [illegible] $U \in \mathbb{C}^{n\times n}$ [illegible] $T \in \mathbb{C}^{n\times n}$ [illegible] $\{\hat{\lambda}_1, \ldots, \hat{\lambda}_n\}$ [illegible] λ [illegible]

$$\sum_{i=1}^{n} |\hat{\lambda}_i - T_{ii}|^2. \tag{4.6}$$

$$P_{\mathcal{M}}(U,T) = U\hat{T}U^*, \tag{4.7}$$

$\hat{T} \in \mathcal{T}$

$$\hat{T}_{ij} = \begin{cases} \hat{\lambda}_i & i = j \\ T_{ij} & \end{cases}$$

Note that $P_{\mathcal{M}}$ maps into the set $\mathcal{M}$.

A given matrix $A \in \mathbb{C}^{n\times n}$ may have a nonunique Schur decomposition and $A = U_1T_1U_1^* = U_2T_2U_2^*$ does imply $P_{\mathcal{M}}(U_1,T_1) = P_{\mathcal{M}}(U_2,T_2)$. For example, if

$$T_1 = \begin{bmatrix} 1 & 1 & 4 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{bmatrix}, \quad T_2 = \begin{bmatrix} 2 & -1 & 3\sqrt{2} \\ 0 & 1 & \sqrt{2} \\ 0 & 0 & 3 \end{bmatrix}, \quad \text{and} \quad U = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & \sqrt{2} \end{bmatrix},$$

then U is unitary and $UT_1U^* = T_2$, [15]. If $\lambda = \{0,0,0\}$, $P_{\mathcal{M}}(U,T_1) \neq P_{\mathcal{M}}(I,T_2)$.

It turns out that this nonuniqueness is not particularly important. The following result shows that for different Schur decompositions of the same matrix, $P_{\mathcal{M}}$ gives points in $\mathcal{M}$ of equal distance from the original matrix.

THEOREM 4.3. $A = U_1T_1U_1^* = U_2T_2U_2^*$ $U_1, U_2 \in \mathbb{C}^{n\times n}$ $T_1, T_2 \in \mathbb{C}^{n\times n}$

$$\|P_{\mathcal{M}}(U_1,T_1) - A\| = \|P_{\mathcal{M}}(U_2,T_2) - A\|.$$

. Suppose $A = UTU^*$, where U is unitary and T is upper triangular. If $\hat{T}$ is the matrix given in Definition 4.2, then by the unitary invariance of the Frobenius norm,

$$\|P_{\mathcal{M}}(U,T) - A\| = \|\hat{T} - T\|.$$

As $\|\hat{T} - T\|^2$ equals the quantity in (4.6), $\|P_{\mathcal{M}}(U,T) - A\|$ depends only on λ and $T_{11}, \ldots, T_{nn}$. The result now follows by noting that the T_{ii}'s are the eigenvalues of A and that (4.6) does not depend on the ordering of the T_{ii}'s. □

The next theorem shows that given $A = UTU^*$, if we restrict attention to matrices of the form $U\tilde{T}U^*$, $\tilde{T} \in \mathcal{T}$, then $P_{\mathcal{M}}(U,T)$ is a point in $\mathcal{M}$ closest to A.

THEOREM 4.4. $A = UTU^* \in \mathbb{C}^{n\times n}$ U T $P_{\mathcal{M}}(U,T)$

$$\|P_{\mathcal{M}}(U,T) - A\| \leq \|U\tilde{T}U^* - A\| \quad \tilde{T} \in \mathcal{T}.$$

. Let $\tilde{T}$ be a matrix in $\mathcal{T}$. The unitary invariance of the Frobenius norm implies the result will be established if we can show

$$\|\hat{T} - T\| \leq \|\tilde{T} - T\|,$$

where $\hat{T}$ is the matrix given in Definition 4.2. Note that

$$\|\tilde{T} - T\|^2 = \sum_{i=1}^{n} |\tilde{T}_{ii} - T_{ii}|^2 + \sum_{i\neq j} |\tilde{T}_{ij} - T_{ij}|^2 \tag{4.8}$$

and that the $\tilde{T}_{ii}$'s are some permutation of the list of eigenvalues λ. The result follows by noting that $\|\hat{T} - T\|^2$ equals the quantity in (4.6) and that this value must be less than or equal to the first summation on the right-hand side of the equality in (4.8). ∎

For completeness, we note that, given $A = UTU^*$, $P_{\mathcal{M}}(U,T)$ may not satisfy

$$\|P_{\mathcal{M}}(U,T) - A\| \leq \|M - A\| \text{ for all } M \in \mathcal{M}.$$

For example if

$$U = \frac{1}{5}\begin{bmatrix} -3 & 4 \\ 4 & 3 \end{bmatrix}, \quad T = \begin{bmatrix} 1 & -3 \\ 0 & 2 \end{bmatrix}, \quad \tilde{U} = \frac{1}{5}\begin{bmatrix} -4 & 3 \\ 3 & 4 \end{bmatrix}, \quad \tilde{T} = \begin{bmatrix} 0 & -3 \\ 0 & 0 \end{bmatrix},$$

and $\lambda = \{0, 0\}$, then one can readily verify that

$$\|P_{\mathcal{M}}(U,T) - UTU^*\| \not\leq \|\tilde{U}\tilde{T}\tilde{U}^* - UTU^*\|.$$

As for the symmetric case, projection onto $\mathcal{N}$ is straightforward.

THEOREM 4.5. [illegible] $A \in \mathbb{C}^{n\times n}$ [illegible] $A_+ \in \mathbb{R}^{n\times n}$ [illegible]

$$(A_+)_{ij} = \max\{\mathrm{Re}(A_{ij}), 0\} \text{ [illegible] } 1 \leq i, j \leq n. \tag{4.9}$$

A_+ [illegible] $\mathcal{N}$ [illegible] A [illegible]. The projection of $z \in \mathbb{C}$ onto the nonnegative real numbers is given by $\max\{\mathrm{Re}(z), 0\}$. The remainder of the proof follows by exactly the same reasoning used in the proof of Theorem 3.3. ∎

Our proposed algorithm for solving the NIEP is the following.

NIEP algorithm:

[illegible] List of desired complex eigenvalues $\lambda = \{\lambda_1, \ldots, \lambda_n\}$.

[illegible] Choose a randomly generated nonnegative matrix $Y \in \mathbb{R}^{n\times n}$.

repeat

1. Calculate a Schur decomposition of Y: $Y = UTU^*$.
2. $X := P_{\mathcal{M}}(U,T)$.
3. $Y := X_+$.

until $\|X - Y\| < \epsilon$.

In the above algorithm, $P_{\mathcal{M}}(U,T)$ is given by Definition 4.2 and X_+ is given by (4.9).

As for the SNIEP algorithm, at each iteration of the NIEP algorithm, X has the desired spectrum λ and Y is nonnegative. If ϵ is small, say $\epsilon = 10^{-14}$, termination of the loop ensures X equals Y (approximately) and hence that Y solves the NIEP.

REMARK 4.6. If each of the members of λ are real and we seek a symmetric nonnegative matrix with spectrum λ, then the NIEP algorithm reduces to the SNIEP algorithm. More precisely, this is true if the members of λ are real, if the initial condition Y is a symmetric nonnegative matrix, and, for Schur decompositions used in the NIEP algorithm, if U is restricted to be real.

Indeed, suppose the current Y is symmetric and nonnegative. For any Schur decomposition of Y, T must be a real diagonal matrix. As we restrict the U matrix to be real, such a decomposition is nothing but a standard eigenvalue-eigenvector

decomposition for a symmetric matrix (though the eigenvalue are not necessarily ordered along the diagonal of T).

As both the elements of λ and the diagonal entries of T are real, the permutation that minimizes (4.6) can be easily characterized. Indeed, in this case (4.6) is minimized if and only if

$$\sum_{i=1}^{n} \hat{\lambda}_i T_{ii} \tag{4.10}$$

is maximized. From Lemma 3.1, (4.10) is maximized if the $\hat{\lambda}_i$'s are ordered in the same way as the T_{ii}'s. This implies that if Y is symmetric, the step of producing a X from Y is the same in both algorithms.

Lastly, projection of a symmetric matrix onto (4.4) gives the same matrix as projection onto (3.3) and hence this step in both algorithms is also the same. This establishes our claim.

We close this section by noting that unlike the SNIEP algorithm, for the NIEP algorithm there is no guarantee that $\|X - Y\|$ is nonincreasing from one iteration to the next.

5. Convergence. In this section we study the convergence properties of the SNIEP and NIEP algorithms. We present a number of results for the SNIEP algorithm, though limit ourselves to a local convergence result for the NIEP algorithm. We start by characterizing the SNIEP algorithm fixed points. All references to "$\mathcal{M}$," "$\mathcal{N}$," and "the algorithm" refer to the SNIEP versions of these objects, unless otherwise stated.

5.1. Fixed points. As there may be more than one projection of a point Y onto the set $\mathcal{M}$, some care needs to be taken in regard to the definition of fixed points of the algorithm. This subsection includes such a definition, as well as a characterization of these points.

DEFINITION 5.1. $X \in \mathcal{S}^n$ [illegible] fixed point [illegible] X_+

$$X_+ = U\tilde{\Lambda}U^T, \tag{5.1}$$

[illegible] $\tilde{\Lambda} = \operatorname{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n)$ $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_n$ [illegible] U [illegible]

$$X = U\Lambda U^T. \tag{5.2}$$

X is a fixed point if and only if there is a projection of X_+ onto $\mathcal{M}$ which equals X.

We would like to point out an important fact regarding this definition of a fixed point. If $X \in \mathcal{M}$ is an infeasible fixed point, that is, $X \in \mathcal{M}$ is a fixed point which is not in the solution set $\mathcal{M} \cap \mathcal{N}$, it does [illegible] necessarily mean that the algorithm cannot make further progress toward a feasible solution from X. This is a consequence of the possible nonuniqueness of the algorithm's matrix decompositions, as we now explain.

If X_+ has distinct eigenvalues, then the orthonormal eigenvectors of X_+ are unique up to multiplication by -1. In this case, any decomposition (5.1) of X_+ will result in the same projected point (5.2). On the other hand, if X_+ has repeated eigenvalues, then there are an infinite number of different decompositions of X_+. If this is the case for an infeasible fixed point X, it may be possible to escape from X by forcing the algorithm to use a different decomposition of X_+.

THEOREM 5.2. [illegible] $X \in \mathcal{M}$ [illegible] V [illegible] $X_+ = V\tilde{\Lambda}V^T$ [illegible] $(V\Lambda V^T)_+ \neq X_+$ [illegible] X_+ [illegible] $\mathcal{N}$ [illegible] X [illegible]. We will show that

$$\|X - X_+\| = \|V\Lambda V^T - X_+\| > \|V\Lambda V^T - (V\Lambda V^T)_+\|. \tag{5.3}$$

Note that if (5.3) holds, then $V\Lambda V^T \in \mathcal{M}$ and $(V\Lambda V^T)_+ \in \mathcal{N}$ are closer together than X and X_+, and the distance reduction property, Theorem 2.3, implies the result.

The equality in (5.3) holds as both X and $V\Lambda V^T$ are projections of X_+ onto $\mathcal{M}$. The inequality in (5.3) follows by noting that, as $\mathcal{N}$ is closed and convex, $(V\Lambda V^T)_+$ is the unique closest point in $\mathcal{N}$ to $V\Lambda V^T$. □

For the main result of this subsection, we will need the following lemma.

LEMMA 5.3. [illegible] $x, y \in \mathbb{R}^n$ [illegible] $x_1 \geq \cdots \geq x_n$ $y_1 \geq \cdots \geq y_n$ [illegible] σ [illegible] $\{1, \ldots, n\}$ [illegible]

$$\sum_i x_i y_i = \sum_i x_i y_{\sigma(i)}, \tag{5.4}$$

[illegible] i [illegible] j

$$i < j \quad \text{[illegible]} \quad y_{\sigma(i)} < y_{\sigma(j)},$$

[illegible]

$$x_i = x_j. \tag{5.5}$$

[illegible]. From Lemma 3.1, for any permutation π of $\{1, \ldots, n\}$,

$$\sum_i x_i y_i \geq \sum_i x_i y_{\pi(i)}. \tag{5.6}$$

Suppose (5.5) does not hold. Then $x_i > x_j$, which implies

$$(x_i - x_j)(y_{\sigma(i)} - y_{\sigma(j)}) < 0,$$

or rearranging the terms,

$$x_i y_{\sigma(i)} + x_j y_{\sigma(j)} < x_i y_{\sigma(j)} + x_j y_{\sigma(i)}.$$

This combined with (5.4) implies there is a permutation that violates (5.6). As this is not possible, (5.5) must hold. □

THEOREM 5.4. [illegible] $X \in \mathcal{M}$ X [illegible]

$$[X, X_+] = 0 \tag{5.7}$$

[illegible]

$$\mathrm{tr}(\Lambda\tilde{\Lambda}) = \mathrm{tr}(\tilde{\Lambda}^2), \tag{5.8}$$

[illegible] $\tilde{\Lambda} = \operatorname{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n)$ $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_n$ [illegible] X_+

[illegible]. ($\Rightarrow$) If X is a fixed point, there exists an orthogonal matrix U such that

$$X_+ = U\tilde{\Lambda}U^T \quad \text{and} \quad X = U\Lambda U^T. \tag{5.9}$$

Hence X and X_+ commute and (5.7) holds.

Equality (5.8) follows from

$$\operatorname{tr}(XX_+) = \operatorname{tr}(X_+^2)$$

and (5.9).

($\Leftarrow$) From (5.7), X and X_+ are simultaneously diagonalizable. Hence there exists an orthogonal matrix U and a diagonal matrix $\tilde{\Lambda}_\sigma$ whose diagonal entries are a permutation of the diagonal entries of $\tilde{\Lambda}$ such that

$$X_+ = U\tilde{\Lambda}_\sigma U^T \quad \text{and} \quad X = U\Lambda U^T.$$

By an argument similar to the one used in the first part of the proof,

$$\operatorname{tr}(\Lambda\tilde{\Lambda}_\sigma) = \operatorname{tr}(\tilde{\Lambda}_\sigma^2). \tag{5.10}$$

Equalities (5.8) and (5.10) imply

$$\operatorname{tr}(\Lambda\tilde{\Lambda}) = \operatorname{tr}(\Lambda\tilde{\Lambda}_\sigma).$$

From Lemma 5.3, if $i < j$ and $(\tilde{\Lambda}_\sigma)_{ii} < (\tilde{\Lambda}_\sigma)_{jj}$, then $\Lambda_{ii} = \Lambda_{jj}$. Hence, the columns of U can always be reordered to get a new U so that (5.9) holds, and hence, X is a fixed point. □

REMARK 5.5. It is interesting to compare the fixed points of the algorithm with those of the SNIEP gradient flow algorithm of [5]. The gradient flow used in [5] is

$$\begin{aligned} \frac{dX}{dt} &= [X, [X, Y]], \\ \frac{dY}{dt} &= 4Y \circ (X - Y). \end{aligned} \tag{5.11}$$

If $(X(t), Y(t))$, $t \geq 0$, is a solution of this differential equation, then $X(t)$ is isospectral, that is, it preserves the spectrum of $X(0)$, and $Y(t)$ is nonnegative for all $t \geq 0$ if $Y(0)$ is. Suppose $X(0)$ is chosen to have the desired spectrum and $Y(0)$ is chosen nonnegative. Then, if $X(t)$ and $Y(t)$ converge to the same point, that point is a solution of the problem.

The fixed points of (5.11) are the points (X, Y) for which the right-hand side is 0:

$$\begin{aligned} [X, [X, Y]] &= 0, \\ Y \circ (X - Y) &= 0. \end{aligned}$$

Note that for any $X, Y \in \mathcal{S}^n$, $[X, [X, Y]] = 0$ if and only if $[X, Y] = 0$: If $[X, [X, Y]] = 0$, then $0 = \operatorname{tr}(Y[X, [X, Y]]) = \operatorname{tr}([X, Y]^T[X, Y])$ and hence $[X, Y] = 0$. For any $X \in \mathcal{S}^n$, $X_+ \circ (X - X_+) = 0$. Hence, if $X \in \mathcal{M}$ is a fixed point of our SNIEP algorithm, then (X, X_+) is a fixed point of the algorithm of [5]. Roughly speaking,

the set of fixed points of the SNIEP algorithm of this paper is a subset of the set of fixed points of the algorithm of [5].

There do exist infeasible fixed points. If Λ contains negative values (the SNIEP is trivial if it does not), then $X = \Lambda$ is an infeasible fixed point. It may or may not be possible to escape from such a fixed point using alternate decompositions of X_+. An example where escape via this technique is not possible is when $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2)$ with $\lambda_1 > 0$ and $\lambda_2 < 0$, in which case Λ_+ has distinct eigenvalues.

If X is a fixed point, then so is PXP^T for any permutation matrix P. In particular, $P\Lambda P^T$ is a fixed point for any permutation matrix P.

The attractive set of the diagonal fixed points includes the matrices with nonpositive off diagonal terms: Suppose X is such a matrix. Then X_+ is diagonal and can be decomposed as $X_+ = P\tilde{\Lambda}P^T$, where $\tilde{\Lambda}$ is diagonal with ordered diagonal entries and P is a permutation matrix. For this decomposition of X_+, Step 2 of the algorithm maps onto the diagonal fixed point $P\Lambda P^T$.

5.2. General convergence properties. The following is a general result regarding convergence to fixed points.

THEOREM 5.6. [illegible] $X_1, X_2, \ldots,$ [illegible] X [illegible]

$$\|X - X_+\| = \lim_{i\to\infty} \|X_i - (X_i)_+\|. \tag{5.12}$$

[illegible] (5.12) [illegible]

. The X_i's are elements of the compact set $\mathcal{M}$ and hence contain a convergent subsequence. Let X_{i_k} denote the kth element in this subsequence and denote the limit of the X_{i_k}'s by $\hat{X}$. For each k, let U_{i_k} and $\tilde{\Lambda}_{i_k}$ be the matrices from the decomposition of $(X_{i_k})_+$ used to produce X_{i_k+1}, that is, $(X_{i_k})_+ = U_{i_k}\tilde{\Lambda}_{i_k}U_{i_k}^T$ and $X_{i_k+1} = U_{i_k}\Lambda U_{i_k}^T$. As the U_{i_k}'s are members of a compact set, without loss of generality we can assume they converge to a point U. This implies the X_{i_k+1}'s converge, and we denote the corresponding limit point by X.

Corollary 2.4 implies $\lim_{i\to\infty}\|X_i - (X_i)_+\|$ exists. This and the fact that projection onto $\mathcal{N}$ is a continuous operation implies

$$\|\hat{X} - \hat{X}_+\| = \lim_{k\to\infty} \|X_{i_k} - (X_{i_k})_+\| = \lim_{k\to\infty} \|X_{i_k+1} - (X_{i_k+1})_+\| = \|X - X_+\|. \tag{5.13}$$

As $(X_{i_k})_+$ converges to $\hat{X}_+$ and the orthogonal matrices U_{i_k} converges to U, it follows that the $\tilde{\Lambda}_{i_k}$'s also converge to, say, $\tilde{\Lambda}$. Hence $\hat{X}_+ = \lim_{k\to\infty}(X_{i_k})_+ = U\tilde{\Lambda}U^T$ and $X = \lim_{k\to\infty} X_{i_k+1} = U\Lambda U^T$. This implies X is a projection of $\hat{X}_+$ onto $\mathcal{M}$.

The equality (5.13) and the fact that X is a projection of $\hat{X}_+$ onto $\mathcal{M}$ imply

$$\|X - X_+\| \geq \|X - \hat{X}_+\|.$$

As X_+ is the unique projection of X onto $\mathcal{N}$, this implies $\hat{X}_+ = X_+$. As X is a projection of $\hat{X}_+ = X_+$ onto $\mathcal{M}$, X is a fixed point.

Suppose now that the limit in (5.12) equals zero. Consider an arbitrary subsequence $X_{i_1}, X_{i_2}, \ldots$, which converges to some point $\tilde{X}$. Note that $\tilde{X}$ is a limit of points in $\mathcal{M}$. The inequality

$$\|(X_{i_k})_+ - \tilde{X}\| \leq \|(X_{i_k})_+ - X_{i_k}\| + \|X_{i_k} - \tilde{X}\|$$

implies it is also a limit of points in $\mathcal{N}$. The last part of the theorem now follows as both $\mathcal{M}$ and $\mathcal{N}$ are closed. ∎

The next theorem gives a local convergence result which holds for both SNIEP and NIEP algorithms. If $\mathcal{M}$ and $\mathcal{N}$ are given, respectively, by (3.2) and (3.3), and $\overset{\circ}{\mathcal{N}}$ denotes the interior of $\mathcal{N}$, then, if the intersection of $\mathcal{M}$ and $\overset{\circ}{\mathcal{N}}$ is nonempty, the SNIEP algorithm converges to a solution from points in an open neighborhood of this intersection set. The analogous result for the NIEP algorithm is also true.

THEOREM 5.7. [illegible] $\mathcal{M}$ [illegible] $\mathcal{N}$ [illegible] (3.2) [illegible] (3.3) [illegible] (4.3) [illegible] (4.4) [illegible] $\mathcal{M} \cap \overset{\circ}{\mathcal{N}}$ [illegible] $\mathcal{M} \cap \overset{\circ}{\mathcal{N}}$ [illegible]. We prove the result for the NIEP algorithm. The proof for the SNIEP algorithm is almost identical.

Suppose $X \in \mathcal{M} \cap \overset{\circ}{\mathcal{N}}$ and let $\epsilon > 0$ be small enough so that the open ball $B(X, \epsilon)$ is a subset of $\mathcal{N}$. Choose $\delta > 0$ such that if $Y \in B(X, \delta)$ and $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n$ are the eigenvalues of Y, then (reordering the $\tilde{\lambda}_i$'s if necessary)

$$\left(\sum_i |\lambda_i - \tilde{\lambda}_i|^2 \right)^{\frac{1}{2}} < \epsilon/2. \tag{5.14}$$

Here the λ_i's are the desired eigenvalues which define the NIEP. Decreasing δ if necessary, we assume $\delta \leq \epsilon/2$.

We now show that if $Y \in B(X, \delta)$, then any projection of Y onto $\mathcal{M}$ is in $B(X, \epsilon)$. As $B(X, \epsilon) \subset \mathcal{N}$, such a projection of Y is a solution of the problem.

Let $Y \in B(X, \delta)$ and suppose it has Schur decomposition $Y = UTU^*$. Consider

$$\|X - P_{\mathcal{M}}(U, T)\| \leq \|X - Y\| + \|Y - P_{\mathcal{M}}(U, T)\|.$$

The first term on the right of the inequality is less than $\epsilon/2$, as is the second term by the definition of $P_{\mathcal{M}}(U, T)$ and (5.14). This completes the result. ∎

As we will see in the next subsection, every feasible $n = 2$ SNIEP has only a finite number of infeasible fixed points. The following result exploits such a situation.

THEOREM 5.8. [illegible] $c > 0$ [illegible] $X_1, X_2, \ldots,$ [illegible] X [illegible] $\|X_{i^*} - (X_{i^*})_+\| < c$ [illegible] i^* [illegible] X_i [illegible]. Let $c > 0$ be such that $\|Z - Z_+\| \geq c$ for all infeasible fixed points Z. Such a c exists as there are only a finite number of infeasible fixed points. By Corollary 2.4, $\|X_i - (X_i)_+\|$ is a nonincreasing function of i and hence must have a limit. Theorem 5.6 implies this limit must be zero. The rest now also follows from Theorem 5.6. ∎

If a limit point in Theorem 5.8 is in $\mathcal{M} \cap \overset{\circ}{\mathcal{N}}$, then Theorem 5.7 implies the sequence of X_i's will converge to a solution (in a finite number of iterations).

Do all feasible SNIEPs have only a finite number of infeasible fixed points? Alternatively, as $\mathcal{M}$ is compact, an equivalent question is: Are all infeasible fixed points of a feasible SNIEP isolated? These are interesting questions, to which we currently do not have an answer.

5.3. Further analysis: $n = 2$ SNIEP. In this subsection we continue our analysis of convergence; in particular we investigate the $n = 2$ SNIEP. Though necessary and sufficient conditions exist for the solvability of the $n = 2$ SNIEP, and there exists an analytic solution when these conditions are met, we believe the analysis presented here is still quite interesting and also gives insight into higher-dimensional problems.

As noted in [5], for $n = 2$, feasible SNIEPs have a very nice geometric interpretation. If the 2×2 symmetric matrices are parameterized by $\mathbb{R}^3$ in the standard way and the eigenvalues defining $\mathcal{M}$ are distinct, then the points with the desired spectrum form a one dimensional ellipse in $\mathbb{R}^3$, and the SNIEP is equivalent to finding a point on this ellipse that is also in the nonnegative orthant of $\mathbb{R}^3$.

As the trace of a matrix equals the sum of its eigenvalues, a necessary condition for solvability is that $\lambda_1 + \lambda_2 \geq 0$. In fact, this condition is also sufficient; if it is met, then a solution of the problem is

$$X = \frac{1}{2}\begin{pmatrix} \lambda_1 + \lambda_2 & \lambda_1 - \lambda_2 \\ \lambda_1 - \lambda_2 & \lambda_1 + \lambda_2 \end{pmatrix}. \tag{5.15}$$

As normal, here we assume the eigenvalues are ordered: $\lambda_1 \geq \lambda_2$.

The feasible cases can be enumerated as follows:

1. $\lambda_1 = \lambda_2 \geq 0$,
2. $\lambda_1 > \lambda_2 \geq 0$,
3. $\lambda_1 > 0 > \lambda_2$, $\lambda_1 \geq |\lambda_2|$.

Theorem 5.10 below characterizes the infeasible fixed points of the algorithm for the different cases listed above. We will need the following lemma.

LEMMA 5.9. [illegible]

$$X = \begin{pmatrix} a & b \\ b & a \end{pmatrix}$$

[illegible] a [illegible] $b \leq 0$ [illegible] $X \in \mathcal{M}$ [illegible]

$$X = \frac{1}{2}\begin{pmatrix} \lambda_1 + \lambda_2 & \lambda_2 - \lambda_1 \\ \lambda_2 - \lambda_1 & \lambda_1 + \lambda_2 \end{pmatrix}. \tag{5.16}$$

[illegible]. ($\Rightarrow$) $X \in \mathcal{M}$ and $b \leq 0$ implies $\lambda_1 = a - b$ and $\lambda_2 = a + b$. Solving for a and b gives (5.16).

($\Leftarrow$) If X is given by (5.16), then its eigenvalues are λ_1 and λ_2, and hence it is a member of $\mathcal{M}$. As $\lambda_2 - \lambda_1 \leq 0$, X_+ is a constant multiple of the identity. If U is any orthogonal matrix such that $X = U\Lambda U^T$, then $X_+ = U\tilde{\Lambda}U^T$ with $\tilde{\Lambda} = X_+$ and hence X is a fixed point. □

THEOREM 5.10. [illegible] 1 $\mathcal{M}$ [illegible] (5.15) [illegible] $\lambda_1 I$ [illegible] 2 [illegible] (5.16) [illegible] 3 [illegible] (5.16)

$$X = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \tag{5.17}$$

[illegible]

$$X = \begin{pmatrix} \lambda_2 & 0 \\ 0 & \lambda_1 \end{pmatrix}. \tag{5.18}$$

. We assume $X \in \mathcal{M}$ is a fixed point given by

$$X = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

and consider all possibilities for a, b, and c.

2. If $a = 0$ or $c = 0$, then $\det(X) \geq 0$ implies $b = 0$, which implies X is feasible. If $a > 0 > c$ or $c > 0 > a$, then $\det(X) < 0$, which is not possible. If $a < 0$ and $c < 0$, then $\operatorname{tr}(X) < 0$, which is not possible. Hence it remains to consider the subcase $a > 0$ and $c > 0$.

Suppose $a > 0$ and $c > 0$. If $b \geq 0$, X is feasible. If $b < 0$, then $X_+ = \operatorname{diag}(a, c)$. If $a \neq c$, then the orthonormal eigenvectors of X_+, up to multiplication by -1, are the standard orthonormal basis vectors for $\mathbb{R}^2$. This would imply X is diagonal, contradicting $b < 0$. If $a = c$, by Lemma 5.9, the only $X \in \mathcal{M}$ which is a fixed point is given by (5.16). As $\lambda_2 - \lambda_1 < 0$, X is infeasible.

3. By considering the equalities $\operatorname{tr}(X) = \lambda_1 + \lambda_2$ and $\det(X) = \lambda_1\lambda_2$, it follows that

$$X = \begin{pmatrix} a & b \\ b & \lambda_1 + \lambda_2 - a \end{pmatrix}, \quad b = \pm\sqrt{(\lambda_1 - a)(a - \lambda_2)}, \quad \text{and that } a \in [\lambda_2, \lambda_1].$$

Suppose $b \leq 0$. Then $X_+ = \operatorname{diag}(a_+, (\lambda_1+\lambda_2-a)_+)$. X_+ has repeated eigenvalues if and only if $a = (\lambda_1+\lambda_2)/2$. If X_+ does have repeated eigenvalues, then the diagonal terms of X are equal and by Lemma 5.9 the only $X \in \mathcal{M}$ which is a fixed point is given by (5.16). As $\lambda_2 - \lambda_1 < 0$, X is infeasible. If X_+ has distinct eigenvalues, then X must be one of the infeasible fixed points (5.17) or (5.18).

Suppose $b > 0$. If $a \in [0, \lambda_1 + \lambda_2]$, X is feasible. It is not possible that $a = \lambda_1$ or $a = \lambda_2$ as then, $b = 0$. Hence it remains to consider $a \in (\lambda_1 + \lambda_2, \lambda_1)$ (the case $a \in (\lambda_2, 0)$ follows from this case by replacing a with $\lambda_1 + \lambda_2 - a$). If X_+ and X are given by (5.1) and (5.2), respectively, then

$$X - X_+ = \begin{pmatrix} 0 & 0 \\ 0 & \lambda_1 + \lambda_2 - a \end{pmatrix} = U(\Lambda - \tilde{\Lambda})U^T.$$

This implies $\Lambda - \tilde{\Lambda}$ has distinct eigenvalues and hence that (up to multiplication of its columns by -1)

$$U = I \quad \text{or} \quad U = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

This implies X is diagonal but this contradicts the fact that $b > 0$ and hence this subcase cannot occur. □

Consider the infeasible fixed point X given by (5.16). The U satisfying (5.1) and (5.2) is unique up to multiplication of its columns by -1. It is given by

$$U = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}.$$

Notice that X_+ is a constant multiple of the identity. If $X_+ = V\tilde{\Lambda}V^T$ is any alternate decomposition of X_+, that is, if V is orthogonal matrix that does not equal U nor U with one or both of its columns multiplied by -1, then $V\Lambda V^T \neq X$, and

Theorem 5.10 implies $(V\Lambda V^T)_+ \neq X_+$. Hence, Theorem 5.2 implies, using almost any decomposition of X_+, the algorithm is able to escape from X.

The only other infeasible fixed points are the diagonal ones given by (5.17) and (5.18). For these fixed points, X_+ has distinct eigenvalues, and alternate decompositions as a means of escape cannot be utilized. Despite this, the next theorem shows that such fixed points are unstable and one can escape from them by adding an arbitrarily small perturbation.

THEOREM 5.11. [illegible] X [illegible] (5.17) [illegible] (5.18) $\lambda_1 > 0 > \lambda_2$ [illegible] $\bar{b}$ [illegible] b [illegible] $0 < b \leq \bar{b}$ [illegible] $\bar{a} = \bar{a}(b)$ [illegible] $\bar{c} = \bar{c}(b)$ [illegible] $|a| \leq \bar{a}$ [illegible] $|c| \leq \bar{c}$ [illegible] X [illegible] $X + P$ [illegible]

$$P = \begin{pmatrix} a & b \\ b & c \end{pmatrix},$$

[illegible] X [illegible]. To prove the theorem we will show that if P is as above and $(X+P)_+ = V\tilde{\Lambda}V^T$, then

$$\|X_+ - X\| > \|(X+P)_+ - V\Lambda V^T\|. \tag{5.19}$$

Inequality (5.19) and Theorem 2.3 together imply we cannot return to X.

We assume the fixed point X is given by (5.17). (The proof of the (5.18) case is identical except for a permutation of matrix rows and columns.)

Suppose $a \geq -\lambda_1$, $b > 0$, and $c \leq -\lambda_2$. Then

$$(X+P)_+ = \begin{pmatrix} \lambda_1 + a & b \\ b & 0 \end{pmatrix}. \tag{5.20}$$

If $\tilde{\Lambda} = \operatorname{diag}(\tilde{\lambda}_1, \tilde{\lambda}_2)$, then

$$\tilde{\lambda}_1, \tilde{\lambda}_2 = \frac{\lambda_1 + a \pm \sqrt{(\lambda_1 + a)^2 + 4b^2}}{2}.$$

By the unitary invariance of the Frobenius norm, (5.19) is equivalent to

$$\lambda_2^2 > (\lambda_1 - \tilde{\lambda}_1)^2 + (\lambda_2 - \tilde{\lambda}_2)^2.$$

Substituting $\tilde{\lambda}_1 + \tilde{\lambda}_2 = \lambda_1 + a$, we have

$$\lambda_2^2 > (\tilde{\lambda}_2 - a)^2 + (\lambda_2 - \tilde{\lambda}_2)^2. \tag{5.21}$$

For now, suppose $a = 0$. Noting that $\tilde{\lambda}_2 < 0$ as $b \neq 0$, straightforward algebraic manipulations imply (5.21) is equivalent to $\lambda_2 < \tilde{\lambda}_2$. Hence, for $a = 0$, (5.19) holds if and only if $\lambda_1 - \sqrt{\lambda_1^2 + 4b^2} > 2\lambda_2$. As $\lambda_1 > 0 > \lambda_2$, this inequality holds for all $b > 0$ small enough.

Each $b > 0$ that satisfies (5.21) when $a = 0$ also satisfies this inequality for all a sufficiently small as the right-hand side of (5.21) depends continuously on a. □

Suppose that for $i = 1, 2, \ldots$, $(a, b) = (a_i, b_i)$, $b_i > 0$, satisfies (5.21) and that the b_i's converge to zero. Examination of (5.21) and the expression for $\tilde{\lambda}_2$ shows that the a_i's must also converge to zero, and hence this theorem is the best we can do.

Theorem 5.11 can be readily extended to infeasible diagonal fixed points of problems of size $n > 2$.

TABLE 6.1
SNIEP: A comparison of performance for different problem sizes n. i denotes the average number of iterations and T denotes the average convergence time in CPU seconds.

n	i	T	% solved
5	19	0.0016	100
10	18	0.0030	100
20	17	0.0075	100
100	12	0.15	100

6. Numerical experiments. This section contains some numerical results for both the SNIEP and NIEP algorithms.

All computational results were obtained using a 3 GHz Pentium 4 machine. The algorithms were coded using Matlab 7.0.

Throughout this section, when we say a matrix is "randomly generated" we mean each entry of that matrix is randomly drawn from the uniform distribution on the interval $[0, 1]$. When dealing with the SNIEP algorithm, all randomly generated matrices are chosen symmetric.

For both algorithms, the initial starting Y is always randomly generated and the convergence tolerance ϵ is set to 10^{-14}.

A final note before presenting the results: Suppose $\mathcal{M} \cap \mathring{\mathcal{N}}$ is nonempty and X is a member of this set. Then for any real orthogonal matrix Q that is sufficiently close to the identity, QXQ^T is also a solution. In particular, if $\mathcal{M} \cap \mathring{\mathcal{N}}$ is nonempty, then there will be multiple solutions. This comment applies to both SNIEPs and NIEPs.

6.1. SNIEP. This subsection starts with some results for randomly generated SNIEPs. To ensure each problem is feasible, each desired spectrum is taken from a randomly generated matrix.

Results for various problem sizes n are given in Table 6.1. For each value of n, 1000 problems were considered. The table contains the average number of iterations required to find a solution, the average time required to find a solution, and the success rate. As can be seen, the algorithm performed extremely well and was able to solve every problem. In all cases, both the average number of iterations and the average solution time was very small.

REMARK 6.1. It is interesting to note that T increases with n, as would be expected, while i decreases. A reason for this could be the following. As already mentioned, for any choice of desired eigenvalues, $\mathcal{M}$ is a smooth manifold. In addition, if the eigenvalues defining $\mathcal{M}$ are distinct, as they will be if they were taken from a randomly generated matrix, then the dimension of $\mathcal{M}$ is $n(n-1)/2$; see [14, Chapter 2]. The dimension of $\mathcal{S}^n$ is $n(n+1)/2$. Hence,

$$\frac{\dim \mathcal{M}}{\dim \mathcal{S}^n} = \frac{n-1}{n+1},$$

which is an increasing function of n. For larger n, $\mathcal{M}$ is "thicker" relative to the ambient space and hence, intuitively, the corresponding SNIEP is easier to solve.

Suppose $X_1, X_2, \ldots$, is a sequence of X's produced by the SNIEP algorithm and that these points converge to a solution $\bar{X}$. Figure 6.1 shows a typical plot of $\|X_i - \bar{X}\|$ versus i. Convergence is clearly linear. This is to be expected: Suppose $\bar{X}$ is a point on the boundary of $\mathcal{N}$ and that the $(X_i)_+$'s lie in a particular face of $\mathcal{N}$. As $\mathcal{M}$ is a manifold, near $\bar{X}$ it looks locally like an affine subspace of $\mathcal{S}^n$. As the face of $\mathcal{N}$

TABLE 6.2

SNIEP: A problem with repeated eigenvalues, $\lambda = \{3-t, 1+t, -1, -1, -1, -1\}$. i denotes the average number of iterations and T denotes the average convergence time in CPU seconds. i and T do not include the attempts that had not converged after 5000 *iterations.*

t	i	T	% solved
0.25	480	0.061	100
0.5	470	0.061	97
0.75	340	0.050	65
0.95	310	0.046	59

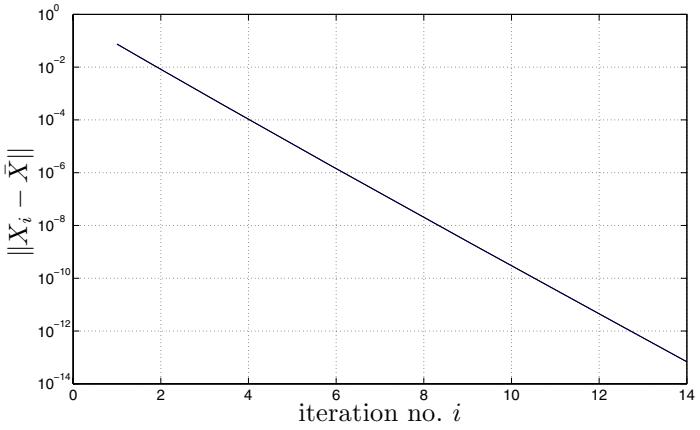


FIG. 6.1. *Linear convergence of the SNIEP algorithm.*

also looks locally like an affine subspace, we would expect local linear convergence as alternating projections between two intersecting affine subspaces converge linearly [11].

Randomly generated problems have properties that are not shared by all SNIEPs. For example, as already mentioned, randomly generated problems have distinct eigenvalues. We next consider a problem with repeated eigenvalues, namely $\lambda = \{3-t, 1+t, -1, -1, -1, -1\}$ for $0 < t < 1$. The $t = 1/2$ version of the problem is also considered in [5], where a numerical solution is sought via the gradient flow approach of that paper. An analytic solution to this problem is given in [20].

Notice that for any value of t the desired eigenvalues sum to zero and hence there exist arbitrarily small perturbations of the spectrum which lead to an infeasible SNIEP. In particular this problem cannot have any solutions in the interior of $\mathcal{N}$. We have tried the SNIEP algorithm on a number of other problems with repeated eigenvalues with excellent results. This is the hardest problem we have encountered so far.

The results of applying the algorithm to the problem for various values of t are given in Table 6.2. They are based on running the algorithm 100 times for each value of t.

First, the results indicate that the SNIEP algorithm is not always successful in finding a solution. However, they also show that the algorithm can still be quite successful if a number of initial conditions are tried. It is interesting to note that the algorithm becomes more sensitive to the choice of the initial condition the larger t is. Notice that as $t \to 1$, the eigenvalues $3-t$ and $1+t$ both converge to the same value,

TABLE 6.3

NIEP: A comparison of performance for different problem sizes n. i denotes the average number of iterations and T denotes the average convergence time in CPU seconds. i and T do not include the problems that had not converged after 5000 *iterations.*

n	i	T	% solved
5	26	0.011	99.7
10	44	0.045	99.8
20	48	0.12	99.8
100	200	12	96.6

and the dimension of the manifold $\mathcal{M}$ (which depends solely on the multiplicities of the eigenvalues) goes from 9 when $0 < t < 1$ to 8 when $t = 1$ [14, Chapter 2].

Aside: Regarding initial conditions, as noted before, both the SNIEP and NIEP algorithms use a nonnegative initial starting point. This is important and, in fact, the performance of neither algorithm is as good if non-nonnegative initial conditions are used.

Here is a solution that was found to the $t = 1/2$ problem:

$$X = \begin{pmatrix} 0 & 0 & 0 & \sqrt{\frac{3}{2}} & 0 & 1 \\ 0 & 0 & 1 & \frac{1}{2} & 1 & 0 \\ 0 & 1 & 0 & \frac{1}{2} & 1 & 0 \\ \sqrt{\frac{3}{2}} & \frac{1}{2} & \frac{1}{2} & 0 & \frac{1}{2} & \sqrt{\frac{3}{2}} \\ 0 & 1 & 1 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & \sqrt{\frac{3}{2}} & 0 & 0 \end{pmatrix}.$$

This solution is different to both the solution in [20] and the solution in [5]. A number of other solutions were also found.

Here is a X_+ corresponding to an infeasible X (again for the $t = 1/2$ problem):

$$X_+ = \begin{pmatrix} \frac{1}{2} & 0 & 0 & 0 & 0 & \sqrt{\frac{3}{2}} \\ 0 & 0 & \frac{7}{8} & \frac{7}{8} & \frac{7}{8} & 0 \\ 0 & \frac{7}{8} & 0 & \frac{7}{8} & \frac{7}{8} & 0 \\ 0 & \frac{7}{8} & \frac{7}{8} & 0 & \frac{7}{8} & 0 \\ 0 & \frac{7}{8} & \frac{7}{8} & \frac{7}{8} & 0 & 0 \\ \sqrt{\frac{3}{2}} & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The eigenvalues of this matrix are $\tilde{\lambda} = \{2\frac{5}{8}, 1\frac{1}{2}, -\frac{7}{8}, -\frac{7}{8}, -\frac{7}{8}, -1\}$.

6.2. NIEP. This subsection starts with some results for randomly generated NIEPs. Again, to ensure each problem is feasible, each desired spectrum is taken from a randomly generated matrix. Results are given in Table 6.3.

As can be seen, the results are again very good, with almost all problems solved.

The results indicate that NIEPs are harder to solve than SNIEPs. Also, the number of iterations, time, and time per iteration are greater. Part of the reason for an increase in time per iteration will be the extra computation required to calculate the least squares matching component of each $P_{\mathcal{M}}(U, T)$ calculation; see (4.6). (For

SNIEPs, the corresponding step is easy: the eigenvalues are real and just need to be sorted in decreasing order.)

For the NIEPs, both i and T increased with n.

The final problem we consider is taken from [8]. It is to find a stochastic matrix with (presumedly randomly generated) spectrum $\lambda = \{1.0000, -0.2608, 0.5046, 0.6438, -0.4483\}$. Furthermore the problem requires the matrix to have zeros in certain positions. In the context of Markov chains, we require the states to form a ring and that each state be linked to at most two immediate neighbors. The zero pattern is given by the zeros of the following matrix:

$$Z = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \end{pmatrix}. \tag{6.1}$$

Our algorithm as it stands is not able to solve this problem though it is able to do so if a simple modification is made. Using Z from (6.1), define

$$\tilde{\mathcal{N}} = \{A \in \mathbb{R}^{n\times n} \mid \ A_{ij} \geq 0 \text{ and } A_{ij} = 0 \text{ if } Z_{ij} = 0\}.$$

$\tilde{\mathcal{N}}$ is still a convex set. In the NIEP algorithm, replacing projection onto $\mathcal{N}$ by projection onto $\tilde{\mathcal{N}}$ gives solutions (nonnegative matrices) with zeros in the desired places. Using the transformation discussed in the introduction of the paper, solutions found by the algorithm can be converted into stochastic matrices with the same spectrum. Note that this transformation preserves zeros.

Using this methodology readily produced many solutions. An example is

$$X = \begin{pmatrix} 0.6931 & 0.2887 & 0 & 0 & 0.0182 \\ 0.1849 & 0.2422 & 0.5729 & 0 & 0 \\ 0 & 0.5476 & 0.3622 & 0.0902 & 0 \\ 0 & 0 & 0.5437 & 0.1233 & 0.3330 \\ 0.3712 & 0 & 0 & 0.6103 & 0.0185 \end{pmatrix}.$$

Another solution is

$$X = \begin{pmatrix} 0.8634 & 0.0431 & 0 & 0 & 0.0936 \\ 0.6224 & 0 & 0.3776 & 0 & 0 \\ 0 & 0.4935 & 0.1564 & 0.3501 & 0 \\ 0 & 0 & 0.1107 & 0.0115 & 0.8778 \\ 0.3452 & 0 & 0 & 0.2467 & 0.4080 \end{pmatrix}.$$

Notice that this latter solution has an extra zero. While this X still solves the problem, by further modifying $\mathcal{N}$ it is possible to ensure zeros appear only in the places specified by (6.1) and nowhere else.

For example, using

$$\tilde{\mathcal{N}} = \{A \in \mathbb{R}^{n\times n} \mid \ A_{ij} = 0 \text{ if } Z_{ij} = 0 \text{ and } A_{ij} \geq \delta \text{ otherwise}\},$$

with $\delta > 0$ a small constant, does the trick. Note that the stochastic matrix transformation leaves positive entries positive.

7. Conclusion. In this paper we have presented two related numerical methods, one for the NIEP, which can also be used to solve the inverse eigenvalue problem for stochastic matrices, and another for the SNIEP. The ideas used in the paper should also be applicable to many other inverse eigenvalue problems, including other problems involving nonsymmetric matrices.

8. Appendix. Local uniqueness and smoothness of projections. This appendix contains some supplementary results regarding projection onto the symmetric version of $\mathcal{M}$; see (3.2). While these results are not used in the main body of the paper, we believe they are interesting and worth mentioning. We would also expect them to be useful for other inverse eigenvalue problems.

THEOREM 8.1. [illegible] $\mathcal{M}$ [illegible] $\mathcal{M}$ [illegible].

[illegible]. To ease the presentation we will assume Λ has only two distinct eigenvalues. (The general case follows by similar reasoning.) Let m be such that $\lambda_m > \lambda_{m+1}$. If $Z \in \mathcal{M}$, then there exists a neighborhood of Z such that each matrix in this neighborhood has distinct mth and $m+1$th (ordered) eigenvalues. Suppose Y is an element in this neighborhood with eigenvalues $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_n$. It follows from the proof of Theorem 3.2 that if X is a projection of Y onto $\mathcal{M}$, then there exist orthonormal vectors $u_1, \ldots, u_n$ such that $Y = \sum_{i=1}^n \tilde{\lambda}_i u_i u_i^T$ and $X = \lambda_1 \sum_{i=1}^m u_i u_i^T + \lambda_{m+1} \sum_{i=m+1}^n u_i u_i^T$. Note that here we have used the fact that $\tilde{\lambda}_m$ and $\tilde{\lambda}_{m+1}$ are distinct. The proof will be complete if we can show X does not depend on the particular decomposition of Y.

Suppose eigenvalue $\tilde{\lambda}_j$ has multiplicity k with $\tilde{\lambda}_j = \cdots = \tilde{\lambda}_{j+k-1}$. If $\hat{u}_j, \ldots, \hat{u}_{j+k-1}$ is another set of orthonormal vectors that span the eigenspace corresponding to $\tilde{\lambda}_j$, then there exists a $k \times k$ orthogonal matrix Θ such that $[\hat{u}_j, \ldots, \hat{u}_{j+k-1}] = [u_j, \ldots, u_{j+k-1}]\Theta$. Consequently,

$$\sum_{i=j}^{j+k-1} \hat{u}_i \hat{u}_i^T = \sum_{i=j}^{j+k-1} u_i u_i^T.$$

The separation of eigenvalues implies the indices $j, \ldots, j+k-1$ are all either less than or equal to m, or, greater than or equal to $m+1$. It follows that X does not depend on the particular decomposition of Y. □

THEOREM 8.2. [illegible] $\mathcal{M}$ [illegible] $\mathcal{M}$ [illegible].

[illegible] (Outline). $\mathcal{M}$ is a submanifold of $\mathcal{S}^n$ and hence each point in $\mathcal{M}$ is in the image of a local parametrization of $\mathcal{M}$. The result can be shown to hold locally by using such a parametrization, using a condition necessary for a point to be a projection (if X is a projection of Y, then $X - Y$ is normal to the tangent space of $\mathcal{M}$ at X), and employing the implicit function theorem. In trying to satisfy the conditions of the implicit function theorem, the requirement that points being projected are sufficiently close to $\mathcal{M}$ appears.

As a consequence of the above mentioned necessary condition not being sufficient, for the proof to work it appears to be that it must be known a priori that in a neighborhood with unique projections, the projection operation is continuous. This is indeed the case as can be shown via a contradiction argument. □

The proof of the above theorem does not use any properties of $\mathcal{M}$ aside from the fact that it is a (closed) submanifold and that, near the set, projections are unique. (The projection result that uniqueness implies continuity holds for projections onto any closed set.) Hence, Theorem 8.2 also holds for any set with these two properties.

Acknowledgment. The author would like to thank Uwe Helmke for bringing the problems in this paper to his attention.

REFERENCES

[1] H. H. Bauschke and J. M. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Rev., 38 (1996), pp. 367–426.
[2] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979. Also published as Classics in Appl. Math. 9, SIAM, Philadelphia, 1994.
[3] L. M. Brègman, *The method of successive projection for finding a common point of convex sets*, Soviet Mathematics, 6 (1965), pp. 688–692.
[4] M. T. Chu, *Inverse eigenvalue problems*, SIAM Rev., 40 (1998), pp. 1–39.
[5] M. T. Chu and K. R. Driessel, *Constructing symmetric nonnegative matrices with prescribed eigenvalues by differential equations*, SIAM J. Math. Anal., 22 (1991), pp. 1372–1387.
[6] M. T. Chu and G. H. Golub, *Structured inverse eigenvalue problems*, Acta Numer., 11 (2002), pp. 1–71.
[7] M. T. Chu and G. H. Golub, *Inverse Eigenvalue Problems: Theory, Algorithms, and Applications*, Oxford University Press, Oxford, 2005.
[8] M. T. Chu and Q. Guo, *A numerical method for the inverse stochastic spectrum problem*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 1027–1039.
[9] M. T. Chu and S. F. Xu, *On computing minimal realizable spectral radii of non-negative matrices*, Numer. Linear Algebra Appl., 12 (2005), pp. 77–86.
[10] P. L. Combettes and H. J. Trussell, *Method of successive projections for finding a common point of sets in metric spaces*, J. Optim. Theory Appl., 67 (1990), pp. 487–507.
[11] F. Deutsch, *Best Approximation in Inner Product Spaces*, Springer-Verlag, New York, 2001.
[12] P. D. Egleston, T. D. Lenker, and S. K. Narayan, *The nonnegative inverse eigenvalue problem*, Linear Algebra Appl., 379 (2004), pp. 475–490.
[13] G. H. Hardy, J. E. Littlewood, and G. Polya, *Inequalities*, Cambridge University Press, Cambridge, 1952.
[14] U. Helmke and J. B. Moore, *Optimization and Dynamical Systems*, Springer-Verlag, London, 1994.
[15] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.
[16] C. R. Johnson, T. J. Laffey, and R. Loewy, *The real and the symmetric nonnegative inverse eigenvalue problems are different*, Proc. Amer. Math. Soc., 124 (1996), pp. 3647–3651.
[17] D. Luenberger, *Optimization by Vector Space Methods*, Wiley, New York, 1969.
[18] H. Minc, *Nonnegative matrices*, Wiley, New York, 1988.
[19] H. Perfect, *Methods of constructing certain stochastic matrices*, Duke Math. J., 20 (1953), pp. 395–404.
[20] G. W. Soules, *Constructing symmetric nonnegative matrices*, Linear and Multilinear Algebra, 13 (1983), pp. 241–251.
[21] K. R. Suleimanova, *Stochastic matrices with real characteristic numbers*, Soviet Math. Dokl., 66 (1949), pp. 343–345. (In Russian).
[22] S. F. Xu, *An Introduction to Inverse Algebraic Eigenvalue Problems*, Peking University Press, Beijing, and Vieweg & Sohn, Braunschweig, Germany, 1998.
[23] K. Yang and R. Orsi, *Pole placement via output feedback: a methodology based on projections*, in Proceedings of the 16th IFAC World Congress, Prague, Czech Republic, 2005.
[24] K. Yang, R. Orsi, and J. B. Moore, *A projective algorithm for static output feedback stabilization*, in Proceedings of the 2nd IFAC Symposium on System, Structure and Control, Oaxaca, Mexico, 2004, pp. 263–268.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 1, pp. 213–228

STRUCTURES PRESERVED BY MATRIX INVERSION*

STEVEN DELVAUX† AND MARC VAN BAREL†

Abstract. In this paper we investigate some matrix structures on $\mathbb{C}^{n\times n}$ that have a good behavior under matrix inversion. The first type of structure is closely related to low displacement rank matrices. Next, we show that for a matrix having a low rank submatrix, the inverse matrix also must have a low rank submatrix, which we can explicitly determine. This allows us to generalize a theorem due to Fiedler and Markham. The generalization consists in the fact that our rank structures may have a certain correction term, which we call the shift matrix $\Lambda_k \in \mathbb{C}^{m\times m}$, for suitable m, and with Fiedler and Markham's theorem corresponding to the limiting cases $\Lambda_k \to 0$ and $\Lambda_k \to \infty I$.

Key words. displacement structures, Hermitian plus low rank, rank structures, lower semiseparable (plus diagonal) matrices, matrix inversion

AMS subject classifications. 15A09, 15A03, 65F05

DOI. 10.1137/040621429

1. Introduction.

The aim of this paper is to handle structures of a matrix $A \in \mathbb{C}^{n\times n}$ that carry over to its inverse.

Section 2 deals with the inversion of displacement structures. The idea is to generalize the classical examples of displacement structures (for example, Toeplitz-like, Cauchy-like, Vandermonde-like, circulant matrices; see [13]), by decoupling the displacement equation. This means that the displacement equation is allowed to involve two variables A and B rather than only one variable A. We will then illustrate these decoupled displacement structures by some examples; one of these examples involves Hermitian plus low rank matrices, for which we provide an alternative characterization. The results of this section can be easily derived from the well-known results in classical low displacement rank matrix theory.

Section 3 handles the inversion of [illegible]. A rank structure on $\mathbb{C}^{n\times n}$ will be defined as a collection of structure blocks $\{\mathcal{B}_k\}_k$: these are low rank submatrices which lie in the bottom left corner of a given matrix $A \in \mathbb{C}^{n\times n}$, together with a certain correction term for the block diagonal positions of A, called the shift matrix. Suppose then that $\mathcal{B}$ is a structure block according to the above definition. It turns out that there can be defined in a natural way an inverse structure block $\mathcal{B}^{-1}$, by just replacing the shift matrix Λ by Λ^{-1} (assuming Λ is nonsingular).

Section 4 handles some generalizations of section 3. A first generalization is the inversion of structure blocks whose shift matrix Λ is singular. To solve this problem, we make a reduction to the case where the shift matrix is block diagonal of the form $\Lambda = \Lambda_{\text{ns}} \oplus 0$, with Λ_{ns} nonsingular. Then it could be expected that the inverse of the shift matrix is block diagonal of the form $\Lambda^{-1} = \Lambda_{\text{ns}}^{-1} \oplus \infty I$, where ∞I is the diagonal

*Received by the editors December 14, 2004; accepted for publication (in revised form) by L. Reichel November 14, 2005; published electronically March 17, 2006. This research was partially supported by the Research Council K.U. Leuven projects OT/00/16 and OT/05/40; by the Fund for Scientific Research–Flanders (Belgium) projects G.0078.01, G.0176.02, G.0184.02, and G.0455.0; and by the Belgian Program on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture, project IUAP V-22. The scientific responsibility rests with the authors.

http://www.siam.org/journals/simax/28-1/62142.html

†Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, B-3001 Leuven (Heverlee), Belgium (Steven.Delvaux@cs.kuleuven.ac.be, Marc.VanBarel@cs.kuleuven.ac.be.)

matrix with diagonal entries "equal to ∞"; we will show that there can be given an exact meaning to this statement.

A second generalization is to absorb permutation matrices into the structure. This allows us to move the structure blocks to an arbitrary matrix position, not necessarily situated in the bottom left matrix corner anymore. Moreover, in the case of shift matrix $\Lambda \to 0$ or $\Lambda \to \infty I$, we obtain in this way an alternative derivation of a theorem due to Fiedler and Markham [6], which we state now.

DEFINITION 1. [illegible] right nullity [illegible] nullity Null A [illegible] $A \in \mathbb{C}^{m\times n}$ [illegible] A [illegible] left nullity [illegible] A [illegible] rank defect [illegible] A [illegible]

THEOREM 2 (see [6]). [illegible] $A \in \mathbb{C}^{n\times n}$ [illegible] I [illegible] J [illegible]

$$\text{Null } A^{-1}(I,J) = \text{Null } A(N\setminus J, N\setminus I), \tag{1}$$

[illegible] $N := \{1,2,\dots,n\}$ [illegible] (1) [illegible] Null [illegible]

It should be emphasized that Fiedler and Markham's theorem will be just a special case of the theory. Moreover, there are several connections between sections 2, 3, and 4, in the sense that the main results for rank structures can be derived as a special case of the decoupled displacement structures of section 2, as will become clear soon.

2. Displacement structures. In this section we handle displacement structure. As a general reference, we can refer to [13, 12, 14] for an overview of the many applications of displacement theory in numerical linear algebra. Some references of historical interest are [10, 11]. For our purposes, however, we will only be interested in matrix inversion.

Let us start with some classical examples of displacement structure. Let A be a Toeplitz matrix, i.e., $A = [a_{i-j}]_{i,j=1}^{n}$. Putting $Z := [\mathbf{e}_2 \dots \mathbf{e}_n \mathbf{0}]$ with $\mathbf{e}_k$ the kth column of the identity matrix, it is easy to check that

$$A - ZAZ^T = \text{Rk } 2 \tag{2}$$

with Rk 2 denoting a matrix of rank at most 2. Therefore, Toeplitz matrices can be embedded in the class of [illegible] matrices, i.e., the class of matrices A satisfying (2).

Let A be a Cauchy matrix, i.e., $A = [\frac{1}{x_i - y_j}]_{i,j=1}^{n}$. It is easy to check that

$$D_{\mathbf{x}}A - AD_{\mathbf{y}} = \text{Rk } 1 \tag{3}$$

with Rk 1 denoting a matrix of rank at most 1. (Here we are using the notation $D_{\mathbf{x}} = \text{diag}(x_i)_i$ for any vector $\mathbf{x}$). Therefore, Cauchy matrices can be embedded in the class of [illegible] matrices, i.e., the class of matrices A satisfying (3).

Now we come to some formal definitions. The main difference with (2) and (3) is that the variable A is decoupled into two variables A and B.

DEFINITION 3. [illegible] $G, H \in \mathbb{C}^{m\times n}$ [illegible] $r \in \mathbb{N}$ [illegible] $A \in \mathbb{C}^{m\times m}$ [illegible] $B \in \mathbb{C}^{n\times n}$

1. [illegible] A [illegible] B [illegible] Stein type displacement equation [illegible] (G,H,r) [illegible]

$$A - GBH^T = \text{Rk } r, \tag{4}$$

[illegible] Rk r [illegible] r

2. A B Sylvester type displacement equation (G, H, r)

$$AG - HB = \text{Rk } r, \tag{5}$$

Rk r r

Note that the dimension requirements in Definition 3 are equivalent to saying that A and B are square and the block matrix

$$\begin{bmatrix} B^{-1} & H^T \\ G & A \end{bmatrix} \tag{6}$$

has compatible matrix dimensions. (Here we assumed that B^{-1} exists). Moreover, this block representation is useful in the sense that the left-hand side of (4) can be realized as a Schur complement in (6). This is the basis of the following inversion result.

THEOREM 4 (Stein type inversion). 3 A B

$$A - GBH^T = \text{Rk } r, \tag{7}$$

$$B^{-1} - H^T A^{-1} G = \widetilde{\text{Rk}}\, \tilde{r}, \tag{8}$$

$\widetilde{\text{Rk}}\, \tilde{r}$ $\tilde{r} := r + n - m$ (based on [13, Lemma 1.5.1]). Let us denote $A \sim B$ if these matrices can be obtained out of each other by elementary Gaussian row and/or column operations. Consider the embedded matrix (6). By the nonsingularity of B^{-1}, we can use it as pivot block for a Gaussian elimination process, and the Schur complement formula yields

$$\begin{bmatrix} B^{-1} & H^T \\ G & A \end{bmatrix} \sim \begin{bmatrix} B^{-1} & 0 \\ 0 & A - GBH^T \end{bmatrix}. \tag{9}$$

Similarly, by the nonsingularity of A, we can use it also as pivot block and obtain

$$\begin{bmatrix} B^{-1} & H^T \\ G & A \end{bmatrix} \sim \begin{bmatrix} B^{-1} - H^T A^{-1} G & 0 \\ 0 & A \end{bmatrix}. \tag{10}$$

The proof can then be finished by comparing (9) and (10) and by using the fact that $\sim$ is a rank-preserving relation. □

5 (displacement nullity). Note that the above theorem states that the displacement rank should be corrected by the quantity $n - m$ under matrix inversion. But since this quantity equals precisely the difference in size between the matrices (7) and (8), it follows that these matrices must have the same nullity (Definition 1). In other words, the quantity that is strictly preserved under matrix inversion is not the displacement rank but is rather the

We come to the second, easier inversion result.

THEOREM 6 (Sylvester type inversion). 3 A B

$$AG - HB = \text{Rk } r, \tag{11}$$

(12) $$GB^{-1} - A^{-1}H = \widetilde{\mathrm{Rk}}\, r,$$

$\widetilde{\mathrm{Rk}}\, r := A^{-1}(\mathrm{Rk}\, r)B^{-1}$. This follows immediately by multiplying (11) on the left with A^{-1} and on the right with B^{-1}. □

7. The above proof remains valid if we add a quadratic and constant term to the structure, i.e., if we replace (11) by

$$AFB + AG - HB + J = \mathrm{Rk}\, r$$

and (12) by

$$F + GB^{-1} - A^{-1}H + A^{-1}JB^{-1} = \widetilde{\mathrm{Rk}}\, r$$

for certain $F, G, H, J \in \mathbb{C}^{m\times n}$. Thus we see that the quadratic term AFB has been transformed into the constant term F under matrix inversion, and vice versa.

Let us illustrate why it is useful to decouple the variable A into two variables A and B. As an illustrative example, we will focus on the following displacement equation.

COROLLARY 8. A B

(13) $$A - B = \mathrm{Rk}\, r;$$

(14) $$A^{-1} - B^{-1} = \widetilde{\mathrm{Rk}}\, r.$$

This corollary is a special case of both the Stein and Sylvester type inversion results. Hence the matrix $\widetilde{\mathrm{Rk}}\, r$ occurring in (14) is given explicitly by $-A^{-1}(\mathrm{Rk}\, r)B^{-1}$ or $-B^{-1}(\mathrm{Rk}\, r)A^{-1}$, by the proof of Theorem 6. Yet another explicit formula for $\widetilde{\mathrm{Rk}}\, r$ is the so-called Sherman–Morrison formula [8, section 2.1], stating that

$$\widetilde{\mathrm{Rk}}\, r = -B^{-1}U(V^TB^{-1}U + I)^{-1}V^TB^{-1},$$

where we assumed a decomposition $\mathrm{Rk}\, r = UV^T$ with $U, V \in \mathbb{C}^{n\times r}$. Note that this formula does not involve A^{-1} anymore.

Now we come to some illustrations of Corollary 8. First we can take B to be unitary. Note that the inverse matrix B^{-1} is again unitary; hence Corollary 8 reveals the following fact

COROLLARY 9. $A = \mathrm{Uni} + \mathrm{Rk}\, r$ A r

We could state a similar property for the property $A = \mathrm{Herm} + \mathrm{Rk}\, r$, i.e., A is Hermitian plus rank at most r. But here we encounter the problem that for a nonsingular matrix $A = \mathrm{Herm} + \mathrm{Rk}\, r$, the Hermitian component Herm does not necessarily have to be nonsingular too. This means that the nonsingularity conditions in Corollary 8 will not always be satisfied. Nevertheless, it turns out that the preservation of Hermitian plus low rank structure under inversion will still be valid, even if this nonsingularity assumption is not satisfied.

To show this, we will give an alternative characterization of the property $A = \mathrm{Herm} + \mathrm{Rk}\, r$.

We recall the following definition.

DEFINITION 10. $A, B \in \mathbb{C}^{n\times n}$ congruent $T \in \mathbb{C}^{n\times n}$ $A = TBT^H$ Herm Inertia(Herm) 3 (π, ν, ζ) π ν ζ Herm Herm

Inertia and congruence are classical tools in linear algebra. For example, states that Inertia(A) = Inertia(B) if and only if A and B are congruent. Moreover, inertia is in the sense that for all Hermitian matrices A and B, we have $\pi(A+B) \le \pi(A) + \pi(B)$ and $\nu(A+B) \le \nu(A) + \nu(B)$: see, for example, [9, Lemma 2] for an easy proof of this property. (Note that by adding these two equations, we get the well-known property of subadditivity of .)

Now we prove the next theorem.

THEOREM 11.

(i) $A = \text{Herm} + \text{Rk } r$ A r.

(ii) $i(A - A^H) = \text{Rk } 2r$ $i := \sqrt{-1}$ Rk $2r$ $2r$ Inertia(Rk $2r$) $= (\pi, \nu, \zeta)$ $\max\{\pi, \nu\} \le r$

(a) First we prove the implication (i) $\Rightarrow$ (ii). Thus let us assume that $A = \text{Herm} + \text{Rk } r = \text{Herm} + \sum_{k=1}^{r} \mathbf{u}_k \mathbf{v}_k^H$ for suitable column vectors $\mathbf{u}_k, \mathbf{v}_k \in \mathbb{C}^n$. The matrix Herm can be eliminated by considering

$$i(A - A^H) = \sum_{k=1}^{r} i(\mathbf{u}_k \mathbf{v}_k^H - \mathbf{v}_k \mathbf{u}_k^H). \tag{15}$$

We will first prove the theorem for rank upper bound $r = 1$. Thus we will prove that for any $\mathbf{u}, \mathbf{v} \in \mathbb{C}^n$,

$$\text{Inertia}(i(\mathbf{u}\mathbf{v}^H - \mathbf{v}\mathbf{u}^H)) = (\pi, \nu, \zeta) \quad \text{with } \max\{\pi, \nu\} \le 1. \tag{16}$$

To prove this, let us write

$$i(\mathbf{u}\mathbf{v}^H - \mathbf{v}\mathbf{u}^H) = \begin{bmatrix} \mathbf{u} & \mathbf{v} \end{bmatrix} \begin{bmatrix} 0 & i \\ -i & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} & \mathbf{v} \end{bmatrix}^H. \tag{17}$$

If the columns $\mathbf{u}$, $\mathbf{v}$ in (17) are linearly dependent, then $i(\mathbf{u}\mathbf{v}^H - \mathbf{v}\mathbf{u}^H)$ has rank at most one, and hence (16) must obviously be true. If the columns $\mathbf{u}$, $\mathbf{v}$ are independent, then from (17), Sylvester's law of inertia implies that Inertia$(i(uv^H - vu^H))$ = Inertia$(\left[\begin{smallmatrix} 0 & i \\ -i & 0 \end{smallmatrix}\right])$ (completed with $n-2$ zero eigenvalues). But the latter 2-by-2 matrix has determinant -1, and hence it has exactly one positive and one negative eigenvalue. This proves (16), i.e., the theorem has been proven now for rank upper bound $r = 1$.

In the general case $r \ge 1$, it follows from (15) that the matrix $i(A - A^H)$ is the sum of all the $i(\mathbf{u}_k \mathbf{v}_k^H - \mathbf{v}_k \mathbf{u}_k^H)$, $k = 1, \ldots, r$. The theorem follows then by (16) and the subadditivity of inertia.

(b) Now we prove the implication (ii) $\Rightarrow$ (i). Thus let us assume that Inertia$(i(A - A^H)) = (\pi, \nu, \zeta)$, where $\max\{\pi, \nu\} \le r$. By symmetry, we may suppose that $\pi \ge \nu$. We define the block diagonal matrix

$$D = \begin{bmatrix} 0 & i \\ -i & 0 \end{bmatrix} \oplus \ldots \oplus \begin{bmatrix} 0 & i \\ -i & 0 \end{bmatrix} \oplus 1 \oplus \ldots \oplus 1, \tag{18}$$

where the first factor occurs precisely ν times and the second factor $\pi - \nu$ times. By construction, D has exactly the same inertia as $i(A - A^H)$ (except for a loss of zero eigenvalues). By Sylvester's law of inertia, there exists a maximal column rank matrix T such that

$$i(A - A^H) = TDT^H. \tag{19}$$

Now we define vectors $\mathbf{u}_k, \mathbf{v}_k$ by setting $T =: [\mathbf{u}_1, \mathbf{v}_1; \ldots; \mathbf{u}_\nu, \mathbf{v}_\nu; \mathbf{u}_{\nu+1}; \ldots; \mathbf{u}_\pi]$ and $\mathbf{v}_k := \frac{i}{2}\mathbf{u}_k$ for $k = \nu + 1, \ldots, \pi$. It is easy to check that

$$\sum_{k=1}^{\pi} i(\mathbf{u}_k \mathbf{v}_k^H - \mathbf{v}_k \mathbf{u}_k^H) = TDT^H. \tag{20}$$

Together with (19), (20) shows that the matrix $A - \sum_{k=1}^{\pi} \mathbf{u}_k \mathbf{v}_k^H$ must be Hermitian. Since moreover $\pi = \max\{\pi, \nu\} \leq r$, this yields us a decomposition $A = \text{Herm} + \text{Rk } r$, hence finishing the proof. ∎

From this theorem, we can derive some further properties of Hermitian plus rank at most r matrices, including an alternative (and complete) proof of their preservation under inversion.

COROLLARY 12. [illegible] $r \in \mathbb{N}$ [illegible]

1. [illegible] $\{A \in \mathbb{C}^{n\times n} |\ A = \text{Herm} + \text{Rk } r\}$ [illegible]
2. [illegible] $\{A \in \mathbb{C}^{n\times n} |\ A = \text{Herm} + \text{Rk } r\}$ [illegible]
3. [illegible] A [illegible] 11 [illegible] $i(A - A^H) = \text{Rk } 2r$ [illegible] Inertia(Rk $2r$) $= (\pi, \nu, \zeta)$ [illegible] $\max\{\pi, \nu\} \leq r$

[illegible]

1. Obviously, for a family of matrices $A_\epsilon \in \mathbb{C}^{n\times n}$, $\epsilon \in \mathbb{C} \setminus \{0\}$ with $\lim_{\epsilon\to 0} A_\epsilon = A \in \mathbb{C}^{n\times n}$, the property Inertia$(i(A_\epsilon - A_\epsilon^H)) = (\pi, \nu, \zeta)$ with $\max\{\pi, \nu\} \leq r$ cannot be lost for the limiting matrix $A = \lim_{\epsilon\to 0} A_\epsilon$. We can then conclude by Theorem 11.
2. We use again Theorem 11. Thus let A be a matrix satisfying $i(A - A^H) = \text{Rk } 2r$ and Inertia(Rk $2r$) $= (\pi, \nu, \zeta)$ with $\max\{\pi, \nu\} \leq r$. From Corollary 8 (or by direct verification), we obtain that
$$i(A^{-1} - A^{-H}) = -\widetilde{\text{Rk}}\ 2r, \tag{21}$$
where $\widetilde{\text{Rk}}\ 2r := A^{-1}(\text{Rk } 2r)A^{-H}$. Since Rk $2r$ and $\widetilde{\text{Rk}}\ 2r$ are congruent, by Sylvester's law of inertia they have the same inertia. In particular, the property $\max\{\pi, \nu\} \leq r$ must carry over to $\widetilde{\text{Rk}}\ 2r$ and hence to $-\widetilde{\text{Rk}}\ 2r$ in (21).
3. It can be easily checked that for a [illegible] matrix A, the eigenvalues of the matrix $i(A - A^H)$ always come in (real) pairs $\lambda, -\lambda$. Hence from $i(A - A^H) = \text{Rk } 2r$ it automatically follows that $\pi = \nu = \max\{\pi, \nu\} \leq r$, leading to the desired simplification of Theorem 11. ∎

3. Inversion of rank structures. In this section we handle the inversion of rank structures. These rank structures are a generalization of those in [2]. We start with the following definition.

DEFINITION 13. [illegible] rank structure [illegible] $\mathbb{C}^{n\times n}$ [illegible] $\mathcal{R} = \{\mathcal{B}_k\}_k$ [illegible] structure block $\mathcal{B}_k$ [illegible] 4 [illegible]

$$\mathcal{B}_k = (i_k, j_k, r_k, \Lambda_k),$$

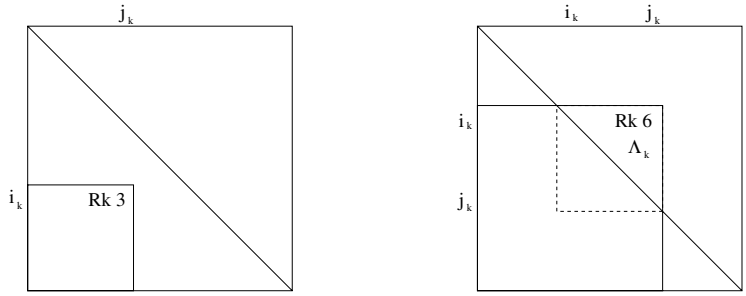


FIGURE 1. *The structure block in the left figure has* $j_k - i_k + 1 \leq 0$ *and is* pure. *The structure block in the right figure has the following meaning: after subtracting the shift matrix* $\Lambda_k \in \mathbb{C}^{4\times 4}$ *from the dashed square submatrix in the middle, the indicated bottom left submatrix must be of rank at most* 6.

i_k j_k r_k $\Lambda_k \in \mathbb{C}^{(j_k-i_k+1)\times(j_k-i_k+1)}$ $\mathcal{B}_k$ $A \in \mathbb{C}^{n\times n}$ k k

$$A =: \begin{bmatrix} A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,1} & A_{3,2} & A_{3,3} \end{bmatrix}_k, \tag{22}$$

$A_{2,2}$ $i_k, \ldots, j_k$

$$\begin{bmatrix} A_{2,1} & A_{2,2} - \Lambda_k \\ A_{3,1} & A_{3,2} \end{bmatrix}_k = \text{Rk } r_k, \tag{23}$$

r_k

$\mathcal{B}_k$ pure structure block $\Lambda_k = 0$ $\mathcal{B}_{\text{pure},k}$ $j_k - i_k + 1 \leq 0$ $\mathcal{B}_k$ 1 (23) $j_k - i_k + 1 < 0$ $\begin{bmatrix} A_{3,1} \end{bmatrix}_k = \text{Rk } r_k$ $A_{3,1}$ $i_k, \ldots, n$ $1, \ldots, j_k$

In case $\Lambda_k = \lambda_k I$ for certain $\lambda_k \in \mathbb{C}$, Definition 13 leads to the structure blocks that were studied in [2]. We proved there that these structure blocks are preserved by the shifted QR-algorithm. For the present paper, we do not need to make this restriction on the shift matrices Λ_k.

Let us illustrate Definition 13. We can use it to describe Hessenberg matrices: $\mathcal{B}_{\text{pure},k} = (k+2, k, 0, 0)$, $k = 1, \ldots, n-2$; upper triangular matrices; lower semiseparable matrices: $\mathcal{B}_{\text{pure},k} = (k, k, 1, 0)$, $k = 1, \ldots, n$; lower semiseparable plus diagonal matrices: $\mathcal{B}_k = (k, k, 1, \lambda_k)$, $k = 1, \ldots, n$. Note that for this last example, the shift matrices $\Lambda_k =: \lambda_k \in \mathbb{C}$ are scalar. Also for upper triangular matrices, we could absorb the diagonal elements λ_k into the structure, if we would want to. Of course, there are also more general or "chaotic" rank structures than the ones that we just mentioned.

Concerning matrix inversion, let us first indicate why we should expect a positive result. Some well-known examples are the following: upper triangular structure is preserved under matrix inversion; lower semiseparable matrices and Hessenberg matrices are each others inverses; the of lower semiseparable plus diagonal matrices, i.e., $\mathcal{B}_{\text{pure},k} = (k+1, k, 1, 0)$, $k = 1, \ldots, n-1$, is preserved under matrix inversion.

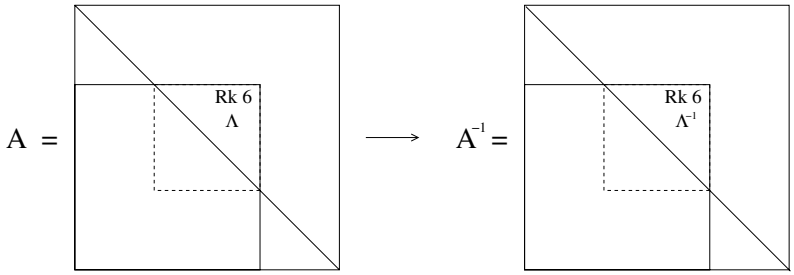


FIGURE 2. *Under matrix inversion, the shift matrix* Λ *of the* Rk 6 *structure block will transform into the new shift matrix* Λ^{-1}. *The rank number itself is preserved.*

Having a second look at these examples, we can note that for upper triangular matrices, the diagonal elements λ_k are precisely transformed into $1/\lambda_k$ under matrix inversion. The same holds for the shift elements λ_k of a lower semiseparable plus diagonal matrix, as we will show. (See also the notes in Remark 15.4.) Since it is sufficient to consider the behavior of a single structure block $\mathcal{B}_k$, from now on we will put $\mathcal{B}_k =: \mathcal{B}$ and drop the index k.

THEOREM 14. [illegible] $A \in \mathbb{C}^{n\times n}$ [illegible] $\mathcal{B} = (i, j, r, \Lambda)$ [illegible] Λ [illegible] $j-i+1 \geq 0$ [illegible] $\mathcal{B}$ [illegible] A [illegible] $\mathcal{B}^{-1} := (i, j, r, \Lambda^{-1})$. [illegible] 2 [illegible]. Making a partition of A as in (22), it follows that $A - B = \text{Rk } r$ where

$$B = \begin{bmatrix} A_{1,1} & A_{1,2} & A_{1,3} \\ 0 & \Lambda & A_{2,3} \\ 0 & 0 & A_{3,3} \end{bmatrix}.$$

Then by (14), or by direct verification, it follows that $A^{-1} - B^{-1} = \widetilde{\text{Rk}}\, r$. By the form of B, this implies that A^{-1} will satisfy the inverse structure block $\mathcal{B}^{-1}$, and the theorem is proved.

Note that we assumed here implicitly that the matrix B is nonsingular, i.e., that the square blocks $A_{1,1}$ and $A_{3,3}$ are nonsingular. The case where this nonsingularity is not satisfied can be handled by a continuity argument. (We add an infinitesimally small correction to the $A_{1,1}$ and $A_{3,3}$ blocks, which will not influence the given structure block). □

[illegible] 15.

1. Note that the above theorem has been stated in terms of [illegible] r. We proved in fact that the structure block $\mathcal{B}^{-1}$ can have at most the same rank as the original structure block $\mathcal{B}$, i.e., $r_{\mathcal{B}^{-1}} \leq r_{\mathcal{B}}$. By applying the same argument to $\mathcal{B}^{-1}$, we obtain also the inequality $r_{\mathcal{B}^{-1}} \geq r_{\mathcal{B}}$, and it follows that the ranks must be [illegible] the same. A similar remark could have been made also for the theorems in section 2.
2. Since the above proof was based on the equality $A - B = \text{Rk } r$, one could use the Sherman–Morrison formula (mentioned above) to compute the inverse matrix. The problem is that we worked here with a single structure block, and that the formula may not always be applicable in case of several structure blocks.
3. As stated in the formulation of the theorem, the above proof works also for the case where Λ is lying just below the main diagonal of A, i.e., for the case

where $j-i+1=0$. One should then consider Λ and Λ^{-1} as empty matrices. Structure blocks with $j-i+1=0$ have already received a lot of attention in the literature, from the point of view of [illegible]; see, for example, [3, 4].

4. The inheritance of lower semiseparable plus diagonal structure under matrix inversion was also proved in [5, Theorem 4.1] and in a more general block context in [7, Theorem 3.1]. Here we should note that the structures of these authors are slightly different from ours, in the sense that they are [illegible] based on structure blocks. For example, the paper [7] considers mp by mp matrices of the form $D+S$, where D and S are block matrices consisting of p by p blocks, D is nonsingular and block diagonal, and S has its block lower triangular part equal to that of a rank r matrix. We mention that this paper gives explicit inversion formulas.

4. Inversion of rank structures: Some extensions. In this section we will extend Theorem 14 in several directions.

4.1. Singular shift matrices. In the statement of Theorem 14, it was assumed that the shift matrix Λ is nonsingular. We will now remove this condition.

The first step is to construct unitary matrices U and V to bring the shift matrix in block diagonal form

$$U^H \Lambda V = \Lambda_{\text{ns}} \oplus 0 \tag{24}$$

with Λ_{ns} nonsingular. This matrix decomposition can be considered as an "incomplete singular value decomposition." The word [illegible] means that we are only concerned with transforming the dependent rows and columns of Λ into zeros, which is a relatively easy operation.

Now by (24), we have that $(I \oplus V^H \oplus I)A^{-1}(I \oplus U \oplus I)$ satisfies the structure block

$$\mathcal{B}^{-1}, \quad \text{where } \mathcal{B} = (i, j, r, \Lambda_{\text{ns}} \oplus 0).$$

So the structure of A^{-1} is known as soon as we know the structure block $\mathcal{B}^{-1}$. Hence from now on, we will suppose that $\Lambda = \Lambda_{\text{ns}} \oplus 0$ with Λ_{ns} nonsingular.

By continuity reasons, we could then expect that $\Lambda^{-1} = \Lambda_{\text{ns}}^{-1} \oplus \infty I$, where ∞I is the diagonal matrix with diagonal entries equal to ∞.

DEFINITION 16. [illegible] $\mathcal{B} = (i, j, r, \Lambda)$ [illegible] $\mathbb{C}^{n\times n}$ [illegible] $\Lambda = \Lambda_{\text{fin}} \oplus \infty I$ [illegible] $\Lambda_{\text{fin}} \in \mathbb{C}^{m\times m}$ [illegible] m [illegible] $\mathcal{B}$ [illegible] [illegible] ∞ [illegible] [illegible] r [illegible] [illegible] 3

Formally, we will use $\mathbb{C} \cup \{\infty\}$ to denote the one-point-compactification of $\mathbb{C}$. Thus by definition, we have $x_\epsilon \to \infty$ if and only if the moduli $|x_\epsilon| \to \infty$.

Of course we have to motivate Definition 16. Thus let us show that it is indeed the correct, i.e., continuous definition for shift matrices ∞I.

THEOREM 17. [illegible] i j r m_1 [illegible] m_2 [illegible] $j-i+1 = m_1+m_2$ [illegible] A_ϵ $\epsilon \in \mathbb{C}\setminus\{0\}$ [illegible]

F1. [illegible] A_ϵ [illegible] $\mathcal{B}_\epsilon = (i, j, r, \Lambda_\epsilon)$ [illegible] $\Lambda_\epsilon = \Lambda_{\epsilon,1} \oplus \Lambda_{\epsilon,2}$ [illegible] $\Lambda_{\epsilon,k} \in \mathbb{C}^{m_k\times m_k}$ $k = 1, 2$.

F2a. $\lim_{\epsilon\to 0} \Lambda_{\epsilon,1} =: \Lambda_{\text{fin}} \in \mathbb{C}^{m_1\times m_1}$ [illegible] $\lim_{\epsilon\to 0} \Lambda_{\epsilon,2} = \infty I$.

F2b. $\lim_{\epsilon\to 0} A_\epsilon =: A \in \mathbb{C}^{n\times n}$

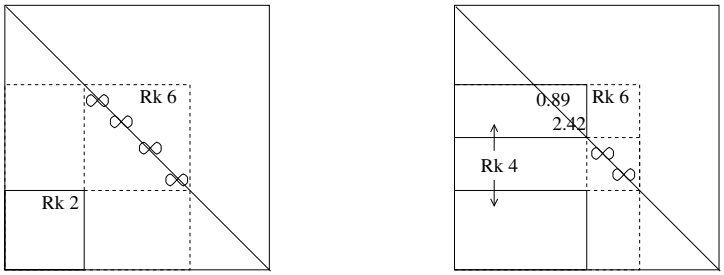


FIGURE 3. *Given the* Rk 6 *structure block* $\mathcal{B}$ *in the left picture, we have here* $\Lambda = \infty I_4$*, and thus by definition* $\mathcal{B}$ *should be identified with the smaller* Rk 2 *structure block in the bottom left corner. Next, consider the* Rk 6 *structure block* $\mathcal{B}$ *in the right picture. We have here* $\Lambda = \mathrm{diag}(0.89, 2.42, \infty, \infty)$*, and thus by definition* $\mathcal{B}$ *should be identified with the smaller* Rk 4 *structure block, consisting of two pieces. Note that the shift submatrix* $\mathrm{diag}(0.89, 2.42)$ *is inherited.*

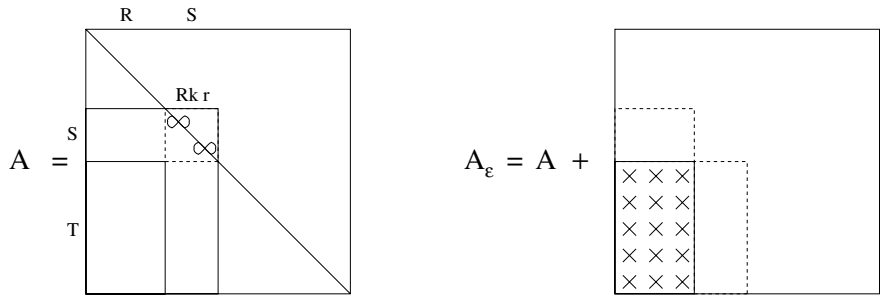


FIGURE 4. *Let* A *satisfy a structure block* $\mathcal{B}$ *with shift matrix* $\Lambda = \infty I$*. We can approximate* Λ *by finite shift matrices by adding an infinitesimally small correction term to* $A(T, R)$*. Note that, moreover, the row space of this correction term is contained in the row space of* $A(S, R)$ *and the column space is contained in the column space of* $A(T, S)$ *(see* (27)*).*

1. [illegible] A [illegible] $\mathcal{B} = (i, j, r, \Lambda)$ [illegible] $\Lambda = \Lambda_{\mathrm{fin}} \oplus \infty I$
2. [illegible] A [illegible] $\mathcal{B} = (i, j, r, \Lambda)$ [illegible] A_ϵ [illegible] F1 [illegible] F2

[illegible]. For simplicity, we restrict the proof to the "pure" case where

Assumption: $\Lambda = \infty I$, i.e., $m_1 = 0$ and $m_2 = j - i + 1$.

Moreover, we define index sets $R = \{1, \ldots, i-1\}$, $S = \{i, \ldots, j\}$, $T = \{j+1, \ldots, n\}$: this is the situation in the left picture of Figure 4.

1. First, suppose that the family A_ϵ satisfies conditions F1 and F2. Define a family $\tilde{A}_\epsilon$ from A_ϵ by setting

$$\tilde{A}_\epsilon(S, S) := A_\epsilon(S, S) - \Lambda_\epsilon.$$

Thus by F2, for $\epsilon \to 0$ the diagonal entries of $\tilde{A}_\epsilon(S, S)$ converge to ∞ while the off-diagonal entries converge to finite values. Hence it is easy to see that $\tilde{A}_\epsilon(S, S)^{-1}$ exists for all ϵ sufficiently small and satisfies

$$\lim_{\epsilon \to 0} \tilde{A}_\epsilon(S, S)^{-1} = 0. \tag{25}$$

Moreover, the nonsingularity of $\tilde{A}_\epsilon(S,S)$ allows us to use it as pivot block for a Gaussian elimination process. Then the Schur complement formula yields (we suppress the tilde whenever allowed)

$$\begin{aligned}(26)\quad &\text{Rank } \tilde{A}_\epsilon(S\cup T, R\cup S)\\ &\qquad = |S| + \text{Rank } \Big(A_\epsilon(T,R) - A_\epsilon(T,S)\tilde{A}_\epsilon(S,S)^{-1}A_\epsilon(S,R)\Big).\end{aligned}$$

But by F1, the left-hand side of (26) cannot exceed r, i.e.,

$$\begin{aligned} r - |S| &\geq \limsup_{\epsilon\to 0} \text{Rank } \Big(A_\epsilon(T,R) - A_\epsilon(T,S)\tilde{A}_\epsilon(S,S)^{-1}A_\epsilon(S,R)\Big)\\ &\geq \text{Rank } \lim_{\epsilon\to 0} \Big(A_\epsilon(T,R) - A_\epsilon(T,S)\tilde{A}_\epsilon(S,S)^{-1}A_\epsilon(S,R)\Big)\\ &= \text{Rank } A(T,R),\end{aligned}$$

where the last transition follows from (25). This shows that the limiting matrix A satisfies the structure block $\mathcal{B} = (i,j,r,\infty I)$.

2. Conversely, suppose that A is a matrix satisfying $\mathcal{B} = (i,j,r,\infty I)$. Define a matrix $\tilde{A}$ from A by setting $\tilde{A}(S,S) = A(S,S) - \frac{1}{\epsilon}I$ (considering ϵ as a symbol). Define a family A_ϵ, $\epsilon \in \mathbb{C}\setminus\{0\}$ from A by

$$(27)\qquad A_\epsilon(T,R) := A(T,R) + A(T,S)\tilde{A}(S,S)^{-1}A(S,R),$$

as in the right picture of Figure 4. Finally, define a family $\tilde{A}_\epsilon$ from A_ϵ by setting $\tilde{A}_\epsilon(S,S) := A_\epsilon(S,S) - \frac{1}{\epsilon}I$. Then for $\epsilon \to 0$ the diagonal entries of $\tilde{A}_\epsilon(S,S)$ converge to ∞ while the off-diagonal entries are just constant values. Hence the derivation of (25) and (26) remains valid. But now by the above definition of A_ϵ, the Schur complement in the right-hand side of (26) simplifies to $A_\epsilon(T,R) - A(T,S)\tilde{A}(S,S)^{-1}A(S,R) = A(T,R)$. Hence (26) collapses to

$$\begin{aligned}\text{Rank}(\tilde{A}_\epsilon(S\cup T, R\cup S)) &= |S| + \text{Rank } A(T,R)\\ &\leq |S| + (r - |S|)\\ &= r,\end{aligned}$$

where the second transition follows from our assumptions on A. Thus we established that the family A_ϵ satisfies F1 and F2a, with $\Lambda_\epsilon := \frac{1}{\epsilon}I$. Finally, the fact that $A = \lim_{\epsilon\to 0} A_\epsilon$ follows from the above definition of A_ϵ and (25). □

18. The above proof was made under the assumption $\Lambda = \infty I$. For the general case, we make an additional partition $S = S_1 \cup S_2$ with S_1 the indices corresponding to Λ_{fin} and S_2 the indices corresponding to ∞I. Then to prove Theorem 17.2, say, we proceed as follows: (i) we observe that the problem reduces to the matrix $A_{\text{pure}} := A - 0 \oplus \Lambda_{\text{fin}} \oplus 0$; (ii) we realize $A_{\text{pure}} = \lim_{\epsilon\to 0} A_{\text{pure},\epsilon}$ by just applying the result which was already proved for Theorem 17.2, but now with index sets $R := R \cup S_1$, $S := S_2$, and $T := T \cup S_1$.

Theorem 17 allows a restatement in terms of topological closure. Let us illustrate this for a particular example. Let $\mathcal{M}$ be the set of "partially lower semiseparable" matrices satisfying $\mathcal{R} = \{\mathcal{B}_k\}_{k\in K}$, where K is a certain index set, and with each structure block $\mathcal{B}_k = (k,k,1,\lambda_k)$, where the shift element λ_k is allowed to take any value in $\mathbb{C}$. We claim that the topological closure $\bar{\mathcal{M}}$ can be obtained in exactly the

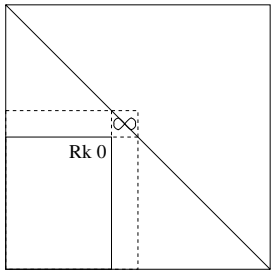


FIGURE 5. *Each structure block $\mathcal{B}_k = (k, k, 1, \infty)$, $k = 1, \ldots, n$, can be seen as a Hessenberg structure block. In particular, each Hessenberg matrix can be approximated by a sequence of lower semiseparable plus diagonal matrices.*

same way, but now allowing $\lambda_k \in \mathbb{C} \cup \{\infty\}$. This means that also [illegible] should be allowed in the structure; see Figure 5.

Indeed, let us show that to each $A \in \bar{\mathcal{M}}$ there corresponds a set of shift elements $\lambda_k \in \mathbb{C} \cup \{\infty\}$. Thus let $A_\epsilon \in \mathcal{M}$ be a converging sequence of matrices, with corresponding shift elements $\lambda_{\epsilon,k}$. By the fact that $\mathbb{C} \cup \{\infty\}$ is compact, for each k there must exist a convergent subsequence $\lambda_{k,\tilde{\epsilon}} \to \lambda_k \in \mathbb{C} \cup \{\infty\}$. The result follows then by Theorem 17.1 (Continuity).

Conversely, let us show that if A satisfies a set of shift elements $\lambda_k \in \mathbb{C} \cup \{\infty\}$, we have $A \in \bar{\mathcal{M}}$. But this follows just by Theorem 17.2 (Approximation by finite shift matrices). The only subtility is that the infinitesimal correction term, constructed in Theorem 17.2 for one structure block $\mathcal{B}_k$ with $\lambda_k = \infty$, should not destroy the other structure blocks $\mathcal{B}_{\tilde{k}}$, $\tilde{k} \neq k$. The fact that this is satisfied follows since the row and column spaces of these correction terms are well behaved; see the explanation in Figure 4.

We will not go further into this.

Now we come back to matrix inversion. We can use the result of Theorem 17 to remove the nonsingularity condition for Λ from the statement of Theorem 14.

COROLLARY 19. [illegible] $A \in \mathbb{C}^{n \times n}$ [illegible] $\mathcal{B} = (i, j, r, \Lambda)$ [illegible] $\Lambda = \Lambda_{\text{ns}} \oplus 0 \oplus \infty I$ [illegible] Λ_{ns} [illegible] A^{-1} [illegible] $\mathcal{B}^{-1} := (i, j, r, \Lambda^{-1})$ [illegible] $\Lambda^{-1} := \Lambda_{\text{ns}}^{-1} \oplus \infty I \oplus 0$ [illegible] $\frac{1}{0} = \infty$ [illegible] $\frac{1}{\infty} = 0$ [illegible]. The case where $\Lambda = \Lambda_{\text{ns}}$ has been proved in Theorem 14. The general case follows by continuity. Indeed, by Theorem 17.2, we can construct a family A_ϵ, $\epsilon \in \mathbb{C} \setminus \{0\}$ with $\lim_{\epsilon \to 0} A_\epsilon = A$, and such that the block ∞I is approximated by finite shift matrices $\frac{1}{\epsilon} I$. Next we add the correction term ϵI to the zero block. Thus each A_ϵ satisfies the structure block $\mathcal{B} = (i, j, r, \Lambda_\epsilon)$ with

$$\Lambda_\epsilon = \Lambda_{\text{ns}} \oplus \epsilon I \oplus \frac{1}{\epsilon} I.$$

Hence

$$\Lambda_\epsilon^{-1} = \Lambda_{\text{ns}}^{-1} \oplus \frac{1}{\epsilon} I \oplus \epsilon I.$$

The theorem follows then by applying Theorem 17.1 (Continuity). □

An interesting special case of Corollary 19 is when $\Lambda = 0$ or $\Lambda = \infty I$. Then the theorem can be interpreted in terms of [illegible] structure blocks; see Figure 6.

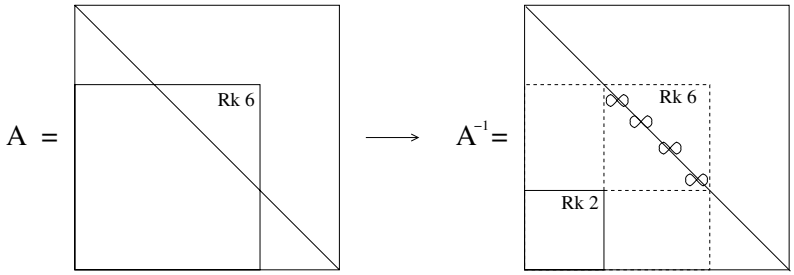


FIGURE 6. *Given the matrix A in the left picture, satisfying the pure structure block $\mathcal{B}_{\text{pure}} = (i, j, 6, 0)$. Then the inverse matrix A^{-1} satisfies the pure structure block $\mathcal{B}_{\text{pure}}^{-1} = (i, j, 6, \infty I)$ in the right picture.*

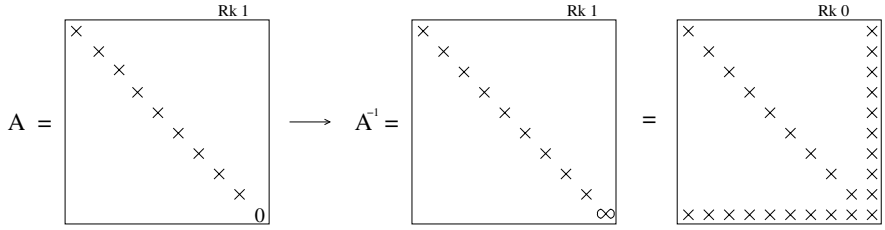


FIGURE 7. *Given the matrix A in the left picture, which is diagonal plus rank one with last diagonal element equal to zero. Then the inverse matrix is arrowhead, and conversely.*

In particular, we see that lower semiseparable and Hessenberg matrices are each other's inverses. This follows since these matrices are both lower semiseparable plus diagonal, with structure blocks $\mathcal{B}_k$ having shift elements $\lambda_k = 0$ for lower semiseparable and $\lambda_k = \infty$ for Hessenberg matrices.

For another example, suppose that A is a diagonal plus rank one matrix with diagonal correction Λ. Then there are two possibilities. If Λ is nonsingular, the inverse matrix A^{-1} is again diagonal plus rank one, with diagonal correction Λ^{-1}; this follows immediately from Theorem 14. Suppose now that Λ is singular, say, $\lambda_n = 0$. Then the corresponding diagonal element of Λ^{-1} is $\frac{1}{\lambda_n} = \infty$, and hence A^{-1} will be an [illegible]; see Figure 7.

Next let us assume that a certain square submatrix Λ of a matrix A is known. For example, let us suppose that

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & \Lambda \end{bmatrix},$$

with $\Lambda \in \mathbb{C}^{k \times k}$, for certain k. Then we claim that

$$A^{-1} - \begin{bmatrix} 0 & 0 \\ 0 & \Lambda^{-1} \end{bmatrix} = \text{Rk } n - k. \tag{28}$$

Indeed, this can be seen by the argument in Figure 8.

We should mention here that also other proofs for (28) are possible. One possibility is to use an argument based on Schur complements. Another possibility is to translate the data to the decoupled displacement equations of section 2 by writing $PAP^T - \Lambda = \text{Rk } 0$, where P is the projection matrix onto the last k columns. Hence Theorem 4 implies $A^{-1} - P^T \Lambda P = \text{Rk } n - k$, which is precisely (28).

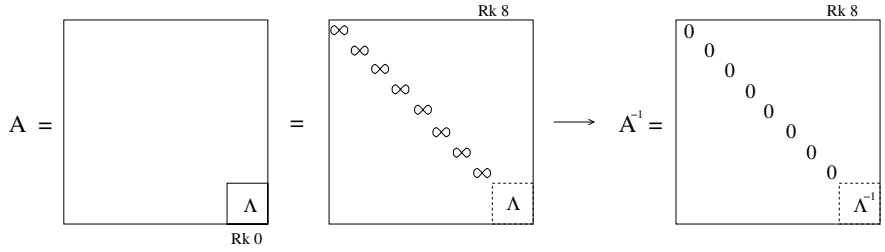


FIGURE 8. *Given the matrix* $A \in \mathbb{C}^{10\times 10}$ *in the leftmost picture, with given* $(2,2)$ *square submatrix* Λ. *By suitably adding shift elements* ∞, *we can express these data in terms of a structure block* $\mathcal{B}$. *Hence the inverse matrix will satisfy the inverse structure block* $\mathcal{B}^{-1}$ *of the rightmost picture, and this is precisely what* (28) *states.*

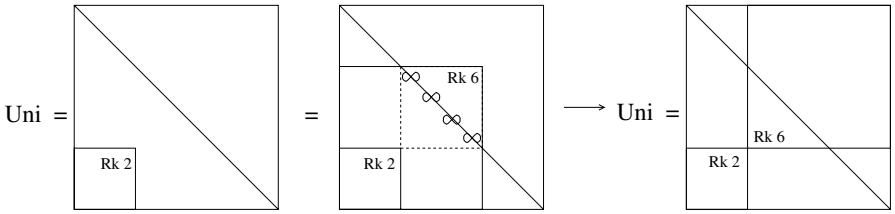


FIGURE 9. *For a unitary matrix* Uni, *the structure blocks always come in pairs. The picture illustrates this for a given pure* Rk 2 *structure block, but the result holds also for nonpure structure blocks.*

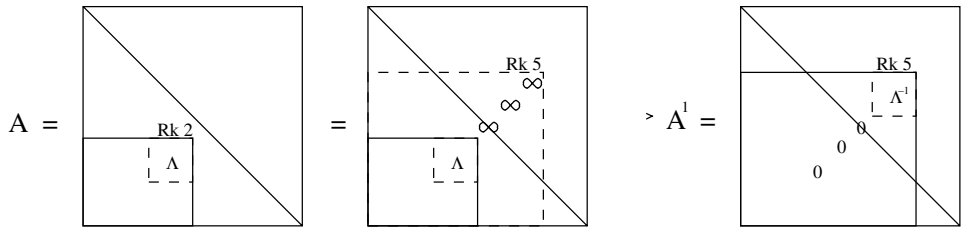


FIGURE 10. *Given the matrix in the left picture, with* Λ *a* 2 *by* 2 *matrix, we can say that the matrix satisfies a structure block by suitable adding some shift elements* ∞, *as indicated in the middle picture. The right picture shows then the form of the inverse matrix.*

As a final example, let us assume that $A = \mathrm{Uni} + \mathrm{Rk}\ r$ is a unitary plus low rank matrix. Then it is straightforward to check that $A^{-1} - A^H = \mathrm{Rk}\ 2r$ is a matrix of rank at most $2r$. Hence by Corollary 19, the structure blocks of A always come in , as illustrated in Figure 9.

4.2. Incorporating permutations. In this final subsection we derive a generalization of Theorem 14 and Corollary 19 by absorbing permutation matrices into the structure.

To do this in a clean way, we will first make some additional observations about the results in the previous subsection. Recall that Theorem 14 and Corollary 19 predict that both the rank and the shape of a structure block are preserved under matrix inversion. But it is important to remember that this holds only when appropriate use is made of shift elements ∞. Consider for example the matrix in the left and middle pictures of Figure 10. This matrix satisfies a structure block $\mathcal{B}$ whose shift matrix has an auxiliary submatrix ∞I_3. Comparing the middle and the right picture

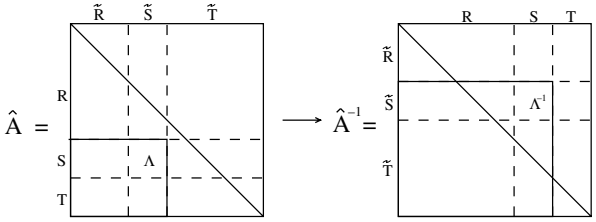


FIGURE 11. *By using suitable permutation matrices, the structures occurring in* (29) *can be brought to the bottom left matrix corner, resulting in a permuted matrix* $\hat{A}$. *The figure shows the position of the index sets* R, S, T; $\tilde{R}$, $\tilde{S}$, *and* $\tilde{T}$ *after permutation. The fact that the theorem is correct for the illustrated case follows from Figure* 10.

of Figure 10, we see indeed that the rank and the shape of $\mathcal{B}$ are preserved under matrix inversion. On the other hand, comparing the [illegible] and the right pictures of Figure 10, it makes more sense to say that the structure block has grown with three rows and columns and that its rank has increased from Rk 2 to Rk 5.

In general, we may be interested in an ∞-free formulation of the inversion theorem. Therefore we recall that each independent shift element ∞ has the effect of decreasing the rank by one and skipping one row and column out of the structure block; the key point is that these operations may influence the rank and shape but not the [illegible] of the structure block. In other words, from the ∞-free point of view, the quantity that is preserved under matrix inversion is not the rank of the structure block but rather its nullity.

We can extend this observation by incorporating permutations.

THEOREM 20 (nullity theorem). [illegible] $n \in \mathbb{N}$ [illegible] $N = \{1, \dots, n\}$ [illegible] $N = R \cup S \cup T = \tilde{R} \cup \tilde{S} \cup \tilde{T}$ [illegible] S $\tilde{S}$ [illegible]

$$\text{Null}\,(A^{-1}_{\Lambda^{-1}}(\tilde{S} \cup \tilde{T}, R \cup S)) = \text{Null}\,(A_{\Lambda}(S \cup T, \tilde{R} \cup \tilde{S})), \tag{29}$$

[illegible] $A^{-1}_{\Lambda^{-1}}$ [illegible] A^{-1} [illegible] $A^{-1}_{\Lambda^{-1}}(\tilde{S}, S) = A^{-1}(\tilde{S}, S) - \Lambda^{-1}$ [illegible] A_{Λ} [illegible] A [illegible] $A_{\Lambda}(S, \tilde{S}) = A(S, \tilde{S}) - \Lambda$ [illegible]. We will prove the theorem for the case $|R| \geq |\tilde{R}|$, i.e., for the case where R has at least the same number of elements as $\tilde{R}$. (The other case can be proved in a similar way). First we apply permutations to bring the structure to the bottom left matrix corner. More specifically, consider the permutation

$$P : \begin{cases} \{1, \dots, |R|\} & \mapsto & R, \\ \{|R|+1, \dots, n-|T|\} & \mapsto & S, \\ \{n-|T|+1, \dots, n\} & \mapsto & T. \end{cases}$$

Similarly, we define a permutation $\tilde{P}$. Then by construction, the right-hand side of (29) is equivalent with the matrix $\hat{A} := P^{-1} A \tilde{P}$ satisfying a (usual) structure block $\mathcal{B} = (i, j, r, \Lambda)$ with $i := |R| + 1$, $j := |R| + |S|$ and with r such the structure block has the required nullity; see the left picture of Figure 11. By the observations in the paragraphs preceding this proof, the inverse matrix $\hat{A}^{-1} = \tilde{P}^{-1} A^{-1} P$ will then satisfy the structure block illustrated in the right picture of Figure 11, with exactly the same nullity as $\mathcal{B}$. But by definition of the permutation matrices, this is equivalent to the left-hand side of (29) having the required nullity, hence proving the theorem. □

5. Conclusion. In this paper we investigated some structures on $\mathbb{C}^{n\times n}$ that have good behavior under matrix inversion. We handled two classes of them, namely, decoupled displacement structures and rank structures. For the case of rank structures, we provided some generalizations to deal with singular shift matrices and generally positioned structure blocks, leading amongst others to a different interpretation of a theorem due to Fiedler and Markham [6], which corresponds to the limiting cases $\Lambda = 0$ and $\Lambda = \infty I$. In [1], we show that these structures also have good behavior under Schur complementation.

REFERENCES

[1] S. Delvaux and M. Van Barel, *Structures preserved by Schur complementation*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 229–252.
[2] S. Delvaux and M. Van Barel, *Structures preserved by the QR-algorithm*, J. Comput. Appl. Math., 187 (2005), pp. 29–40.
[3] P. Dewilde and A.-J. van der Veen, *Time-Varying Systems and Computations*, Kluwer Academic Publishers, Boston, 1998.
[4] Y. Eidelman and I. C. Gohberg, *Fast inversion algorithms for diagonal plus semiseparable matrices*, Integral Equations Operator Theory, 27 (1997), pp. 165–183.
[5] D. Fasino, L. Gemignani, N. Mastronardi, and M. Van Barel, *Orthogonal rational functions and structured matrices*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 810–829.
[6] 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.
[7] I. C. Gohberg and M. A. Kaashoek, *Time varying linear systems with boundary conditions and integral operators,* I. *The transfer operator and its properties*, Integral Equations Operator Theory, 7 (1984), pp. 325–391.
[8] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, 1996.
[9] D. A. Gregory, B. Heyink, and K. N. Vander Meulen, *Inertia and biclique decompositions of joins of graphs*, J. Combin. Theory Ser. B, 88 (2003), pp. 135–151.
[10] G. Heinig and K. Rost, *Algebraic Methods for Toeplitz-like Matrices and Operators*, Mathematical Research, Akademie Verlag, Berlin, 1984.
[11] T. Kailath, S.-Y. Kung, and M. Morf, *Displacement ranks of matrices and linear equations*, J. Math. Anal. Appl., 68 (1979), pp. 395–407.
[12] T. Kailath and A. H. Sayed, *Displacement structure: Theory and applications*, SIAM Rev., 37 (1995), pp. 297–386.
[13] T. Kailath and A. H. Sayed, editors. *Fast Reliable Algorithms for Matrices with Structure*, SIAM, Philadelphia, 1999.
[14] V. Olshevsky, ed. *Fast Algorithms for Structured Matrices: Theory and Applications*, Contemp. Math. 323, AMS, Providence, RI, 2003.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 1, pp. 229–252

STRUCTURES PRESERVED BY SCHUR COMPLEMENTATION*

STEVEN DELVAUX† AND MARC VAN BAREL†

Abstract. In this paper we investigate some matrix structures on $\mathbb{C}^{m\times n}$ that are preserved by Schur complementation. The first type of structure is closely related to low displacement rank matrices. Next, we show that for a matrix having a low rank submatrix, the Schur complement also must have a low rank submatrix, which we can explicitly determine. This property holds even if the low rank submatrix contains a certain correction term that we call the shift matrix.

Key words. displacement structures, rank structures, lower semiseparable (plus diagonal) matrices, Schur complements

AMS subject classifications. 15A09, 15A03, 65F05

DOI. 10.1137/040621417

1. Introduction.

In this paper we will handle several matrix structures that are preserved by Schur complementation, as a continuation of [1] where we handled structures preserved by matrix inversion. Nevertheless, all results will be developed independently of [1].

Section 2 deals with the preservation of [illegible]. As in [1], the idea is to generalize the classical examples of displacement structures (such as Toeplitz-like, Cauchy-like, Vandermonde-like, circulant matrices; see [11]) by "decoupling" the displacement equation. This means that the displacement equation is allowed to involve two variables A and B rather than only one variable A. We will then illustrate this definition by some examples.

Section 3 handles the preservation of what we call [illegible]. As in [2, 1], such a structure is defined as a collection of [illegible]: these are low rank submatrices of a given matrix $A \in \mathbb{C}^{m\times n}$, together with a certain correction term Λ called the shift matrix. We will prove that these rank structures are preserved under Schur complementation, and we provide some examples to illustrate this. These examples include the preservation of higher-order semiseparable plus diagonal matrices under Schur complementation, which was the basis of a fast solver in [4].

Section 4 considers [illegible] of a matrix A. Each Möbius transformation can be realized as the Schur complement of a very special block matrix, and hence this connection can be used to translate the preservation results of Schur complements into properties of Möbius transformations.

*Received by the editors December 24, 2004; accepted for publication (in revised form) by H. J. Woerdeman August 30, 2005; published electronically March 17, 2006. This research was partially supported by the Research Council K.U.Leuven, project OT/00/16 (SLAP: Structured Linear Algebra Package), by the Fund for Scientific Research–Flanders (Belgium), projects G.0078.01 (SMA: Structured Matrices and their Applications), G.0176.02 (ANCILA: Asymptotic aNalysis of the Convergence behavior of Iterative methods in numerical Linear Algebra), G.0184.02 (CORFU: Constructive study of Orthogonal Functions) and G.0455.0 (RHPH: Riemann-Hilbert problems, random matrices and Padé–Hermite approximation), 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 scientific responsibility rests with the authors.

http://www.siam.org/journals/simax/28-1/62141.html

†Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, B-3001 Leuven (Heverlee), Belgium (Steven.Delvaux@cs.kuleuven.ac.be, Marc.VanBarel@cs.kuleuven.ac.be).

For further reference, let us recall here some basic definitions and properties of Schur complements [5].

DEFINITION 1. [illegible] $A \in \mathbb{C}^{m\times n}$ [illegible] $k \leq \min\{m,n\}$ [illegible] k-partitioning [illegible] A [illegible]

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}_k, \tag{1.1}$$

[illegible] $A_{1,1} \in \mathbb{C}^{k\times k}$ [illegible] Schur complement [illegible] k [illegible]

$$S_{A,k} := A_{2,2} - A_{2,1}A_{1,1}^{-1}A_{1,2},$$

[illegible] $A_{1,1}$ [illegible] $A_{\mathrm{c},1}$ [illegible] $A_{\mathrm{r},1}$ [illegible] (1.1)

Schur complements are related to Gaussian elimination steps on A with pivot block $A_{1,1}$, in the sense that

$$L_{Gauss} \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix} R_{Gauss} = \begin{bmatrix} A_{1,1} & 0 \\ 0 & S_{A,k} \end{bmatrix}, \tag{1.2}$$

where

$$L_{Gauss} := \begin{bmatrix} I & 0 \\ -A_{2,1}A_{1,1}^{-1} & I \end{bmatrix}, \quad R_{Gauss} := \begin{bmatrix} I & -A_{1,1}^{-1}A_{1,2} \\ 0 & I \end{bmatrix},$$

which are unit block lower and upper triangular matrices, respectively. Hence the following lemma should not come as a surprise.

LEMMA 2. [illegible] $L \in \mathbb{C}^{l\times m}$ $A \in \mathbb{C}^{m\times n}$ [illegible] $R \in \mathbb{C}^{n\times p}$ [illegible]

$$L = \begin{bmatrix} L_{1,1} & 0 \\ L_{1,2} & L_{2,2} \end{bmatrix}, \quad A = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}, \quad R = \begin{bmatrix} R_{1,1} & R_{1,2} \\ 0 & R_{2,2} \end{bmatrix},$$

[illegible] $L_{1,1}$ $A_{1,1}$ [illegible] $R_{1,1}$ [illegible] $\mathbb{C}^{k\times k}$ [illegible]

$$S_{LAR,k} = L_{2,2}S_{A,k}R_{2,2}.$$

[illegible]. First, let us prove the property for the case $R = I$. Then we can expand the matrix LA as

$$\begin{bmatrix} L_{1,1}A_{1,1} & L_{1,1}A_{1,2} \\ L_{1,2}A_{1,1} + L_{2,2}A_{2,1} & L_{1,2}A_{1,2} + L_{2,2}A_{2,2} \end{bmatrix},$$

from which it follows that

$$\begin{aligned} S_{LA,k} &= L_{1,2}A_{1,2} + L_{2,2}A_{2,2} - (L_{1,2}A_{1,1} + L_{2,2}A_{2,1})A_{1,1}^{-1}L_{1,1}^{-1}(L_{1,1}A_{1,2}) \\ &= L_{2,2}[A_{2,2} - A_{2,1}A_{1,1}^{-1}A_{1,2}] = L_{2,2}S_{A,k}, \end{aligned}$$

as we had to prove. In a similar way, one can prove the property for the case $L = I$. The general result follows then by composing these two results. ∎

Note that in particular, it follows from the above lemma that the left multiplication of A with a matrix $\left[\begin{smallmatrix} X & 0 \\ X & I \end{smallmatrix}\right]$ or the right multiplication with the transpose of such a matrix always preserves the Schur complement of A.

We also may recall the so-called [illegible] of Schur complements, i.e., the fact that $S_{S_{A,k},l} = S_{A,k+l}$ whenever all involved Schur complements are defined. The underlying reason is that for a Gaussian elimination step applied on A, the same result is obtained if the Gaussian elimination step is split into two separate steps with smaller pivot blocks. Alternatively, one can prove this property by direct computation.

2. Displacement structures. In this section we handle the preservation of displacement structures under Schur complementation. As a general reference, we can refer to [11, 10] for an overview of the many applications of displacement theory in numerical linear algebra. Some references of historical interest are [7, 9].

2.1. Sylvester-type displacement. First we handle Sylvester-type displacement equations. As a classical example, let A be a Hankel matrix, i.e., $A = [a_{i+j}]_{i,j=1}^{n}$. Putting $Z := [\mathbf{e}_2 \dots \mathbf{e}_n \mathbf{0}]$ with $\mathbf{e}_k$ the kth column of the identity matrix, it is easy to check that

$$AZ^T - ZA = \text{Rk } 2, \tag{2.1}$$

where Rk 2 denotes a matrix of rank at most 2.

Generalizing, we can come to a more general definition. The main difference with (2.1) is that the variable A is "decoupled" into two variables A and B.

DEFINITION 3. [illegible] A B F [illegible] G [illegible] $r \in \mathbb{N}$ [illegible] A [illegible] B [illegible] Sylvester-type displacement equation [illegible] (F, G, r) [illegible]

$$AF - GB = \text{Rk } r, \tag{2.2}$$

[illegible] Rk r [illegible] r

Here we suppose (2.2) to be well defined or, equivalently, we suppose the block matrix

$$\left[\begin{array}{cc} A & G \\ F^T & B^T \end{array}\right] \tag{2.3}$$

to have compatible matrix dimensions. Moreover, this block representation (2.3) is useful in several other aspects, as will become clear soon.

Let us show that for F^T and G block lower triangular, Sylvester-type displacement structure is preserved under Schur complementation. This generalizes the corresponding property for the case $A = B$ (see [11]).

THEOREM 4 (Sylvester-type inheritance). [illegible] $k, l \in \mathbb{N}$ [illegible]

$$\left[\begin{array}{c|c} A & G \\ \hline F^T & B^T \end{array}\right] = \left[\begin{array}{cc|cc} A_{1,1} & A_{1,2} & G_{1,1} & 0 \\ A_{2,1} & A_{2,2} & G_{2,1} & \tilde{G} \\ \hline F_{1,1}^T & 0 & B_{1,1}^T & B_{2,1}^T \\ F_{1,2}^T & \tilde{F}^T & B_{1,2}^T & B_{2,2}^T \end{array}\right], \tag{2.4}$$

[illegible] $A_{1,1} \in \mathbb{C}^{k \times k}$ [illegible] $B_{1,1} \in \mathbb{C}^{l \times l}$ [illegible]

$$AF - GB = \text{Rk } r, \tag{2.5}$$

[illegible] Rk r [illegible] r [illegible]

$$S_{A,k}\tilde{F} - \tilde{G}S_{B,l} = \widetilde{\text{Rk}}\, r, \tag{2.6}$$

[illegible] $\widetilde{\text{Rk}}\, r$ [illegible] r

[illegible]. We will prove the theorem for A and B square and nonsingular. (For the general case, see the paragraph following this proof.) Multiplying (2.5) on the left with A^{-1} and on the right with B^{-1}, it follows that

$$FB^{-1} - A^{-1}G = \text{Rk } r, \tag{2.7}$$

with Rk r a new matrix of rank at most r. Now we use the fact that for any matrix A, the $(2,2)$ block element of A^{-1} is precisely the inverse of the Schur complement $S_{A,k}$. (The proof follows by inverting both sides of (1.2).) Using this, and using the partitioning in (2.4), it follows by evaluating the $(2,2)$ block element of (2.7) that

$$\tilde{F}S_{B,l}^{-1} - S_{A,k}^{-1}\tilde{G} = \text{Rk } r,$$

with Rk r a new matrix of rank at most r. Hence by multiplying on the left with $S_{A,k}$ and on the right with $S_{B,l}$, we obtain the desired equation (2.6). ∎

Although the above proof of inheritance of structure is rather "clean," it only works for square and nonsingular matrices. (One could use a reduction to square matrices and a "continuity argument" to remove these restrictions, but we will not do this here, due to the complexity of the argument.) Furthermore, the proof of Theorem 4 gives rather complicated formulae for the new $\widetilde{\text{Rk}}\ r$ matrix in the right-hand side of (2.6).

To address these questions, one can proceed in a more direct way by directly computing the Schur complements.

Let us work this out. Thus we start with the equation

$$AF - GB = \text{Rk } r =: UV, \tag{2.8}$$

with U having r columns and V having r rows. Let us recall the general property

$$A - A_{\text{c},1}A_{1,1}^{-1}A_{\text{r},1} = 0 \oplus S_{A,k}, \tag{2.9}$$

where "$\oplus$" denotes the operator putting its arguments as diagonal blocks in a block diagonal matrix (as usual), and where we used the notation of Definition 1. Keeping in mind this property and the partitioning in (2.4), we have

$$\begin{aligned} 0 \oplus \widetilde{\text{Rk}}\ r &:= 0 \oplus (S_{A,k}\tilde{F} - \tilde{G}S_{B,l}) \\ &= (0 \oplus S_{A,k})F - G(0 \oplus S_{B,l}) \\ &= (A - A_{\text{c},1}A_{1,1}^{-1}A_{\text{r},1})F - G(B - B_{\text{c},1}B_{1,1}^{-1}B_{\text{r},1}) \\ &= UV - A_{\text{c},1}A_{1,1}^{-1}A_{\text{r},1}F + GB_{\text{c},1}B_{1,1}^{-1}B_{\text{r},1}, \end{aligned} \tag{2.10}$$

where the last transition follows from (2.8). Still keeping in mind (2.8) and the partitioning in (2.4), we can further work this out as

$$\begin{aligned} &= UV - A_{\text{c},1}A_{1,1}^{-1}(U_{\text{r},1}V + \underline{G_{1,1}B_{\text{r},1}}) + (\underline{A_{\text{c},1}F_{1,1}} - UV_{\text{c},1})B_{1,1}^{-1}B_{\text{r},1} \\ &= UV - A_{\text{c},1}A_{1,1}^{-1}U_{\text{r},1}V - UV_{\text{c},1}B_{1,1}^{-1}B_{\text{r},1} + \underline{A_{\text{c},1}A_{1,1}^{-1}U_{\text{r},1}V_{\text{c},1}B_{1,1}^{-1}B_{\text{r},1}} \\ &= (U - A_{\text{c},1}A_{1,1}^{-1}U_{\text{r},1})(V - V_{\text{c},1}B_{1,1}^{-1}B_{\text{r},1}). \end{aligned} \tag{2.11}$$

We see from this that $\widetilde{\text{Rk}}\ r$ is indeed a matrix of rank at most r, which we can explicitly determine. Moreover, the only condition for the above derivation to be valid is the nonsingularity of $A_{1,1}$ and $B_{1,1}$, i.e., the existence of the Schur complements $S_{A,k}$ and $S_{B,l}$.

2.2. Stein-type displacement. We come to a second type of displacement structure.

DEFINITION 5. [illegible] A B G [illegible] H [illegible] $r \in \mathbb{N}$ [illegible] A [illegible] B [illegible] Stein-type displacement equation [illegible] (G, H, r)

$$A - GBH = \text{Rk } r, \tag{2.12}$$

[illegible] Rk r [illegible] r

Here we suppose (2.12) to be well defined or, equivalently, we suppose the block matrix

$$\begin{bmatrix} A & G \\ H & B^T \end{bmatrix} \tag{2.13}$$

to have compatible matrix dimensions. Moreover, this block representation (2.13) is useful in several other aspects, as will become clear soon.

As in the Sylvester case, for H^T and G block lower triangular, Stein-type displacement structure will be preserved under Schur complementation. This generalizes again the corresponding property for the case $A = B$ (see [11]).

THEOREM 6 (Stein-type inheritance). [illegible] $k, l \in \mathbb{N}$ [illegible]

$$\left[\begin{array}{c|c} A & G \\ \hline H & B^T \end{array}\right] = \left[\begin{array}{cc|cc} A_{1,1} & A_{1,2} & G_{1,1} & 0 \\ A_{2,1} & A_{2,2} & G_{2,1} & \tilde{G} \\ \hline H_{1,1} & H_{1,2} & B_{1,1}^T & B_{2,1}^T \\ 0 & \tilde{H} & B_{1,2}^T & B_{2,2}^T \end{array}\right], \tag{2.14}$$

[illegible] $A_{1,1} \in \mathbb{C}^{k\times k}$ [illegible] $B_{1,1} \in \mathbb{C}^{l\times l}$ [illegible]

$$A - GBH = \operatorname{Rk} r, \tag{2.15}$$

[illegible]

$$S_{A,k} - \tilde{G} S_{B,l} \tilde{H} = \widetilde{\operatorname{Rk}}\, \tilde{r}, \tag{2.16}$$

[illegible] $\tilde{r} := r - k + l$ [illegible]. We will prove the theorem for $A \in \mathbb{C}^{m\times m}$ and $B \in \mathbb{C}^{n\times n}$ square and nonsingular. (For the general case, see the paragraph following this proof.) From [1, Theorem 2], it follows that

$$B^{-1} - HA^{-1}G = \operatorname{Rk}(r + n - m). \tag{2.17}$$

Now we recall the fact that for any matrix A, the $(2,2)$ block element of A^{-1} is precisely the inverse of the Schur complement $S_{A,k}$. Using this and the partitioning in (2.14), it follows by evaluating the $(2,2)$ block element of (2.17) that

$$S_{B,l}^{-1} - \tilde{H} S_{A,k}^{-1} \tilde{G} = \operatorname{Rk}(r + n - m),$$

with $\operatorname{Rk}(r+n-m)$ still a matrix of rank at most $r+n-m$. Hence by applying again [1, Theorem 2], we obtain the desired equation (2.16), i.e.,

$$S_{A,k} - \tilde{G} S_{B,l} \tilde{H} = \widetilde{\operatorname{Rk}}\,(r + n - m + (m - k) - (n - l)) =: \widetilde{\operatorname{Rk}}\, \tilde{r},$$

with $\tilde{r} := r - k + l$. □

Again, the above proof was valid only for A and B square and nonsingular. Instead of showing theoretically that these restrictions are not essential (by using a reduction to square matrices, together with a "continuity argument" to remove the nonsingularity condition), let us indicate how to prove the theorem by a direct approach.

We start with the equation

$$A - GBH = \operatorname{Rk} r =: UV, \tag{2.18}$$

with U having r columns and V having r rows. In a way similar to the derivation of (2.10), we obtain

$$0 \oplus \widetilde{\mathrm{Rk}}\, r = UV - A_{\mathrm{c},1}A_{1,1}^{-1}A_{\mathrm{r},1} + GB_{\mathrm{c},1}B_{1,1}^{-1}B_{\mathrm{r},1}H. \tag{2.19}$$

Then keeping in mind (2.18) and the partitioning in (2.14), we can further work out the right-hand side of (2.19) as

$$\begin{aligned} &= UV - (UV_{\mathrm{c},1} + GB_{\mathrm{c},1}H_{1,1})A_{1,1}^{-1}(U_{\mathrm{r},1}V + G_{1,1}B_{\mathrm{r},1}H) \\ &\qquad + GB_{\mathrm{c},1}B_{1,1}^{-1}B_{\mathrm{r},1}H \\ &= U(I - V_{\mathrm{c},1}A_{1,1}^{-1}U_{\mathrm{r},1})V - UV_{\mathrm{c},1}A_{1,1}^{-1}G_{1,1}B_{\mathrm{r},1}H \\ &\qquad - GB_{\mathrm{c},1}H_{1,1}A_{1,1}^{-1}U_{\mathrm{r},1}V - GB_{\mathrm{c},1}(H_{1,1}A_{1,1}^{-1}G_{1,1} - B_{1,1}^{-1})B_{\mathrm{r},1}H. \end{aligned} \tag{2.20}$$

To proceed further, we will assume that the following assumption holds.

. $k = l$.

Then we claim that there exist matrices $X_1 \in \mathbb{C}^{k\times r}$, $X_3 \in \mathbb{C}^{r\times k}$, $X_2, X_4 \in \mathbb{C}^{r\times r}$ that satisfy the embedding relation

$$\begin{bmatrix} B_{1,1}^{-1} & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} H_{1,1} & X_1 \\ V_{\mathrm{c},1} & X_2 \end{bmatrix} \begin{bmatrix} A_{1,1}^{-1} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} G_{1,1} & U_{\mathrm{r},1} \\ X_3 & X_4 \end{bmatrix}, \tag{2.21}$$

where all involved matrices are square of size $k + r$. Assuming this for the moment, then we have the componentwise equations

$$\begin{cases} B_{1,1}^{-1} &= H_{1,1}A_{1,1}^{-1}G_{1,1} + X_1X_3, \\ 0 &= H_{1,1}A_{1,1}^{-1}U_{\mathrm{r},1} + X_1X_4, \\ 0 &= V_{\mathrm{c},1}A_{1,1}^{-1}G_{1,1} + X_2X_3, \\ I &= V_{\mathrm{c},1}A_{1,1}^{-1}U_{\mathrm{r},1} + X_2X_4. \end{cases}$$

Hence (2.20) can be rewritten as

$$\begin{aligned} &= UX_2X_4V + UX_2X_3B_{\mathrm{r},1}H + GB_{\mathrm{c},1}X_1X_4V + GB_{\mathrm{c},1}X_1X_3B_{\mathrm{r},1}H \\ &= (UX_2 + GB_{\mathrm{c},1}X_1)(X_4V + X_3B_{\mathrm{r},1}H). \end{aligned} \tag{2.22}$$

We see from this that $\widetilde{\mathrm{Rk}}\, r$ is indeed a matrix of rank at most r, which we can explicitly determine in terms of X_1, X_2, X_3, and X_4.

To prove the solvability of the embedding relation (2.21) is beyond the scope of the paper. We may notice that it suffices to find X_i, $i = 1, \ldots, 4$, which solve the equivalent embedding relation

$$\begin{bmatrix} A_{1,1} & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} G_{1,1} & U_{\mathrm{r},1} \\ X_3 & X_4 \end{bmatrix} \begin{bmatrix} B_{1,1} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} H_{1,1} & X_1 \\ V_{\mathrm{c},1} & X_2 \end{bmatrix}, \tag{2.23}$$

where the $(1,1)$ block element of this equation is nothing but the equality $A_{1,1} = G_{1,1}B_{1,1}H_{1,1} + U_{\mathrm{r},1}V_{\mathrm{c},1}$, which is satisfied by (2.18). To prove that the other block elements of this equation also can be satisfied, we refer to [10], where completely similar problems are handled.

Finally, we recall our above assumption that $k = l$. For $k \neq l$, we should additionally look at two special cases. The first is when $k \neq 0$ and $l = 0$: then $B_{1,1}$ is the

empty matrix, and hence the partitioning in (2.14) implies $G = \left[\begin{smallmatrix} 0 \\ \tilde{G} \end{smallmatrix}\right]$ and $H = \left[\begin{smallmatrix} 0 & \tilde{H} \end{smallmatrix}\right]$. Thus the displacement equation $A - GBH = \text{Rk } r$ can be rewritten as

$$A - (0 \oplus \tilde{G}B\tilde{H}) = \text{Rk } r. \tag{2.24}$$

Applying on (2.24) the block unit lower and upper triangular transformations L_{Gauss} and R_{Gauss} appearing in (1.2), we obtain

$$A_{1,1} \oplus (S_{A,k} - \tilde{G}B\tilde{H}) = \text{Rk } r,$$

with Rk r a new matrix of rank at most r. Hence it follows that indeed $S_{A,k} - \tilde{G}B\tilde{H} = \widetilde{\text{Rk}}\ \tilde{r}$ with $\tilde{r} := r - k$.

The second special case is when $l \neq 0$ and $k = 0$, and then it can be seen by a similar argument that indeed $A - \tilde{G}S_{B,l}\tilde{H} = \widetilde{\text{Rk}}\ \tilde{r}$ with $\tilde{r} := r + l$.

The general case $k \neq l$ follows then by combining the results for $k = l$ together with the above two special cases, by using the "transitivity" of Schur complements. We will not go further into this.

2.3. Stein–Sylvester hybrid displacement. The reader will have noticed a lot of similarity between the Sylvester- and Stein-type displacement. In fact, there exist also [illegible], in the sense described by Kailath and Sayed [10, section 7.4].

To introduce these structures, let us start from the Sylvester-type displacement equation $AF - GB = \text{Rk } r$. Suppose we can factor

$$A := E\tilde{A}, \quad B := \tilde{B}H, \tag{2.25}$$

for certain block lower triangular matrices E, H^T with $E_{1,1}$ and $H_{1,1}$ square and nonsingular. Then Lemma 2 implies that

$$S_{A,k} := E_{2,2}S_{\tilde{A},k}, \quad S_{B,l} := S_{\tilde{B},l}H_{2,2},$$

and, moreover, it is easy to see that by substituting (2.25) into (2.11), the latter transforms into the expression

$$\widetilde{\text{Rk}}\ r = (U - E\tilde{A}_{\text{c},1}\tilde{A}_{1,1}^{-1}E_{1,1}^{-1}U_{\text{r},1})(V^T - V_{\text{c},1}^T H_{1,1}^{-1}\tilde{B}_{1,1}^{-1}\tilde{B}_{\text{r},1}H). \tag{2.26}$$

Similarly, suppose that in the [illegible]-type displacement equation $A - GBH = \text{Rk } r$ we can factor $A := E\tilde{A}F$ for certain block lower triangular matrices E, F^T with $E_{1,1}$ and $F_{1,1}$ square and nonsingular. Then Lemma 2 implies that

$$S_{A,k} = E_{2,2}S_{\tilde{A},k}F_{2,2},$$

and, moreover, it is easy to see that (2.22) remains invariant; the only change is that the embedding relation (2.21) must be updated by substituting $A_{1,1}^{-1} := F_{1,1}^{-1}\tilde{A}_{1,1}^{-1}E_{1,1}^{-1}$.

Note that in both cases, we were led to an equation of the form

$$EAF + GBH = \text{Rk } r.$$

In particular, the block matrix

$$\begin{bmatrix} E & G & 0 \\ A^T & 0 & F \\ 0 & B^T & H \end{bmatrix} \tag{2.27}$$

must have compatible matrix dimensions. We can then summarize the above facts in the following theorem.

THEOREM 7 (Stein–Sylvester hybrid inheritance). [illegible] $k, l \in \mathbb{N}$ [illegible] k [illegible] A [illegible] l [illegible] B [illegible] 1 [illegible] (2.27) [illegible] A [illegible] B [illegible] E F^T G [illegible] H^T [illegible] $\{E_{1,1}, G_{1,1}\}$ [illegible] $\{F_{1,1}, H_{1,1}\}$ [illegible]

$$\left[\begin{array}{c|c|c} E & G & 0 \\ \hline A^T & 0 & F \\ \hline 0 & B^T & H \end{array}\right] = \left[\begin{array}{cc|cc|cc} E_{1,1} & 0 & G_{1,1} & 0 & 0 & 0 \\ E_{2,1} & \tilde{E} & G_{2,1} & \tilde{G} & 0 & 0 \\ \hline A_{1,1}^T & A_{2,1}^T & 0 & 0 & F_{1,1} & F_{1,2} \\ A_{1,2}^T & A_{2,2}^T & 0 & 0 & 0 & \tilde{F} \\ \hline 0 & 0 & B_{1,1}^T & B_{2,1}^T & H_{1,1} & H_{1,2} \\ 0 & 0 & B_{1,2}^T & B_{2,2}^T & 0 & \tilde{H} \end{array}\right].$$

[illegible]

$$EAF + GBH = \text{Rk } r, \tag{2.28}$$

[illegible]

$$\tilde{E} S_{A,k} \tilde{F} + \tilde{G} S_{B,l} \tilde{H} = \widetilde{\text{Rk}}\, \tilde{r}, \tag{2.29}$$

[illegible] $\tilde{r} := r$ [illegible] $\{E_{1,1}, H_{1,1}\}$ [illegible] $\{F_{1,1}, G_{1,1}\}$ [illegible] [illegible] $\tilde{r} := r - k + l$ [illegible] $\{E_{1,1}, F_{1,1}\}$ [illegible] $\{G_{1,1}, H_{1,1}\}$ [illegible] [illegible]. This follows from the paragraph preceding the statement of the theorem. We even showed there how to update the explicit formulae for the new $\widetilde{\text{Rk}}\, \tilde{r}$ matrix, if so desired. □

As an application of Stein–Sylvester hybrid displacement structure, we will use it to establish a converse to the reasoning in the proof of Theorem 6; i.e., we will show how the preservation of structure under matrix inversion [1, Theorem 2] is a consequence of the preservation of structure under Schur complementation. Thus let $A \in \mathbb{C}^{m \times m}$ and $B \in \mathbb{C}^{n \times n}$ be nonsingular matrices satisfying $A - GBH = \text{Rk } r$ for arbitrary G and H. Hence

$$\left[\begin{array}{cc} A - GBH & G - G \\ H - H & 0 \end{array}\right] = \text{Rk } r$$

or, by a small calculation,

$$\left[\begin{array}{cc} I & 0 \\ 0 & H \end{array}\right] \left[\begin{array}{cc} A & -I \\ I & 0 \end{array}\right] \left[\begin{array}{cc} I & 0 \\ 0 & G \end{array}\right] - \left[\begin{array}{cc} G & 0 \\ 0 & I \end{array}\right] \left[\begin{array}{cc} B & -I \\ I & 0 \end{array}\right] \left[\begin{array}{cc} H & 0 \\ 0 & I \end{array}\right] = \text{Rk } r.$$

But now the Schur complements in this last equation are precisely the inverse matrices A^{-1} and B^{-1}. Hence by Theorem 7, it follows that

$$HA^{-1}G - B^{-1} = \widetilde{\text{Rk}}\, \tilde{r},$$

with $\tilde{r} := r + n - m$, as we had to prove.

REMARK 8.

1. [illegible] [1, 2] [illegible] 6 [illegible] 6 [illegible] [1, 2]
2. [illegible] $A = B$ [illegible] A^{-1} [illegible] generalized Schur algorithm [11] [illegible] $A \neq B$ [illegible]

2.4. Computational aspects. The preservation results of this section are in the first place [illegible] oriented, in the sense that there seems to be no analogue if $A \neq B$ for the so-called generalized Schur algorithm [11].

To state the problem more precisely, let us first make some assumptions. Suppose that F^T, G (for Sylvester-type displacement) and H^T, G (for Stein-type displacement) are not just [illegible] lower triangular but completely lower triangular matrices. Then the preservation of structure holds for any choice of indices $k = l$. Moreover, by the transitivity of Schur complements we are allowed to recursively pull off rows and columns of A and B, one at a time, so that we can assume that $k = 1 = l$.

Now let us recall the explicit formulae (2.11) and (2.22) that we obtained for the new low rank matrix $\widetilde{\text{Rk}}\, r$. These formulae involved information about the first row and column of the matrices A and B. (For the Stein type, this dependence also appeared in an indirect way, via the embedding relation (2.21).) The ideal situation would be the following: we use the given displacement equation in order to determine these first rows and columns, next update the generators of the $\widetilde{\text{Rk}}\, r$ matrix, and then repeat this procedure in a recursive way on the Schur complements $S_{A,k}$ and $S_{B,l}$ (which we do not actually compute but only store in a "coded" form by means of the subsequent $\widetilde{\text{Rk}}\, r$ matrices). Repeating this procedure, at the end we would obtain information about the LDU decompositions [5] of both the matrices A and B.

Unfortunately, the above scheme cannot possibly work since the given displacement equation does not contain enough information to determine the first block rows and columns of [illegible] the matrices A and B.

The situation may be different if an additional connection is given between A and B. For example, it could be that (i) a factorization $B = LDU$ is given, and we want to compute the LDU decomposition of A; (ii) we have a relation in the style $B = A$ (leading to the generalized Schur algorithm as described in [11]), or $B = A^T$. Thus only in such cases can we hope for the above scheme to work.

3. Rank structures. In this section we handle the preservation of rank structures. The following result could already have been mentioned in the previous section. It is a special case of the preservation of both Sylvester- and Stein-type displacement structures.

COROLLARY 9. [illegible] $k \in \mathbb{N}$ [illegible] A [illegible] B [illegible] [illegible] $S_{A,k}$ [illegible] $S_{B,k}$ [illegible]

$$A - B = \text{Rk}\, r,$$

[illegible]

$$S_{A,k} - S_{B,k} = \widetilde{\text{Rk}}\, r.$$

Note that there appears only one index k in the statement of the above corollary, rather than two indices k and l as in the previous section. This is because we have here F^T and G equal to the identity matrix, which by Theorem 4 has to be block lower triangular w.r.t. the indices k and l; hence $k = l$ is the only relevant choice.

It turns out that the analogy between matrix inversion and Schur complementation goes still deeper, as proved by the following specification of Corollary 9.

THEOREM 10 (Sherman–Morrison-like formula). [illegible] 9 [illegible] $\operatorname{Rk} r =: UV$ [illegible] U [illegible] r [illegible] V [illegible] r [illegible]

$$\widetilde{\operatorname{Rk}}\, r = (U_2 - A_{2,1}A_{1,1}^{-1}U_1)(V_2 - V_1A_{1,1}^{-1}A_{1,2}), \tag{3.1}$$

[illegible]

$$\begin{bmatrix} I_r & V \\ U & A \end{bmatrix} = \begin{bmatrix} I_r & V_1 & V_2 \\ U_1 & A_{1,1} & A_{1,2} \\ U_2 & A_{2,1} & A_{2,2} \end{bmatrix}, \tag{3.2}$$

[illegible] I_r [illegible] r [illegible] $A_{1,1} \in \mathbb{C}^{k\times k}$ [illegible]. It is possible to prove this by plugging in the Sherman–Morrison formula [5, section 2.1] into the formulae for Sylvester displacement equations obtained in (2.11). But let us give here a direct proof. Consider again the matrix (3.2). We will compute in two different ways the Schur complement induced by the leading $(r+k)$ by $(r+k)$ submatrix. The first way is to apply first I_r as pivot, resulting in the partial Schur complement $A - UV$. Taking now the Schur complement w.r.t. the leading k by k submatrix gives

$$S_{A-UV,k}. \tag{3.3}$$

The second way is to use first $A_{1,1}$ as pivot to eliminate the $(3,2)$ and $(2,3)$ elements, resulting in

$$\begin{bmatrix} I_r & V_1 & V_2 - V_1A_{1,1}^{-1}A_{1,2} \\ U_1 & A_{1,1} & 0 \\ U_2 - A_{2,1}A_{1,1}^{-1}U_1 & 0 & S_{A,k} \end{bmatrix}.$$

If we then apply I_r as pivot to eliminate the $(3,1)$ and $(1,3)$ elements, we obtain the final Schur complement

$$S_{A,k} - (U_2 - A_{2,1}A_{1,1}^{-1}U_1)(V_2 - V_1A_{1,1}^{-1}A_{1,2}). \tag{3.4}$$

Equating (3.3) and (3.4) gives the desired formula (3.1). □

Now we return to the simpler statement of Corollary 9. Suppose then that we take B to be an arbitrary Hermitian matrix. By the general property $S_{B^H,k} = (S_{B,k})^H$, the Schur complement $S_{B,k}$ also must be Hermitian. Hence Corollary 9 reveals the following fact.

COROLLARY 11. [illegible] $A = \text{Herm} + \operatorname{Rk} r$ [illegible] A [illegible] r [illegible]

The rest of this section is devoted to the preservation of what we call rank structures. First we recall some definitions from [1]. We will use here the subscript [illegible] to distinguish these definitions from the actual definition of rank structures, which is given later.

DEFINITION 12 (see [1]). weak structure block $\mathcal{B}_{\text{weak}}$ $\mathbb{C}^{n\times n}$
4

$$\mathcal{B}_{\text{weak}} = (i, j, r, \Lambda),$$

i j r
$\Lambda \in \mathbb{C}^{(j-i+1)\times(j-i+1)}$ $\mathcal{B}_{\text{weak}}$
$j-i+1 \geq 0$ $A \in \mathbb{C}^{n\times n}$

$$A =: \begin{bmatrix} A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,1} & A_{3,2} & A_{3,3} \end{bmatrix}, \tag{3.5}$$

$A_{2,2}$ $i, \ldots, j$

$$\begin{bmatrix} A_{2,1} & A_{2,2} - \Lambda \\ A_{3,1} & A_{3,2} \end{bmatrix} = \text{Rk } r, \tag{3.6}$$

r 3.1

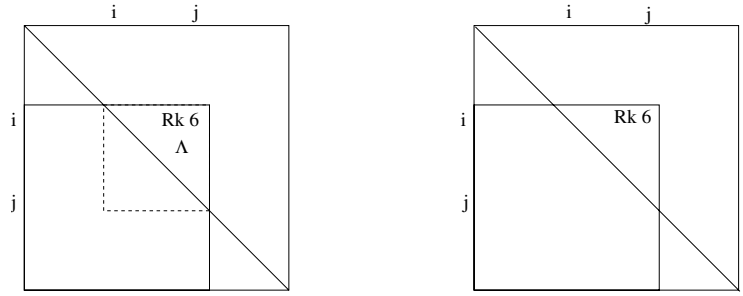


FIG. 3.1. *The structure block $\mathcal{B}_{\text{weak}}$ in the left-hand picture has the following meaning: after subtracting the shift matrix $\Lambda \in \mathbb{C}^{4\times 4}$ from the dashed square submatrix in the middle, the indicated bottom left submatrix must be of rank at most 6. The structure block $\mathcal{B}_{\text{weak,pure}}$ in the right-hand picture is a special case of this, with $\Lambda = 0$.*

$\Lambda = \Lambda_{\text{fin}} \oplus \infty I$ Λ_{fin}
$\mathcal{B}_{\text{weak}}$ ”
∞ r
3.2
$\Lambda = 0 \oplus \infty I$ pure $\mathcal{B}_{\text{weak,pure}}$

THEOREM 13 (see [1, Corollary 16]). $A \in \mathbb{C}^{n\times n}$
$\mathcal{B}_{\text{weak}} = (i, j, r, \Lambda)$ $\Lambda = \Lambda_{\text{ns}} \oplus 0 \oplus \infty I$ Λ_{ns}
A^{-1} $\mathcal{B}^{-1}_{\text{weak}} :=$
(i, j, r, Λ^{-1}) $\Lambda^{-1} := \Lambda^{-1}_{\text{ns}} \oplus \infty I \oplus 0$ $\frac{1}{0} = \infty$ $\frac{1}{\infty} = 0$

Let us recall also that by absorbing permutation matrices into the structure, structure blocks can be moved to any matrix position, not necessarily situated in the bottom left matrix corner anymore [1].

Now we come to the actual definition of structure blocks in the context of Schur complements. Such structure blocks will be denoted as just $\mathcal{B}$, hence dropping the subscript

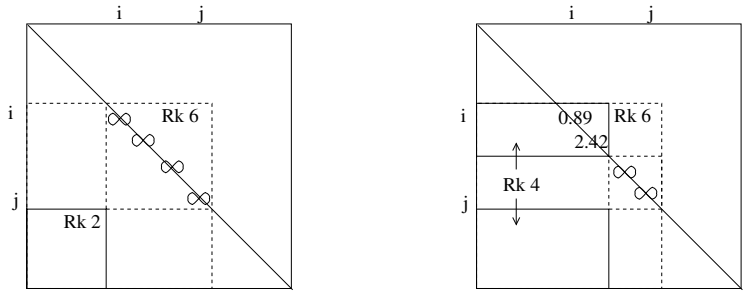


FIG. 3.2. *The structure block* $\mathcal{B}_{\text{weak,pure}}$ *in the left picture has shift matrix* $\Lambda = \infty I_4$. *Hence by definition, it should be identified with the* Rk 2 *structure block in the bottom left corner. The structure block* $\mathcal{B}_{\text{weak}}$ *in the right picture has* $\Lambda = \text{diag}(0.89, 2.42, \infty, \infty)$. *Hence it should be identified with the smaller* Rk 4 *structure block, consisting of two pieces. Note that the shift submatrix* $\Lambda_{\text{fin}} := \text{diag}(0.89, 2.42)$ *is inherited.*

DEFINITION 14. $A \in \mathbb{C}^{m\times n}$ $k \in \mathbb{N}$ k A 1 k

$$\mathcal{B} = (I, J, I_\Lambda, J_\Lambda, r, \Lambda),$$

I J r $\Lambda \in \mathbb{C}^{|I_\Lambda|\times|J_\Lambda|}$ $I_\Lambda \subseteq I$ $J_\Lambda \subseteq J$ $I = I_1 \cup I_2$ $J = J_1 \cup J_2$ k

$$\Lambda = \begin{bmatrix} \Lambda_{1,1} & \Lambda_{1,2} \\ \Lambda_{2,1} & \Lambda_{2,2} \end{bmatrix},$$

$\Lambda_{1,1}$ $|J_1| - |I_1| + 1$ A $\mathcal{B}$

$$\tilde{A}(I, J) = \text{Rk } r,$$

$\tilde{A}$ A $\tilde{A}(I_\Lambda, J_\Lambda) = A(I_\Lambda, J_\Lambda) - \Lambda$ 3.3

Figure 3.3 illustrates that a given structure block $\mathcal{B}$ can be considered as a collection of four individual parts w.r.t. the given k-partitioning. Therefore we will sometimes refer to $\mathcal{B}$ as a structure block: this will be clear from the context. Moreover, note that the size restriction in Definition 14 expresses precisely that the top left part of the structure block is a structure block according to Definition 12, at least up to permutation.

Now we come to the preservation of structure blocks under Schur complementation.

THEOREM 15. $A \in \mathbb{C}^{m\times n}$ k A $\mathcal{B}$ k 14 $\Lambda_{1,1}$ $S_{A,k}$ $S_{\mathcal{B}} := (I_2, J_2, I_{\Lambda,2}, J_{\Lambda,2}, r, S_\Lambda)$ $S_\Lambda := \Lambda_{2,2} - \Lambda_{2,1}\Lambda_{1,1}^{-1}\Lambda_{1,2}$ 3.4

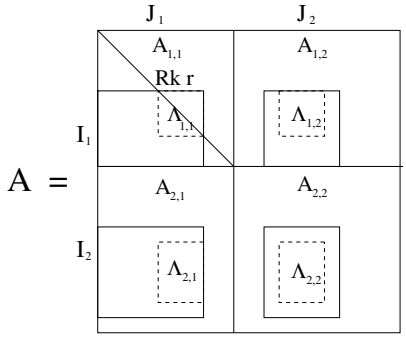


FIG. 3.3. *Consider a matrix A together with a k-partitioning of A, which is visualized by the horizontal and vertical lines in the figure. The figure shows an example of a structure block $\mathcal{B}$ satisfied by this matrix. The meaning is that after subtracting the shift matrix Λ (consisting of four parts) from the dashed matrix positions, the indicated submatrix $A(I, J)$ (also consisting of four parts) must be of rank at most r.*

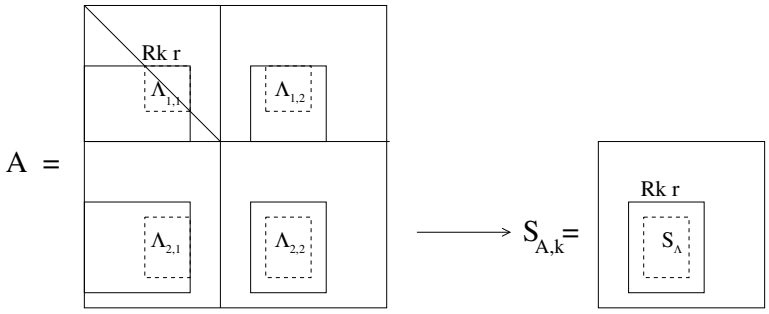


FIG. 3.4. *Consider the matrix in the left-hand side, satisfying the huge structure block $\mathcal{B}$, consisting of four parts. Then the Schur complement $S_{A,k} = A_{2,2} - A_{2,1}A_{1,1}^{-1}A_{1,2}$ inherits this structure block, with new shift matrix given by $S_\Lambda := \Lambda_{2,2} - \Lambda_{2,1}\Lambda_{1,1}^{-1}\Lambda_{1,2}$.*

. By definition of structure block, there exists a matrix B having the form

$$B = P \left[\begin{array}{ccc|ccc} X & X & X & X & X & X \\ 0 & \Lambda_{1,1} & X & 0 & \Lambda_{1,2} & X \\ 0 & 0 & X & 0 & 0 & X \\ \hline X & X & X & X & X & X \\ 0 & \Lambda_{2,1} & X & 0 & \Lambda_{2,2} & X \\ 0 & 0 & X & 0 & 0 & X \end{array}\right] \tilde{P}$$

for certain permutation matrices $P = P_1 \oplus P_2$ and $\tilde{P} = \tilde{P}_1 \oplus \tilde{P}_2$, such that $A - B = \operatorname{Rk} r$. By Corollary 9, it follows that $S_{A,k} - S_{B,k} = \widetilde{\operatorname{Rk}}\, r$. But by the form of B, it is easy to see that its Schur complement satisfies

$$S_{B,k} = P_2 \left[\begin{array}{ccc} X & X & X \\ 0 & S_\Lambda & X \\ 0 & 0 & X \end{array}\right] \tilde{P}_2,$$

where $S_\Lambda = \Lambda_{2,2} - \Lambda_{2,1}\Lambda_{1,1}^{-1}\Lambda_{1,2}$. It follows that $S_{A,k}$ satisfies the structure block $S_{\mathcal{B}} = (I_2, J_2, I_{\Lambda,2}, J_{\Lambda,2}, r, S_\Lambda)$, as we had to prove. ∎

As an illustrative example, suppose that

$$A_{i,j} = \begin{bmatrix} \times & \times & \times \\ 1 & 1+\lambda_{i,j} & \times \\ 1 & 1 & \times \end{bmatrix}$$

for $i, j = 1, 2$. Then we claim that $A_{2,2} - A_{2,1}A_{1,1}^{-1}A_{1,2}$ (if $A_{1,1}^{-1}$ exists) will satisfy the structure block $\mathcal{B}_{\text{weak}} : (i, j, r, \lambda) = (2, 2, 1, S_\lambda)$, with new shift element defined by $S_\lambda := \lambda_{2,2} - \lambda_{2,1}\lambda_{1,1}^{-1}\lambda_{1,2}$ (if $\lambda_{1,1}^{-1}$ exists). Indeed, the proof follows immediately from Theorem 15 by working with the embedded matrix

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

and observing that the given data can be translated in terms of a huge structure block $\mathcal{B}$ on A.

Note that in this last example, it was necessary that the low rank blocks $\left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right]$ of the several matrices $A_{i,j}$ were "compatible" with each other. If this were not the case, it could be that the rank upper bound of the huge structure block $\mathcal{B}$ (and hence of the Schur complement $S_{A,k}$) must be increased.

A way to avoid the latter problem is to choose several of the low rank blocks equal to zero. Suppose, for example, that $A_{1,1} := T$ is a given matrix, satisfying a given structure block $\mathcal{B}_{\text{weak}}$. Suppose that we choose $A_{1,2}$, $A_{2,2}$, and $A_{2,1}$ with sparse bottom left parts as illustrated in Figure 3.5. Then it is clear that the structure block $\mathcal{B}_{\text{weak}}$ can be extended to a huge structure block $\mathcal{B}$ in the matrix A, with new shift matrix

$$\Lambda = \begin{bmatrix} \Lambda_{1,1} & -I \\ I & 0 \end{bmatrix}.$$

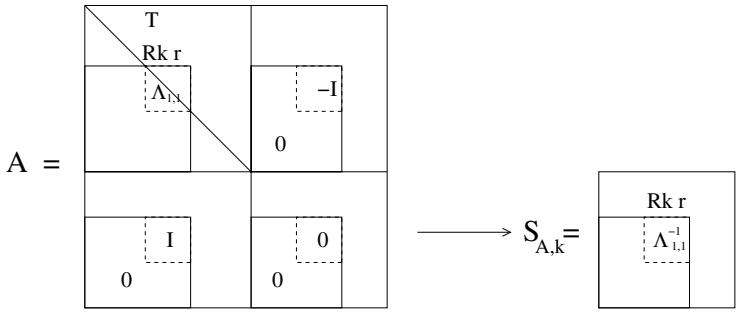


FIG. 3.5. *Consider the matrix $A_{1,1} := T$ in the top left position, satisfying a structure block $\mathcal{B}_{\text{weak}}$. Then the data in the figure imply a huge structure block $\mathcal{B}$ in the matrix A. Hence the Schur complement $S_{A,k}$ will inherit the structure block $S_{\mathcal{B}}$, with new shift matrix $S_\Lambda \equiv \Lambda_{1,1}^{-1}$.*

We can derive two things from this example. First, consider the special case where $A_{1,2} = -I$, $A_{2,2} = 0$, and $A_{2,1} = I$. Then Figure 3.5 shows that $S_A \equiv T^{-1}$ satisfies the structure block $S_{\mathcal{B}} \equiv \mathcal{B}_{\text{weak}}^{-1}$, which is precisely the structure block inversion result of [1, Theorem 11]. Second, we can interpret Figure 3.5 in the following way: it can be used to generate matrices $A_{1,2}$, $A_{2,2}$, and $A_{2,1}$ such that $A_{2,2} - A_{2,1}T^{-1}A_{1,2}$ inherits the structure of T^{-1}. We can interpret this as a set of [illegible] for T^{-1}.

To conclude this section, we want to relax the nonsingularity condition in Theorem 15. At the same time we want to introduce shift elements equal to ∞, in the sense of Definition 12. Here we will restrict ourselves to the case where the following assumption holds.

. $\Lambda_{1,1} = \Lambda_{\text{ns}} \oplus \infty I \oplus 0_l$, where Λ_{ns} is square and nonsingular, and with 0_l being the zero matrix of size l by l. The other parts $\Lambda_{1,2}$, $\Lambda_{2,1}$, and $\Lambda_{2,2}$ are not allowed to contain elements equal to ∞.

Now let us write

$$\Lambda = \left[\begin{array}{c|c} \Lambda_{1,1} & \Lambda_{1,2} \\ \hline \Lambda_{2,1} & \Lambda_{2,2} \end{array}\right] = \left[\begin{array}{cc|c} \Lambda_{1,1}^{TL} & 0 & \Lambda_{1,2}^{T} \\ 0 & 0_l & \Lambda_{1,2}^{B} \\ \hline \Lambda_{2,1}^{L} & \Lambda_{2,1}^{R} & \Lambda_{2,2} \end{array}\right], \tag{3.7}$$

where $\Lambda_{1,1}^{TL} := \Lambda_{\text{ns}} \oplus \infty I$, and where the superscripts T, B, L, and R denote the top, bottom, left, and right parts of the corresponding matrices. It is easy to see that the Schur complement of (3.7) can be written as a "dyadic decomposition,"

$$S_\Lambda = S_{\text{fin}} + S_\infty, \tag{3.8}$$

where S_{fin} and S_∞ are the Schur complements of the respective matrices

$$\left[\begin{array}{cc} \Lambda_{1,1}^{TL} & \Lambda_{1,2}^{T} \\ \Lambda_{2,1}^{L} & \Lambda_{2,2} \end{array}\right], \quad \left[\begin{array}{cc} 0_l & \Lambda_{1,2}^{B} \\ \Lambda_{2,1}^{R} & 0 \end{array}\right]. \tag{3.9}$$

Here S_{fin} contains only finite elements, and hence this will just be a . correction term to the structure of $S_{A,k}$. The problem is instead to determine the meaning of S_∞.

To achieve this, we will suppose that operations have been applied on the second block row and column of A, such that

$$\left[\begin{array}{c|c} 0_l & \Lambda_{1,2}^{B} \\ \hline \Lambda_{2,1}^{R} & 0 \end{array}\right] = \left[\begin{array}{c|cc} 0_l & 0 & \Lambda_{\text{ind col}} \\ \hline 0 & 0 & 0 \\ \Lambda_{\text{ind row}} & 0 & 0 \end{array}\right], \tag{3.10}$$

where $\Lambda_{\text{ind col}}$ contains independent columns and $\Lambda_{\text{ind row}}$ contains independent rows. (Here the row and column operations which we applied on A to achieve (3.10) have a well-determined effect on the Schur complement $S_{A,k}$ by virtue of Lemma 2.)

Then we have the following result.

THEOREM 16. . $A \in \mathbb{C}^{m\times n}$ k A $\mathcal{B}$ k $\Lambda_{1,1} = \Lambda_{\text{ns}} \oplus \infty I \oplus 0_l$ S_{fin} S_∞ (3.8) (3.10)

$$S_{A,k}(I_{\Lambda,2}, J_{\Lambda,2}) := S_{A,k}(I_{\Lambda,2}, J_{\Lambda,2}) - S_{\text{fin}}$$

$I_{\Lambda,2}$ $J_{\Lambda,2}$ (3.10) $S_{A,k}$ $r - l$ 3.6

. If necessary, we can virtually add extra rows and columns to A until the blocks $\Lambda_{\text{ind row}}$ and $\Lambda_{\text{ind col}}$ in (3.10) become square and nonsingular (of size l by l). Then since S_∞ was defined as the Schur complement of (3.10), and by approximating 0_l as $0_l = \lim_{\epsilon\to 0} \epsilon I$, we obtain

$$S_\infty = \left[\begin{array}{cc} 0 & 0 \\ 0 & \Lambda_{\text{ind row}}(\infty I_l)\Lambda_{\text{ind col}} \end{array}\right].$$

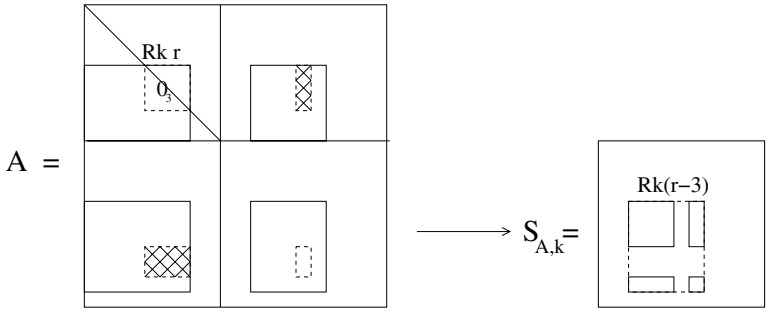


FIG. 3.6. *Consider the matrix in the left-hand side, satisfying the huge structure block* $\mathcal{B}$ *with* $\Lambda_{1,1} = 0_3$. *For the other parts of the shift matrix, the places where the nonzero elements act are indicated with a cross. (We assume here for simplicity of illustration that the finite correction term* S_{fin} *in* (3.8) *is equal to zero.) Then the Schur complement* $S_{A,k}$ *satisfies an* $\text{Rk}(r-3)$ *structure block consisting of four parts, as indicated.*

But by our knowledge of the meaning of shift elements ∞, this means that we should drop from $S_{A,k}$ all l rows and columns where ∞ is standing, and decrease the rank upper bound r by this same number l. The theorem now follows. □

As an illustrative example, suppose that

$$A_{i,j} = \begin{bmatrix} \times & \times & \times \\ 1 & 1 & \times \\ 1 & 1 & \times \end{bmatrix}$$

for $i, j = 1, 2$. Then we claim that $A_{2,2} - A_{2,1}A_{1,1}^{-1}A_{1,2}$ (if $A_{1,1}^{-1}$ exists) will satisfy the structure block $\mathcal{B}_{\text{weak}} : (i, j, r) = (2, 2, 0)$, i.e., that

$$A_{2,2} - A_{2,1}A_{1,1}^{-1}A_{1,2} = \begin{bmatrix} \times & \times & \times \\ 0 & 0 & \times \\ 0 & 0 & \times \end{bmatrix}.$$

Indeed, this follows from Theorem 16 by working with the embedded matrix A (as usual), and by observing that the given data can be translated in terms of a huge structure block $\mathcal{B}$ with shift matrix

$$\begin{bmatrix} \lambda_{1,1} & \lambda_{1,2} \\ \lambda_{2,1} & \lambda_{2,2} \end{bmatrix} \equiv \begin{bmatrix} 0_1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Thus indeed the rank upper bound r decreases by the value $l = 1$.

Note that in this last example, it was again necessary that the low rank blocks $\left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right]$ of the several matrices $A_{i,j}$ were compatible with each other.

A way to avoid the latter problem is to choose several of the low rank blocks equal to zero. Suppose, for example, that $A_{1,1} := T$ is a given matrix satisfying a given structure block $\mathcal{B}_{\text{weak,pure}}$ with $\Lambda = 0_3$. Suppose that we choose $A_{1,2}$, $A_{2,2}$, and $A_{2,1}$ with zero bottom left parts as illustrated in Figure 3.7. Then it is clear that the structure block $\mathcal{B}_{\text{weak,pure}}$ can [illegible] be extended to a huge structure block $\mathcal{B}_{\text{pure}}$ in the matrix A, with new shift matrix

$$\begin{bmatrix} 0_3 & 0 \\ 0 & 0 \end{bmatrix}.$$

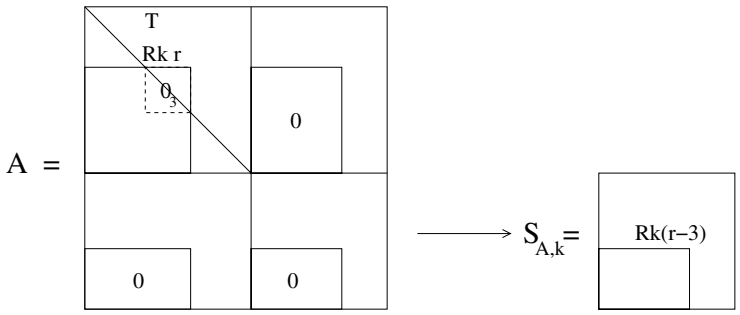


FIG. 3.7. *Specification of Figure* 3.5 *in the case of zero shift matrices.*

We can derive two things from this example. First, consider the special case where $A_{1,2} = -I$, $A_{2,2} = 0$, and $A_{2,1} = I$. Then Figure 3.7 shows that $S_A \equiv T^{-1}$ satisfies the structure block $S_{\mathcal{B}} \equiv \mathcal{B}^{-1}_{\text{weak,pure}}$, which is precisely the structure block inversion result of [1, Corollary 16] (concerning shift matrices $\Lambda_{\text{ns}} \oplus 0 \oplus \infty I$). Second, we can interpret Figure 3.7 in the following way: it can be used to generate matrices $A_{1,2}$, $A_{2,2}$, and $A_{2,1}$ such that $A_{2,2} - A_{2,1}T^{-1}A_{1,2}$ inherits the structure of T^{-1}. We can interpret this as a set of [illegible] for T^{-1}. Other examples of structure-preserving transformations will be given in the next section, in the context of Möbius transformations.

We conclude this section with a final remark.

REMARK 17.

1. [illegible] $\mathcal{B}$ [illegible] 2 [illegible] A [illegible] $\begin{bmatrix} X & 0 \\ X & I \end{bmatrix}$ [illegible] $\Lambda_{1,1}$ [illegible] $\Lambda_{1,1}$ [illegible] 14 [illegible]
2. [illegible] [1] $I_1 = \emptyset$ $J_1 = \{1, \ldots, k\}$ [illegible] $\Lambda_{2,1} = 0$ [illegible] 3.8 [illegible] $\Lambda_{1,1}$ [illegible] 14 [illegible] " [illegible] 3.8

Concerning Remark 17(2), note that the preservation under Schur complementation of semiseparable plus diagonal structure (not including diagonal elements) was also shown in [4], where it was used as the basis for a fast solver. The word "semiseparable" refers in this case to a matrix whose strictly lower and strictly upper triangular part both come from a (possibly different) rank-r matrix, $r \in \mathbb{N}$.

4. Möbius and Cayley transformations. In this section we will focus on Möbius transformations, as an illustration of the results on Schur complements in the previous section. Möbius transformations appear also under the name of [illegible]. As a general reference, we can refer to [8] for the treatment of Möbius transformations with scalar coefficients, and to [12, 13] for the general case of matrix-valued coefficients. Most of the results which we state without proof can be found there.

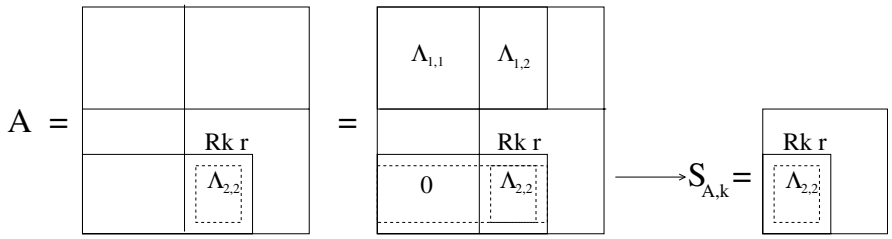


FIG. 3.8. *Consider the structure block in the leftmost picture. Then we can "enlarge" this structure block by redefining* $I_1 := \{1, \ldots, k\}$, $\Lambda_{1,1} := A_{1,1}$, *and* $\Lambda_{1,2} := A_{1,2}|_{I_1 \times J_2}$, *as illustrated in the middle picture. In this way the resulting structure block will still be of rank at most* r, *and the size restrictions on* $\Lambda_{1,1}$ *occurring in Definition* 14 *are restored. But then it follows that* $S_{A,k}$ *satisfies the new structure block* $S_{\mathcal{B}} = (I_2, J_2, I_{\Lambda_2}, J_{\Lambda_2}, r, \Lambda_{2,2})$.

We start with a definition.

DEFINITION 18. $P, Q, R, S \in \mathbb{C}^{n \times n}$ Möbius transformation $\mathbb{C}^{n \times n}$

$$\mathcal{M} : A \mapsto (PA + Q)(RA + S)^{-1}.$$

dual Möbius transformation

$$A \mapsto (AP + R)^{-1}(AQ + S).$$

$$\begin{bmatrix} P & Q \\ R & S \end{bmatrix}$$

$\mathcal{M}$ $\mathcal{M}$ invertible

Unless explicitly mentioned, we will always work with usual Möbius transformations, rather than with their dual versions.

Note that the Möbius transformation is defined only on the domain $\mathcal{D} := \{A \in \mathbb{C}^{n \times n} | \det(RA + S) \neq 0\}$. Since the domain $\mathcal{D}$ is defined by the nonvanishing of an algebraic equation, it is either empty (a case we exclude) or a dense subset of $\mathbb{C}^{n \times n}$.

The use of the matrix associated with $\mathcal{M}$ follows by rewriting $\mathcal{M}(A) = N_{\mathcal{M}(A)} D_{\mathcal{M}(A)}^{-1}$ with

$$\begin{bmatrix} N_{\mathcal{M}(A)} \\ D_{\mathcal{M}(A)} \end{bmatrix} := \begin{bmatrix} P & Q \\ R & S \end{bmatrix} \begin{bmatrix} A \\ I \end{bmatrix}. \tag{4.1}$$

This matrix representation is useful in several aspects. For example, it can be checked that for given Möbius transformations $\mathcal{M}_1$ and $\mathcal{M}_2$, the composed map $A \mapsto \mathcal{M}_2(\mathcal{M}_1(A))$ is again a Möbius transformation, with associated matrix

$$\begin{bmatrix} P_2 & Q_2 \\ R_2 & S_2 \end{bmatrix} \begin{bmatrix} P_1 & Q_1 \\ R_1 & S_1 \end{bmatrix}.$$

Since the identity map $A \mapsto A$ is a special case of a Möbius transformation, with associated matrix $\left[\begin{smallmatrix} I & 0 \\ 0 & I \end{smallmatrix}\right]$, it follows that the $\mathcal{M}^{-1}$

will have as its associated matrix precisely $\left[\begin{smallmatrix} P & Q \\ R & S \end{smallmatrix}\right]^{-1}$. Moreover, denoting with $\mathcal{D}$ the domain of $\mathcal{M}$, it can be shown that $\mathcal{M}^{-1}$ has as its domain $\mathcal{M}(\mathcal{D})$ and as its range $\mathcal{D}$.

We may note here that $\mathcal{M}^{-1}$ can also be obtained by directly solving for A in terms of $\mathcal{M}(A)$ in Definition 18. This yields $\mathcal{M}^{-1} : B \mapsto -(BR - P)^{-1}(BS - Q)$, which is not a Möbius transformation anymore but rather a dual Möbius transformation in the sense of Definition 18. In particular, it follows that every invertible Möbius transformation can be expressed as a dual Möbius transformation too.

Now note that $\mathcal{M}(A)$ can be realized as the Schur complement of

$$\begin{bmatrix} RA + S & -I \\ PA + Q & 0 \end{bmatrix}. \tag{4.2}$$

In particular, we can prove the following result.

THEOREM 19. [illegible] $\mathcal{M}$ [illegible] $\mathcal{D}$ [illegible] $A \in \mathcal{D}$ $A + \operatorname{Rk} r \in \mathcal{D}$ [illegible] $\operatorname{Rk} r$ [illegible] r [illegible]

$$\mathcal{M}(A + \operatorname{Rk} r) = \mathcal{M}(A) + \widetilde{\operatorname{Rk}}\, r,$$

[illegible] $\widetilde{\operatorname{Rk}}\, r$ [illegible] r [illegible]. Let us write $\operatorname{Rk} r = UV^H$ with $U, V \in \mathbb{C}^{n\times r}$. Note that $\mathcal{M}(A + UV^H)$ can be realized as the Schur complement of the following matrix:

$$\begin{bmatrix} RA + S & -I \\ PA + Q & 0 \end{bmatrix} + \begin{bmatrix} RU \\ PU \end{bmatrix} \begin{bmatrix} V^H & 0 \end{bmatrix}.$$

Since the latter matrix is an Rk r correction of (4.2), the result follows by Corollary 9. □

We come to a second topic.

DEFINITION 20. [illegible] $E, F, G \in \mathbb{C}^{n\times n}$ [illegible] E [illegible] G [illegible] quadratic transformation [illegible] $\mathbb{C}^{n\times n}$ [illegible]

$$\mathcal{Q} : A \mapsto A^H E A + A^H F^H + F A + G.$$

[illegible]

$$\begin{bmatrix} E & F^H \\ F & G \end{bmatrix} \tag{4.3}$$

[illegible]

The use of the matrix associated with the quadratic transformation follows by rewriting

$$\mathcal{Q}(A) = \begin{bmatrix} A^H & I \end{bmatrix} \begin{bmatrix} E & F^H \\ F & G \end{bmatrix} \begin{bmatrix} A \\ I \end{bmatrix}. \tag{4.4}$$

Moreover, note that by our assumption that E and G are Hermitian, both the middle matrix and the right-hand side of (4.4) must be Hermitian. In particular, it makes sense to speak about the [illegible] of $\mathcal{Q}(A)$.

To establish some inertia results, we will first prove the following lemma. Here we agree to label the eigenvalues of a matrix A in nonincreasing order as $\lambda_{A,k}$, $k = 1, \ldots, n$.

LEMMA 21. Herm T $H :=$ $T(\text{Herm})T^H$ $\tilde{H} = (T + \text{Rk } r)\text{Herm}(T + \text{Rk } r)^H$ Rk r r

$$\lambda_{H,k+r} \leq \lambda_{\tilde{H},k} \leq \lambda_{H,k-r}.$$

. We will prove the interlacing property under the slightly weaker condition that H and $\tilde{H}$ are both Hermitian and

$$\tilde{H} = H + UV^H + \tilde{V}U^H, \tag{4.5}$$

with $U, V, \tilde{V}$ in $\mathbb{C}^{n\times r}$. We recall that for any Hermitian matrix Herm, the eigenvalues can be determined by the Courant–Fisher characterization

$$\lambda_{\text{Herm},k} = \max_{\dim \mathcal{V}=k} \min_{x\in\mathcal{V}} \frac{x^H \text{Herm } x}{x^H x}, \tag{4.6}$$

with $\mathcal{V}$ running over all k-dimensional linear subspaces of $\mathbb{C}^n$. Taking such a fixed subspace $\mathcal{V}$ and taking $\text{Herm} = \tilde{H}$, we obtain by (4.5) that $\min_{x\in\mathcal{V}} \frac{x^H\tilde{H}x}{x^Hx} \leq \min_{x\in\mathcal{V}\cap\mathcal{U}} \frac{x^HHx}{x^Hx}$, where $\mathcal{U}$ denotes the $(n-r)$-dimensional linear subspace of $\mathbb{C}^n$ containing all vectors for which $U^Hx = 0$. Then since $\dim(\mathcal{V}\cap\mathcal{U}) \geq k-r$, we derive by (4.6) that

$$\min_{x\in\mathcal{V}} \frac{x^H\tilde{H}x}{x^Hx} \leq \lambda_{H,k-r}.$$

By taking the maximum over all $\mathcal{V}$, it follows that $\lambda_{\tilde{H},k} \leq \lambda_{H,k-r}$, as we had to prove. The other inequality follows by symmetry. ∎

REMARK 22. 21 (π, ν, ζ) $UV^H+\tilde{V}U^H$ (4.5) " $\max\{\pi, \nu\} \leq r$ 21 [6] [6]

COROLLARY 23.

1. Inertia$(\mathcal{Q}(A+\text{Rk } r))-$Inertia$(\mathcal{Q}(A)) = (\Delta\pi, \Delta\nu, \Delta\zeta)$ $\max\{|\Delta\pi|, |\Delta\nu|\} \leq r$
2. (4.3) (π, ν, ζ) $\pi = n + \delta$ $\delta \geq 1$ $\mathcal{Q}(A)$ δ A

1. Note that in (4.4), going over to $\mathcal{Q}(A + \text{Rk } r)$ corresponds to a rank-r correction of the factor $\left[\begin{smallmatrix} A \\ I \end{smallmatrix}\right]$. Hence the result follows from the previous lemma.
2. We will again use (4.4). First, by adding n extra columns, the matrix $\left[\begin{smallmatrix} A \\ I \end{smallmatrix}\right]$ can always be completed to a nonsingular $2n \times 2n$ matrix. Replacing $\left[\begin{smallmatrix} A \\ I \end{smallmatrix}\right]$ by this completed version, and replacing $\left[\begin{smallmatrix} A^H & I \end{smallmatrix}\right]$ by the Hermitian transpose of it, Sylvester's law of inertia implies that the right-hand side of (4.4) must still have inertia (π, ν, ζ) with $\pi = n + \delta$. The result then follows by again removing the n added columns (which is a rank-n perturbation) and applying the previous lemma. ∎

Since both Möbius and quadratic transformations allow what we called a matrix representation, we can expect that these transformations have a good behavior w.r.t. each other. Let us first introduce the following definition.

DEFINITION 24. [illegible] $\mathcal{Q}$ [illegible] $\mathbb{C}^{n\times n}$ [illegible] r [illegible] $A \in \mathbb{C}^{n\times n}$ [illegible] quadratic relation [illegible] $(\mathcal{Q}, r)$ [illegible]

$$\mathcal{Q}(A) = \operatorname{Rk} r$$

[illegible] $\operatorname{Rk} r$ [illegible] r

Since $\mathcal{Q}(A)$ is always Hermitian, we can in fact add any inertia condition on $\operatorname{Rk} r$ if so desired.

Now let $\mathcal{M}$ be a Möbius transformation with domain $\mathcal{D}$, let $A \in \mathcal{D}$, and let $B := N_B D_B^{-1} = \mathcal{M}(A)$. Then

$$\begin{aligned}
& \mathcal{Q}(B) = \operatorname{Rk} r \\
\Leftrightarrow\quad & \begin{bmatrix} B^H & I \end{bmatrix} \begin{bmatrix} E & F^H \\ F & G \end{bmatrix} \begin{bmatrix} B \\ I \end{bmatrix} = \operatorname{Rk} r \\
\Leftrightarrow\quad & \begin{bmatrix} N_B^H & D_B^H \end{bmatrix} \begin{bmatrix} E & F^H \\ F & G \end{bmatrix} \begin{bmatrix} N_B \\ D_B \end{bmatrix} = \widetilde{\operatorname{Rk}}\, r \\
\overset{\text{by (4.1)}}{\Leftrightarrow}\quad & \begin{bmatrix} A^H & I \end{bmatrix} \begin{bmatrix} P & Q \\ R & S \end{bmatrix}^H \begin{bmatrix} E & F^H \\ F & G \end{bmatrix} \begin{bmatrix} P & Q \\ R & S \end{bmatrix} \begin{bmatrix} A \\ I \end{bmatrix} = \widetilde{\operatorname{Rk}}\, r \\
\Leftrightarrow\quad & \begin{bmatrix} A^H & I \end{bmatrix} \begin{bmatrix} \tilde{E} & \tilde{F}^H \\ \tilde{F} & \tilde{G} \end{bmatrix} \begin{bmatrix} A \\ I \end{bmatrix} = \widetilde{\operatorname{Rk}}\, r \\
\Leftrightarrow\quad & \widetilde{\mathcal{Q}}(A) = \widetilde{\operatorname{Rk}}\, r,
\end{aligned}$$

where we define $\widetilde{\operatorname{Rk}}\, r := D_B^H(\operatorname{Rk} r)D_B$ (being a matrix with same rank and inertia as $\operatorname{Rk} r$), and where we define the quadratic transformation $\widetilde{\mathcal{Q}}$ by its associated matrix

$$\begin{bmatrix} \tilde{E} & \tilde{F}^H \\ \tilde{F} & \tilde{G} \end{bmatrix} := \begin{bmatrix} P & Q \\ R & S \end{bmatrix}^H \begin{bmatrix} E & F^H \\ F & G \end{bmatrix} \begin{bmatrix} P & Q \\ R & S \end{bmatrix}. \tag{4.7}$$

Let us give some illustrations to the above series of equivalences. First consider the class of Hermitian matrices defined by $\mathcal{Q}(A) := -iA + iA^H = \operatorname{Rk} 0$. Note that the matrix associated with $\mathcal{Q}$ is given by $\left[\begin{smallmatrix} 0 & iI \\ -iI & 0 \end{smallmatrix}\right]$. Then consider the class of J-unitary matrices, defined by $\mathcal{Q}(A) := A^H JA - J = \operatorname{Rk} 0$, where $J = I_r \oplus -I_s$ is a fixed signature matrix. The matrix associated with $\mathcal{Q}$ is given by $\left[\begin{smallmatrix} J & 0 \\ 0 & -J \end{smallmatrix}\right]$. Now we may observe that these two matrices obtained for Hermitian and J-unitary quadratic relations have the same inertia, namely $(n, n, 0)$. Hence by Sylvester's law of inertia, there exists a nonsingular congruence transformation which maps these matrices into each other. Indeed, one can check that

$$\begin{bmatrix} 0 & iI \\ -iI & 0 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} I & I \\ -iJ & iJ \end{bmatrix} \begin{bmatrix} J & 0 \\ 0 & -J \end{bmatrix} \begin{bmatrix} I & iJ \\ I & -iJ \end{bmatrix} \frac{1}{\sqrt{2}}.$$

Comparing this with (4.7), we conclude that

$$\mathcal{M} : A \mapsto (A + iJ)(A - iJ)^{-1} \tag{4.8}$$

is a Möbius transformation mapping the class of Hermitian matrices into the class of J-unitary matrices (at least for every Hermitian matrix belonging to the domain $\mathcal{D}$ of $\mathcal{M}$). Moreover, one can check that

$$\mathcal{M}^{-1} : A \mapsto i(A+I)(JA-J)^{-1}.$$

REMARK 25. (4.8)

For the rest of this section, we will restrict ourselves to Möbius transformations with scalar coefficients, i.e., with P, Q, R, and S being scalar multiples of the identity matrix. We will denote these multiples as pI, qI, rI, and sI.

As in the general case, note that $\mathcal{M}(A)$ can be realized as the Schur complement of the embedded matrix

$$\begin{bmatrix} rA+sI & -I \\ pA+qI & 0 \end{bmatrix}. \tag{4.9}$$

Now assume that A satisfies a weak structure block $\mathcal{B}_{\text{weak}} = (i, j, r, \Lambda)$. We want to show that this weak structure block can be extended to a huge structure block in (4.9), with same rank upper bound r. For this, let $A_{\text{pure}} := (A-(0\oplus\Lambda\oplus 0))|_{\mathcal{B}_{\text{weak}}}$. Note that $\text{Rank}\begin{bmatrix} rA_{\text{pure}} \\ pA_{\text{pure}} \end{bmatrix} = \text{Rank } A_{\text{pure}}$. This means that $\mathcal{B}_{\text{weak}}$ can indeed be extended to a huge structure block, denoted $\mathcal{B}_{\text{huge}}$, and with corresponding shift matrix

$$\Lambda_{\text{huge}} := \begin{bmatrix} r\Lambda+sI & -I \\ p\Lambda+qI & 0 \end{bmatrix}.$$

(Here we did not show all the zero blocks of Λ_{huge}; see Figure 4.1 for a more accurate picture.)

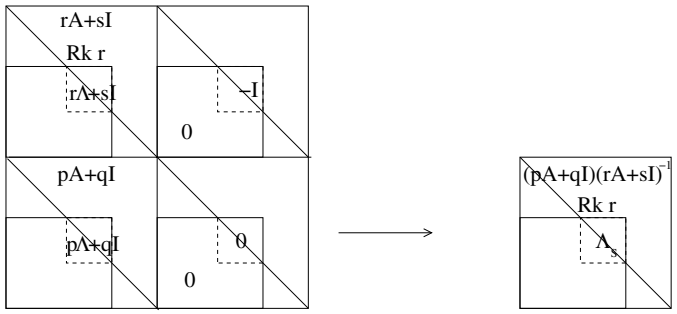


FIG. 4.1. *Consider a matrix A satisfying an* Rk r *weak structure block with shift matrix Λ. Let $p, q, r, s \in \mathbb{C}$ be arbitrary numbers. Then the Möbius transformation $(pA+qI)(rA+sI)^{-1}$ can be obtained in the form of a Schur complement of an embedded matrix A, as illustrated. Hence it will inherit the* Rk r *weak structure block, with new shift matrix $(p\Lambda+qI)(r\Lambda+sI)^{-1}$.*

Now by Theorem 15, the Schur complement of (4.9) must inherit $\mathcal{B}_{\text{huge}}$, with new shift matrix being the Schur complement of Λ_{huge}. Reformulating this in terms of the original data, we obtain the following theorem.

THEOREM 26. [illegible] $A \in \mathbb{C}^{n\times n}$ [illegible] $\mathcal{B}_{\text{weak}}$ [illegible] $\mathcal{M}(A)$ [illegible] $\mathcal{B}_{\text{weak}}$ [illegible] $\mathcal{M}(\Lambda)$ [illegible] Λ

We may mention that Theorem 26 was already shown in [14] for the case of lower semiseparable plus diagonal matrices. See also [3].

For this same example, note that for a [illegible] of matrices A, B satisfying the same weak structure block $\mathcal{B}_{\text{weak}} = (i, j, r, \Lambda)$, the "decoupled scalar Möbius transformation" $(pA+qI)(rB+sI)^{-1}$ will essentially inherit the weak structure block $\mathcal{B}_{\text{weak}}$, but now with rank upper bound only bounded by $2r$. The reason for this is the identity $\text{Rank}\begin{bmatrix} rA_{\text{pure}} \\ pB_{\text{pure}} \end{bmatrix} \leq \text{Rank } A_{\text{pure}} + \text{Rank } B_{\text{pure}}$, which is much weaker than in the case $A_{\text{pure}} = B_{\text{pure}}$.

An interesting case of a scalar Möbius transformation is by taking $J = I$ in (4.8). Then (4.8) reduces to the well-known [illegible] $\mathcal{C} : A \mapsto (A + iI)(A-iI)^{-1}$, mapping Hermitian into unitary matrices. Unlike the general situation, the domain $\mathcal{D}$ of the Cayley transformation contains the [illegible] class of Hermitian matrices.

The Cayley transformation can be used to derive several properties of unitary matrices. We already remarked in [1] that the weak structure blocks of such a matrix always come in pairs, i.e., that the presence of one such weak structure block always implies the presence of a second weak structure block, which we can easily determine. The underlying reason was the fact that $\text{Uni}^{-1} = \text{Uni}^H$ for any unitary matrix Uni, together with the inversion theorem for weak structure blocks. Another way to see this is by using the Cayley transformation. This transformation can be used to establish the property that the structure blocks of a unitary matrix always come in pairs, from the corresponding property that the structure blocks of a [illegible] matrix always come in pairs (for obvious reasons). We will not go further into this.

The Cayley transformation can also be used as a tool to prove a result similar to the following theorem from [1].

THEOREM 27. [illegible] $r \in \mathbb{N}$ [illegible]

(i) $A = \text{Herm} + \text{Rk } r$ [illegible] A [illegible] r.

(ii) $\mathcal{Q}(A) := i(A - A^H) = \text{Rk } 2r$ [illegible] $\text{Rk } 2r$ [illegible] $2r$ [illegible] $\text{Inertia}(\text{Rk } 2r) = (\pi, \nu, \zeta)$ [illegible] $\max\{\pi, \nu\} \leq r$

We obtain the following, similar formulation.

THEOREM 28. [illegible] $r \in \mathbb{N}$ [illegible]

(i) $A = \text{Uni} + \text{Rk } r$ [illegible] A [illegible] r.

(ii) $\mathcal{Q}(A) := A^H A - I = \text{Rk } 2r$ [illegible] $\text{Rk } 2r$ [illegible] $2r$ [illegible] $\text{Inertia}(\text{Rk } 2r) = (\pi, \nu, \zeta)$ [illegible] $\max\{\pi, \nu\} \leq r$

[illegible] The implication (i) $\Rightarrow$ (ii) is a special case of Corollary 23(1). For the implication (ii) $\Rightarrow$ (i), suppose that A is such that $A^H A - I = \text{Rk } 2r$ with $\text{Inertia}(\text{Rk } 2r) = (\pi, \nu, \zeta)$ with $\max\{\pi, \nu\} \leq r$. Denoting by $\mathcal{D} \subseteq \mathbb{C}^{n\times n}$ the domain of the Cayley transformation $\mathcal{C}$, we will suppose that $A \in \mathcal{C}(\mathcal{D})$, the domain of the inverse Cayley transformation $\mathcal{C}^{-1}$. (This can always be realized by multiplying with a suitable number $e^{i\theta}$, $\theta \in \mathbb{R}$.) Then we claim that $B := \mathcal{C}^{-1}(A)$ will satisfy $iB - iB^H = \widetilde{\text{Rk } 2r}$, with $\widetilde{\text{Rk } 2r}$ having the same inertia as $\text{Rk } 2r$. Indeed, this follows from the series of equivalences preceding (4.7). Hence by Theorem 27, we can factorize $B = \text{Herm} + \text{Rk } r$. The result follows then by applying $\mathcal{C}$ on both sides of this equation. (Here it is essential that for the application of Theorem 19, $\mathcal{C}(\text{Herm})$ is always defined, independent of the precise form of this Hermitian component Herm!) ∎

COROLLARY 29. [illegible] $r \in \mathbb{N}$ [illegible] r [illegible]

By combining Theorems 27 and 28 with an affine transformation $A \mapsto pA + qI$, these theorems can in fact be generalized to normal matrices with eigenvalues lying on a fixed [illegible] in $\mathbb{C}$, i.e., either a straight line or a circle in the complex plane. We must then work with a quadratic transformation on $\mathbb{C}^{n\times n}$ of the form $\mathcal{Q}(A) := eA^HA + \bar{f}A^H + fA + gI = 0$ for suitable numbers $e, g \in \mathbb{R}$ and $f \in \mathbb{C}$.

We may note here that these theorems [illegible] be generalized to arbitrary, non-scalar quadratic relations. For example, it can be shown that the class of J-unitary plus rank at most 1 matrices is [illegible] topologically closed.

5. Conclusion. In this paper, we investigated some structures that have a good behavior under Schur complementation. We handled two classes: displacement and rank structures. For displacement structures we derived in a direct way the preservation of structure, leading to formulae which extend the classical displacement tools. For the case of rank structures we showed how the preservation results could be used as a general framework to specify structure-preserving operations. In particular, we considered the Möbius transformation of a matrix and derived several structure preservation results.

REFERENCES

[1] S. DELVAUX AND M. VAN BAREL, *Structures preserved by matrix inversion*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 213–228.

[2] S. DELVAUX AND M. VAN BAREL, *Structures preserved by the QR-algorithm*, J. Comput. Appl. Math., 187 (2006), pp. 29–40.

[3] L. GEMIGNANI, *A unitary Hessenberg QR-based algorithm via semiseparable matrices*, J. Comput. Appl. Math., 184 (2005), pp. 505–517.

[4] I. C. GOHBERG, T. KAILATH, AND I. KOLTRACHT, *Linear complexity algorithms for semiseparable matrices*, Integral Equations Operator Theory, 8 (1985), pp. 780–804.

[5] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[6] D. A. GREGORY, B. HEYINK, AND K. N. VANDER MEULEN, *Inertia and biclique decompositions of joins of graphs*, J. Combin. Theory Ser. B, 88 (2003), pp. 135–151.

[7] G. HEINIG AND K. ROST, *Algebraic Methods for Toeplitz-Like Matrices and Operators*, Math. Res. 19, Akademie-Verlag, Berlin, 1984.

[8] P. HENRICI, *Applied and Computational Complex Analysis*, Wiley-Interscience, New York, 1974.

[9] T. KAILATH, S.-Y. KUNG, AND M. MORF, *Displacement ranks of matrices and linear equations*, J. Math. Anal. Appl., 68 (1979), pp. 395–407.

[10] T. KAILATH AND A. H. SAYED, *Displacement structure: Theory and applications*, SIAM Rev., 37 (1995), pp. 297–386.

[11] T. KAILATH AND A. H. SAYED, EDS., *Fast Reliable Algorithms for Matrices with Structure*, SIAM, Philadelphia, 1999.

[12] V. P. POTAPOV, *The multiplicative structures of J-contractive matrix functions*, in Amer. Math. Soc. Transl. Ser. (2) 15, 1960, pp. 131–243.

[13] V. P. POTAPOV, *Linear fractional transformations of matrices*, in Amer. Math. Soc. Transl. Ser. (2) 138, 1988, pp. 21–35.

[14] M. VAN BAREL, D. FASINO, L. GEMIGNANI, AND N. MASTRONARDI, *Orthogonal rational functions and structured matrices*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 810–829.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 1, pp. 253–263

BALANCING REGULAR MATRIX PENCILS*

DAMIEN LEMONNIER† AND PAUL VAN DOOREN†

Abstract. We present a new diagonal balancing technique for regular matrix pencils $\lambda B - A$, which aims at reducing the sensitivity of the corresponding generalized eigenvalues. It is inspired by the balancing technique of a square matrix A and has a comparable complexity. The diagonally scaled pencil has row and column norms that are balanced in a precise sense. We also show that balancing a pencil boils down to making it closer to some standardized normal pencil. We give numerical examples illustrating that the sensitivity of generalized eigenvalues of a pencil may significantly improve after balancing.

Key words. generalized eigenvalues, normal pencils, balancing

AMS subject classifications. 15A18, 15A22, 65F15, 65F35

DOI. 10.1137/S0895479804440931

1. Introduction. A matrix A with a norm that is several orders of magnitude larger than the modulus of its eigenvalues typically has eigenvalues that are sensitive to perturbations in the entries of A. It is shown in [4] that the Frobenius norm of a matrix can then often be reduced via a diagonal scaling of the type $D^{-1}AD$. Such a scaling can be performed in exact arithmetic if the diagonal elements are constrained to be integer powers of the base of the finite precision arithmetic (typically 2 or 10). As a consequence the eigenvalues do not change, but their sensitivity can significantly be reduced. Such a diagonal scaling is therefore typically used before running any eigenvalue algorithm.

In this paper we introduce a similar scaling method for square pencils $\lambda B - A$ with a determinant $\det(\lambda B - A)$ that is not identically zero for all values of λ. For such pencils—which are called [illegible]—one can define generalized eigenvalues via the zeros of the polynomial $\det(\lambda B - A)$. Our scaling method can be viewed as a natural extension of the balancing algorithm of [4] to regular matrix pencils and is aimed at reducing the sensitivity of the generalized eigenvalues of the pencil. This new method differs from that of Ward [7], whose aim it is to make the pencil entries have magnitudes as close to unity as possible, whereas our aim is to make the pencil as close as possible to some standardized normal pencil.

We first recall the classical balancing method for matrices and some of its properties. We then introduce the new balancing method for pencils and derive its analogous properties. We briefly discuss the complexity of the algorithm and finally give some numerical results illustrating the performance of the new scaling method.

2. Normal matrices and balancing. Normal matrices are known to have orthogonal eigenvectors and hence well conditioned eigenvalues [4]. Therefore if one has to compute eigenvalues of an arbitrary $n \times n$ matrix A, it is recommended to make

*Received by the editors February 10, 2004; accepted for publication (in revised form) by N. J. Higham November 23, 2005; published electronically March 17, 2006. This paper presents research results of the Belgian Programme on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office. The scientific responsibility rests with the authors.

http://www.siam.org/journals/simax/28-1/44093.html

†Université Catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium (lemonnier@csam.ucl.ac.be, vdooren@csam.ucl.ac.be).

it [illegible] to a normal matrix by an error free transformation. Diagonal [illegible] transformations with positive diagonal elements that are integer powers of the base can be performed exactly since they only amount to integer changes in the exponents of the matrix entries. In order to preserve the eigenvalues one performs diagonal similarities $D^{-1}AD$.

The basic question is thus how to characterize a diagonal scaling $D^{-1}AD$ that makes a matrix closer to a normal matrix. For this we consider two equivalent characterizations of normal matrices. A matrix A is normal iff

(1) A has [illegible] eigenvectors or, equivalently, its Schur form A_S

$$A_S := U^*AU, \quad U^*U = I_n \tag{2.1}$$

is a diagonal matrix Λ_A;

(2) the so-called defect from normality

$$\gamma(A) := \sum_{i=1}^{n} \sigma_i^2 - \sum_{i=1}^{n} |\lambda_i|^2 \tag{2.2}$$

is zero, where σ_i and λ_i are the singular values and the eigenvalues of A, respectively.

The defect from normality $\gamma(A)$ is always nonnegative [2], which easily follows from (2.1) and the fact $\gamma(A) = \gamma(A_S)$, since unitary similarities do not change the eigenvalues nor the singular values of a matrix. Let the [illegible] of A be the set of matrices similar to A. Then $\gamma(A)$ is the [illegible] between A and any normal matrix in the orbit of A.

THEOREM 2.1. [illegible]

$$\inf_T \|T^{-1}AT\|_F \tag{2.3}$$

[illegible] N_a [illegible] A [illegible] A [illegible] [illegible] T [illegible] $N_a = T^{-1}AT$. [illegible] T [illegible] [illegible]. Use the (complex) Schur decomposition $A_S = U^*AU$ and choose a unitary matrix Q such that the matrix $R := U^*TQ$ is triangular. Since unitary transformations do not change the Frobenius norm, the above minimization is then equivalent to

$$\inf_R \|R^{-1}A_SR\|_F,$$

which has the diagonal part Λ_A of A_S as solution. The transformation matrix R will be bounded for a diagonalizable matrix A, and it will be unbounded otherwise. (Also see [5] for more details). □

It then follows that $\sum_{i=1}^{n} \sigma_i^2$ is the Frobenius norm squared of A_S and therefore also the sum of the entries squared of A_S, while $\sum_{i=1}^{n} |\lambda_i|^2$ is just the sum of the diagonal entries squared of A_S. A diagonal scaling $D^{-1}AD$, on the other hand, does not change the λ_i but does modify the σ_i. So one can reduce the gap γ by scaling A in order to diminish its Frobenius norm. This is exactly what the balancing algorithm [4] does: it solves

$$\inf_D \|D^{-1}AD\|_F. \tag{2.4}$$

Let e_i denote the ith unit vector; then it is shown in [4] that the optimal scaling is achieved when $D^{-1}AD$ satisfies

$$\|(D^{-1}AD)e_i\|_2^2 = \|e_i^T(D^{-1}AD)\|_2^2 \qquad \forall\ i = 1 \dots n \tag{2.5}$$

and an algorithm is provided for computing an approximate solution D with elements that are powers of the base of the finite precision arithmetic. Each step of that algorithm decreases the Frobenius norm of the scaled matrix and hence also the distance to the normal matrices with the same spectrum as A.

The aim of this paper is to generalize these balancing ideas to regular matrix pencils $(\lambda B - A)$. In other words we will try to answer the following questions: (1) what is the property of regular pencils that is equivalent to normality in the standard eigenvalue problem, and (2) how to scale an arbitrary pencil so that it gets as close as possible to achieving this property?

3. Normal pencils. We first recall a definition of normal pencils, given in [1].

DEFINITION 3.1. $n \times n$ $\lambda B - A$

$$U_l^*(\lambda B - A)U_r = \lambda\Lambda_B - \Lambda_A,$$

U_l, U_r Λ_A, Λ_B

In order to relate this to a defect we recall the definition of generalized singular values of two square matrices A and B.

DEFINITION 3.2. σ_{ri} σ_{li} $\lambda B - A$ $\lambda B^T B - A^T A$ $\lambda B B^T - A A^T$

Since the invertibility of B is not essential in these definitions, we first make the simplifying assumption that B is invertible. It then follows easily that

$$\sigma_{ri} = \sigma_i(AB^{-1}), \quad \sigma_{li} = \sigma_i(B^{-1}A).$$

When B is invertible, it is shown in [1] that the pencil $\lambda B - A$ is normal iff both AB^{-1} and $B^{-1}A$ are normal. A good candidate for the $\lambda B - A$ appears then to be

$$\Gamma(A, B) := \sum_{i=1}^{n} \sigma_{ri}^2 + \sum_{i=1}^{n} \sigma_{li}^2 - 2\sum_{i=1}^{n} |\lambda_i|^2,$$

where λ_i are the generalized eigenvalues of the pencil. Clearly $\Gamma(A, B) = \gamma(AB^{-1}) + \gamma(B^{-1}A)$, which is always positive and is zero iff both AB^{-1} and $B^{-1}A$ are normal and hence iff the pencil $\lambda B - A$ is normal.

If B is not invertible, we need another "defect from normality" function since $\Gamma(A, B)$ is then the difference between two infinite quantities. We can then consider a transformed pencil

$$\lambda\widehat{B} - \widehat{A} := \lambda(cB - sA) - (sB + cA), \quad c^2 + s^2 = 1. \tag{3.1}$$

It is well known (see, e.g., [1]) that for a regular pencil $\lambda B - A$ there always exists a choice (c, s) for which $\widehat{B}$ is invertible. Since the above transformation does affect the left and right eigenvectors of a pencil, it follows that

$$\lambda\widehat{B} - \widehat{A} \text{ is normal} \iff \lambda B - A \text{ is normal.}$$

Rather than minimizing $\Gamma(A, B)$ one can thus minimize $\Gamma(\widehat{A}, \widehat{B})$ which will reach a minimum when both $\lambda\widehat{B} - \widehat{A}$ and $\lambda B - A$ are normal pencils. Notice however that the value of this defect then changes although normality is preserved. Without loss of generality, we assume from now on that B is invertible.

But orthogonality of the left and right eigenvectors is not sufficient to guarantee a low sensitivity of the generalized eigenvalues of a regular pencil because eigenvalues can now be arbitrarily large or small, irrespective of the norm of A and B. Let x_i and y_i be, respectively, the right and left eigenvectors of a given eigenvalue λ_i,

$$Ax_i = \lambda_i Bx_i, \quad y_i^* A = \lambda_i y_i^* B,$$

and define the corresponding Rayleigh components:

$$\alpha_i := y_i^* Ax_i/(\|y_i\|_2\|x_i\|_2), \quad \beta_i := y_i^* Bx_i/(\|y_i\|_2\|x_i\|_2), \quad \lambda_i = \alpha_i/\beta_i.$$

In [6] it is shown that a perturbation in A and B of relative size ϵ,

$$\|\delta A\|_2 \le \epsilon\|A\|_2, \quad \|\delta B\|_2 \le \epsilon\|B\|_2, \tag{3.2}$$

yields a perturbed eigenvalue $\tilde{\lambda}_i$ such that the chordal distance

$$\chi(\lambda_i, \tilde{\lambda}_i) \;:=\; \frac{|\alpha_i\tilde{\beta}_i - \tilde{\alpha}_i\beta_i|}{\sqrt{|\alpha_i|^2 + |\beta_i|^2}\sqrt{|\tilde{\alpha}_i|^2 + |\tilde{\beta}_i|^2}} \tag{3.3}$$

between the original and the perturbed eigenvalue is bounded by

$$\chi(\lambda_i, \tilde{\lambda_i}) \;\le\; \epsilon\frac{(\|A\|_2^2 + \|B\|_2^2)^{1/2}}{(|\alpha_i|^2 + |\beta_i|^2)^{1/2}} \;+\; O(\epsilon^2)$$

and that there exist perturbations δA and δB, for which this bound is met. The quantity

$$\kappa(\lambda_i) := \frac{(\|A\|_2^2 + \|B\|_2^2)^{1/2}}{(|\alpha_i|^2 + |\beta_i|^2)^{1/2}} \tag{3.4}$$

is thus a valid relative condition number for λ_i in the sense that it measures how a perturbation of relative size ϵ in A and B affects λ_i in the (intrinsically relative) chordal metric. The reason why such a "relative" metric is to be preferred for pencils is linked to the fact that eigenvalues are now given by ratios of computed quantities. (See [6] for more details.)

When using the QZ-algorithm to compute the generalized eigenvalues of the pencil $\lambda B - A$ one obtains the so-called Schur form of this pencil:

$$A_S := Q^* AZ,\ B_S := Q^* BZ, \quad Q^*Q = I_n,\ Z^*Z = I_n, \tag{3.5}$$

where A_S and B_S are both upper triangular. This algorithm typically induces errors δA and δB in A and B that are of the order of (3.2), where ϵ is the machine accuracy of the computer. Since the orthogonal transformations Q and Z do not affect the quantities used in the definition (3.4) of $\kappa(\lambda_i)$, we can as well analyze the effect of perturbations in the coordinate system of the Schur form. The right and left

eigenvectors x_i, y_i can then be normalized as follows:

$$x_i := \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_{i-1} \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad y_i := \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \eta_{i+1} \\ \vdots \\ \eta_n \end{bmatrix}.$$

If we denote the diagonal entries of the triangular matrices A_S, B_S by a_{ii}, b_{ii}, respectively, we then obtain the equalities

$$\alpha_i = a_{ii}/n_i, \quad \beta_i = b_{ii}/n_i, \quad n_i := (\|y_i\|_2 \|x_i\|_2) \geq 1.$$

Since

$$\|A\|_2 = \|A_S\|_2 \geq \max_i |a_{ii}|, \quad \|B\|_2 = \|B_S\|_2 \geq \max_i |b_{ii}|$$

we finally obtain the inequality

$$\kappa(\lambda_i) \geq \frac{(\max_i |a_{ii}|^2 + \max_i |b_{ii}|^2)^{1/2}}{(|a_{ii}|^2 + |b_{ii}|^2)^{1/2}} \tag{3.6}$$

with [illegible] holding only for a normal pencil since then x_i and y_i have norm 1 and A_S and B_S are diagonal. But normal pencils can still have a quantity $\kappa(\lambda_i)$ that can be very large if the pairs (a_{ii}, b_{ii}) vary a lot in norm. This is not the case for the following subclass of normal pencils.

DEFINITION 3.3. [illegible] $\lambda B - A$ [illegible] U_l, U_r [illegible] Λ_A, Λ_B [illegible] $\gamma \neq 0$

$$U_l^*(\lambda B - A)U_r = \lambda\Lambda_B - \Lambda_A, \quad |\Lambda_A|^2 + |\Lambda_B|^2 = \gamma^2 I.$$

For this class of pencils we obviously have

$$1 \leq \kappa(\lambda_i) = \frac{(\max_i |a_{ii}|^2 + \max_i |b_{ii}|^2)^{1/2}}{(|a_{ii}|^2 + |b_{ii}|^2)^{1/2}} \leq \sqrt{2} \tag{3.7}$$

with the lower bound $\kappa(\lambda_i) = 1$ met for each i in the particular case where $\Lambda_A = \alpha I$ and $\Lambda_B = \beta I$. Obviously the class of standard normal pencils is nearly optimal in terms of eigenvalue sensitivity since $\kappa(\lambda_i) \leq \sqrt{2}$ for each eigenvalue λ_i.

The following theorem explains which pencils can be transformed to standard normal form using left and right transformations.

THEOREM 3.4. [illegible]

$$T_l^{-1}(\lambda B - A)T_r = \lambda\Lambda_B - \Lambda_A, \quad |\Lambda_A|^2 + |\Lambda_B|^2 = \gamma^2 I \quad \text{with } \gamma \in \mathbb{R}_0.$$

[illegible]. If $\lambda B - A$ has a full set of right and left eigenvectors x_i, y_i, then putting x_i as the columns of T_r and y_i^* as the rows of T_l^{-1} will diagonalize $T_l^{-1}(\lambda B - A)T_r$. A simple additional diagonal scaling—which can be absorbed in either T_r or T_l—will ensure that moreover $|\Lambda_A|^2 + |\Lambda_B|^2 = \gamma^2 I$ for some arbitrary real positive γ. □

[illegible] 3.1. For nondiagonalizable (regular) pencils, the theorem remains valid in the limit, but then T_l, T_r are unbounded. In this case we have that $\lambda\Lambda_B - \Lambda_A$ belongs to the closure of the orbit of $\lambda B - A$ under left and right transformations T_l, T_r [5].

4. Balancing pencils. We now look for scaling transformations that make a given pencil get closer to a normal one. We could use a scaling of the type

$$D_l^{-1}(\lambda B - A)D_r, \tag{4.1}$$

where D_r, D_l are [illegible]. This does not modify the generalized eigenvalues of the pencil, but the defect from normality $\Gamma(A, B)$ becomes now

$$\Gamma(D_l^{-1}AD_r, D_l^{-1}BD_r) = \sum_{i=1}^{n} \sigma_i^2(D_l^{-1}AB^{-1}D_l) + \sum_{i=1}^{n} \sigma_i^2(D_r^{-1}B^{-1}AD_r) - 2\sum_{i=1}^{n} |\lambda_i|^2.$$

It follows from section 2 that the optimal D_r, D_l are solutions of

$$\inf_{D_r} \|D_r^{-1}B^{-1}AD_r\|_F, \quad \inf_{D_l} \|D_l^{-1}AB^{-1}D_l\|_F.$$

But such an approach would require us to invert the matrix B (at least implicitly) and it is unclear how to proceed when B is singular.

We now define a new optimization problem inspired by Theorem 3.4 that avoids the inversion of B. It uses the so-called Frobenius inner product for regular pencils defined in a $2n^2$-dimensional space of two $n \times n$ complex matrices:

$$\langle \lambda B_1 - A_1 \ , \ \lambda B_2 - A_2 \rangle_F \ := \ \mathrm{tr}(A_1 A_2^* + B_1 B_2^*).$$

It follows then that $\|\lambda B - A\|_F^2 := \langle \lambda B - A, \lambda B - A \rangle_F = \|A\|_F^2 + \|B\|_F^2$, where $\|.\|_F$ denotes the usual Frobenius matrix norm.

THEOREM 4.1. [illegible]

$$\inf_{\det(T_l^{-1}T_r)=1} \|T_l^{-1}(\lambda B - A)T_r\|_F \tag{4.2}$$

[illegible] $\lambda B - A$ [illegible] T_r, T_l [illegible]. Using the Schur decomposition $\lambda B_S - A_S = Q^*(\lambda B - A)Z$ we define triangular matrices $R_r := Z^*T_rQ_r$ and $R_l := Q^*T_lQ_l$, where Q_r are Q_l are chosen to be unitary and $\det Q_l^*Q_r = 1$. Since unitary transformations do not change the Frobenius norm, the above minimization is then equivalent to

$$\inf_{\det(R_l^{-1}R_r)=1} \|R_l^{-1}(\lambda B_S - A_S)R_r\|_F,$$

where now all matrices are upper triangular. Moreover, if we factor $R_r = D_rU_r$ and $R_l = D_lU_l$, where U_r and U_l are unit upper triangular and D_r and D_l are diagonal, then the problem splits in two subproblems. Clearly U_r and U_l affect only the elements above the diagonal of $\|R_l^{-1}(\lambda B_S - A_S)R_r\|_F$ and these can all be put equal to zero if the pencil is diagonalizable (e.g., when there are no repeated eigenvalues). In such a case the problem reduces further to

$$\inf_{\det(D_l^{-1}D_r)=1} \|D_l^{-1}(\lambda \Lambda_B - \Lambda_A)D_r\|_F,$$

which is easily solved using a Lagrange multiplier approach. The solution

$$D_l^{-2}D_r^2(\Lambda_B^*\Lambda_B + \Lambda_A^*\Lambda_A) = \gamma^2 I, \quad \gamma^{2n} = \det(\Lambda_B^*\Lambda_B + \Lambda_A^*\Lambda_A)$$

is equivalent to the condition that $D_l^{-1}(\lambda\Lambda_B - \Lambda_A)D_r$ is a standard normal pencil. If the pencil is not diagonalizable, it is still possible to find unbounded diagonal scalings $D_r,\ D_l$ that will make the elements that are above the diagonal in the Schur form tend to zero. □

The above theorem suggests to use the same minimization problem but now restricted to [illegible]

$$\inf_{\det(D_l^{-1}D_r)=1} \|D_l^{-1}(\lambda B - A)D_r\|_F \tag{4.3}$$

as a technique to balance regular pencils. We will show that this has a unique minimum that is attained when

$$\|(D_l^{-1}AD_r)e_j\|_2^2 + \|(D_l^{-1}BD_r)e_j\|_2^2 = \|e_i^T(D_l^{-1}AD_r)\|_2^2 + \|e_i^T(D_l^{-1}BD_r)\|_2^2 = \gamma^2$$

for all i and j. This leads to the following generalization of (2.5).

DEFINITION 4.2. [illegible] $n \times n$ [illegible] $\lambda B - A$ [illegible]

$$\|Ae_j\|_2^2 + \|Be_j\|_2^2 = \|e_i^T A\|_2^2 + \|e_i^T B\|_2^2 = \gamma^2 \quad \forall\ i, j. \tag{4.4}$$

The following theorem proves that every balanced pencil can be seen as the solution of an optimization problem very similar to (4.3).

THEOREM 4.3. [illegible] $D_l^{-1}(\lambda B - A)D_r$ [illegible] D_l, D_r [illegible]

$$\inf_{\det(D_l^{-1}D_r)=c} \|D_l^{-1}(\lambda B - A)D_r\|_F.$$

[illegible]. Denote the ith diagonal entry of D_r and D_l by d_{ri} and d_{li}, respectively, and let $a_{ij},\ b_{ij}$ be the entries of the matrices $A,\ B$. We want to minimize

$$\inf_{d_{li},\ d_{rj}} \sum_{i,j=1}^{n} (|a_{ij}|^2 + |b_{ij}|^2)\left(\frac{d_{rj}}{d_{li}}\right)^2, \quad \text{where} \quad \left(\frac{\prod d_{lk}}{\prod d_{rk}}\right)^2 = c^2.$$

With the change variables $d_{ri}^2 = \exp(u_{ri})$ and $d_{li}^2 = \exp(-u_{li})$ and when putting $m_{ij} := |a_{ij}|^2 + |b_{ij}|^2$, this becomes

$$\inf_{u_{li},\ u_{rj}} \sum_{i,j=1}^{n} m_{ij}\exp(u_{li} + u_{rj}), \quad \text{where} \quad \sum_k (u_{lk} + u_{rk}) = 2\ln c.$$

This is a [illegible] minimization problem with a linear constraint. Its solution can be found via the use of a Lagrange multiplier Γ:

$$\inf_{u_{li},\ u_{rj}} \sum_{i,j=1}^{n} m_{ij}\exp(u_{li} + u_{rj}) + \Gamma\left(2\ln c - \sum_k (u_{lk} + u_{rk})\right).$$

This unconstrained minimization has therefore a minimum iff the first order conditions are satisfied. These are $\sum_k (u_{lk} + u_{rk}) = 2\ln c$ and

$$\sum_{i=1}^{n} m_{ij}\exp(u_{li} + u_{rj}) = \sum_{j=1}^{n} m_{ij}\exp(u_{li} + u_{rj}) = \Gamma \quad \forall\ i, j.$$

Putting $\Gamma = \gamma^2$ and rephrasing it in the original variables, this amounts to

$$\|e_i^T(D_l^{-1}AD_r)\|_2^2 + \|e_i^T(D_l^{-1}BD_r)\|_2^2 \;=\; \|(D_l^{-1}AD_r)e_j\|_2^2 + \|(D_l^{-1}BD_r)e_j\|_2^2 = \gamma^2$$

for all i, j. The optimal pencil $D_l^{-1}(\lambda B - A)D_r$ is therefore balanced. The converse statement is easily checked in a similar manner. □

REMARK 4.1. Notice that if the pencil $\lambda B - A$ can be permuted to a block triangular pencil, then so can the matrix M with elements m_{ij}. One then easily checks that the scalings of the above theorem can be unbounded for this so-called reducible case. This case is typically excluded in the scaling problem, since then the generalized eigenvalue problem can de deflated to smaller dimensional ones [7]. When such permutations do not exist, the scaling problem has a bounded solution.

REMARK 4.2. The above theorem does not prove that the diagonal scaling procedure will always improve the sensitivity of the eigenvalue problem but the bound (3.4) for $\kappa(\lambda_i)$ suggests that this will be the case. We will illustrate by numerical experiments that the scaling typically improves the sensitivity of the eigenvalues.

REMARK 4.3. The above theorem also allows us to choose the parameter γ in (4.4) since modifying the constant c in the condition $\det(D_l^{-1}D_r) = c$ automatically scales all the column and row norms. This is used in the numerical method described below.

5. Numerical method. In order to balance a pencil, we will use a very simple method rather than using convex optimization techniques. This method consists in alternatively updating D_r and D_l such that the compound matrices $\left[\begin{smallmatrix}A\\B\end{smallmatrix}\right]D_r$ and $D_l^{-1}[A \quad B]$ have column norms and row norms equal to 1, respectively. By doing so we converge linearly to a balanced pencil with $\gamma = 1$ in (4.4). The proposed method is essentially a "coordinate descent" method where one alternates between computing the optimum in the "coordinates" of D_r and D_l. The convergence is slow but when we restrict ourselves to powers of the base (2 or 10) for the diagonal elements of D_r and D_l, stagnation typically occurs after two or three updates of both D_r and D_l. Each joint update of D_l and Dr in fact requires only $4n^2$ floating point operations if one uses the matrix M with elements $m_{ij} := |a_{ij}|^2 + |b_{ij}|^2 : 2n^2$ to compute the row and column norms and $2n^2$ to perform the two scalings. (A MATLAB code is given in the appendix for the base 2.) The scaling procedure has therefore a marginal cost in comparison to the eigenvalue computation. As in the standard eigenvalue problem one has to test also if there exist permutations that reduce the pencil to a block triangular form so that lower dimensional eigenvalue problems can be isolated. Such a procedure is needed to guarantee that the diagonal scaling will remain bounded but the complexity is also quadratic in n (see [7]).

6. Numerical examples. In the following tables we compare the precision of the computed eigenvalues without scaling, after applying our proposed scaling procedure and after applying Ward's method [7], which is currently implemented in LAPACK. We consider in Table 6.1 randomly generated diagonalizable pencils $T_l^{-1}(\lambda\Lambda_B - \Lambda_A)T_r$ (where $\lambda\Lambda_B - \Lambda_A$ is in standard normal form), in Table 6.2 randomly generated nondiagonalizable pencils $T_l^{-1}(\lambda J_B - J_A)T_r$ (where $J_B^{-1}J_A$ is in Jordan normal form), and in Table 6.3 pencils with elements of strongly varying order of magnitude. We used normally distributed random numbers for the free elements of $\Lambda_A, \Lambda_B, J_A$, and J_B. We imposed the normalization in $\lambda\Lambda_B - \Lambda_A$ by choosing Λ_B to satisfy $\Lambda_B^2 + \Lambda_A^2 = \gamma^2 I$ and the Jordan structure in $\lambda J_B - J_A$ by choosing some repeated consecutive elements on the diagonals of J_A and J_B and assigning corresponding off-diagonal 1's in J_A. The condition number of the random matrices T_l and T_r was

TABLE 6.1
Comparison for randomly generated diagonalizable pencils.

$n = 10$	c_{orig}	c_{bal}	c_{ward}	c_{ward}/c_{bal}
$\kappa(T_r) = 3.27e+07, \kappa(T_l) = 2.58e+11$	3.01e-03	7.00e-13	2.61e-09	3.72+03
$\kappa(T_r) = 8.24e+12, \kappa(T_l) = 4.21e+10$	3.69e-01	3.20e-12	1.00e-09	3.12e+02
$\kappa(T_r) = 6.81e+08, \kappa(T_l) = 1.75e+07$	7.81e-09	8.84e-14	1.01e-11	1.15e+02
$\kappa(T_r) = 1.06e+07, \kappa(T_l) = 7.82e+08$	1.56e-07	4.90e-13	4.16e-13	8.50e-01
$\kappa(T_r) = 1.46e+05, \kappa(T_l) = 4.08e+05$	2.67e-10	3.52e-15	3.92e-15	1.12e+00
$\kappa(T_r) = 1.92e+03, \kappa(T_l) = 7.72e+02$	6.78e-13	3.04e-15	2.07e-14	6.08e+00
$\kappa(T_r) = 3.95e+01, \kappa(T_l) = 1.75e+01$	2.23e-15	2.20e-15	6.52e-15	2.97e+00
$\kappa(T_r) = 1.00e+00, \kappa(T_l) = 1.00e+00$	4.79e-16	4.79e-16	4.94e-14	1.03e+02

TABLE 6.2
Randomly generated nondiagonalizable pencils.

$n = 10$	c_{orig}	c_{bal}	c_{ward}	c_{ward}/c_{bal}
$\kappa(T_r) = 1.15e+09, \kappa(T_l) = 3.27e+09$	4.88e-01	4.88e-01	4.88e-01	1.00e+00
$\kappa(T_r) = 4.68e+02, \kappa(T_l) = 4.79e+03$	1.30e-01	1.30e-01	1.30e-01	1.00e+00

TABLE 6.3
Pencils with elements of strongly varying order of magnitude.

c_{orig}	c_{bal}	c_{ward}	c_{ward}/c_{bal}
4.38e-10	4.30e-15	1.02e-05	2.37e+09
1.25e-13	1.90e-15	1.92e-03	1.01e+12
9.16e-12	6.13e-16	1.17e-10	1.92e+05

controlled by taking the kth power of normally distributed random numbers $r_{i,j}$ as their elements. A larger power k then typically yields a larger condition number for the transformation. For these experiments we used the QZ-algorithm [3] applied to different pencils of size $n = 10$. We computed the chordal distances $c_i := \chi(\lambda_i, \tilde{\lambda}_i)$ for all eigenvalues λ_i and compared in each table the quantities $c := \|[c_1, \ldots, c_n]\|_2$ for the original pencil (c_{orig}), for the balanced pencil constructed by our algorithm (c_{bal}), and for the balanced pencil using Ward's method (c_{ward}). In Tables 6.1 and 6.2 we also give the condition numbers $\kappa(T_r)$ and $\kappa(T_l)$.

In Table 6.1 we focus on diagonalizable pencils. When $\kappa(T_r) = \kappa(T_l) = 1$ we observe that balancing does not improve the precision of the calculated eigenvalues, but otherwise it does, in general, significantly improve the accuracy of the calculated eigenvalues. Recall also that we restrict the diagonal elements of the balancing transformations D_r, D_l to be powers of two. From the table it appears that the proposed balancing algorithm has a positive effect on the precision of the computed eigenvalues. The comparison factor c_{ward}/c_{bal} shows that in general the new method outperforms Ward's algorithm.

In Table 6.2 we look at nondiagonalizable pencils. We imposed the first example to have two Jordan blocks of size 2 and the second example to have one Jordan block of size 3. The eigenvalue sensitivity is in principle infinite and the calculated eigenvalues have little in common with the true eigenvalues. The table shows that both balancing algorithms do not alter the precision of the computed eigenvalues. In Table 6.3 we look at pencils with entries of strongly varying size: the largest ones are of the order of 1, the smallest ones are much smaller. Ward's method tries to

make the size of these elements equal and in doing so, it applies a scaling that often deteriorates the sensitivity instead of improving it. The new method, on the other hand, usually significantly improves the sensitivity.

7. Conclusion. In this paper we presented a new balancing method for matrix pencils. From the point of view of the sensitivity of the eigenvalues we showed that the standard normal pencils are near optimal and that they can be viewed as a natural extension of normal matrices. A diagonal balancing method was then proposed that makes a given pencil as close as possible to a standard normal one. Moreover we showed that the complexity of the new method is comparable to that of the classical balancing of matrices. We also gave numerical evidence that the accuracy of computed generalized eigenvalues may significantly improve after balancing a pencil and that the method often outperforms the method of Ward implemented in LAPACK.

Appendix.

```
function [Dl, Dr, iter] = baleig(A,B,max_iter)

% Performs two-sided scaling Dl\A*Dr, Dl\B*Dr in order to improve
% the sensitivity of generalized eigenvalues. The diagonal matrices
% Dl and Dr are constrained to powers of 2 and are computed iteratively
% until the number of iterations max_iter is met or until the norms are
% between 1/2 and 2. Convergence is often reached after 2 or 3 steps.
% The diagonals of the scaling matrices are returned in Dl and Dr
% and so is iter, the number of iterations steps used by the method.

n=size(A,1); Dl=ones(1,n); Dr=ones(1,n); M=abs(A).^2+abs(B).^2;

for iter=1:max_iter,
    emax=0;emin=0;
    for i=1:n;
        % scale the rows of M to have approximate row sum 1
        d=sum(M(i,:));e=-round(log2(abs(d))/2);
        M(i,:)=pow2(M(i,:),2*e);
        % apply the square root scaling also to Dl
        Dl(i)=pow2(Dl(i),-e);
        if e > emax, emax=e; end; if e < emin, emin=e; end
    end
    for i=1:n;
        % scale the columns of M to have approximate column sum 1
        d=sum(M(:,i));e=-round(log2(abs(d))/2);
        M(:,i)=pow2(M(:,i),2*e);
        % apply the square root scaling also to Dr
        Dr(i)=pow2(Dr(i),e);
        if e > emax, emax=e; end; if e < emin, emin=e; end
    end
    % Stop if norms are all between 1/2 and 2
    if (emax<=emin+2), break; end
end
```

Acknowledgment. We would like to acknowledge the help in Theorem 4.3 of Yurii Nesterov, who pointed out that this was a convex optimization problem. We also thank Daniel Kressner, who sent us a 3-by-3 example from his thesis for which Ward's scaling significantly deteriorates the sensitivity of the computed eigenvalues. The examples in Table 6.3 are inspired by this.

REFERENCES

[1] J.-P. Charlier and P. Van Dooren, *A Jacobi-like algorithm for computing the generalized Schur form of a regular pencil*, J. Comput. Appl. Math., 27 (1989), pp. 17–36.

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

[3] C. B. Moler and G. W. Stewart, *An algorithm for generalized matrix eigenvalue problems*, SIAM J. Numer. Anal., 10 (1973), pp. 241–256.

[4] B. N. Parlett and C. Reinsch, *Balancing a matrix for calculation of eigenvalues and eigenvectors*, Numer. Math., 13 (1969), pp. 293–304.

[5] A. Pokrzywa, *On perturbations and the equivalence orbit of a matrix pencil*, Linear Algebra Appl., 82 (1986), pp. 99–121.

[6] G. W. Stewart, *Perturbation theory for the generalized eigenvalue problem*, in Recent Advances in Numerical Analysis, C. de Boor and G. Golub, eds., Academic Press, New York, 1978, pp. 193–206.

[7] R. C. Ward, *Balancing the generalized eigenvalue problem*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 141–152.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 1, pp. 264–284

MODIFIED GRAM–SCHMIDT (MGS), LEAST SQUARES, AND BACKWARD STABILITY OF MGS-GMRES*

CHRISTOPHER C. PAIGE†, MIROSLAV ROZLOŽNÍK‡, AND ZDENĚK STRAKOŠ‡

Abstract. The generalized minimum residual method (GMRES) [Y. Saad and M. Schultz, *SIAM J. Sci. Statist. Comput.*, 7 (1986), pp. 856–869] for solving linear systems $Ax = b$ is implemented as a sequence of least squares problems involving Krylov subspaces of increasing dimensions. The most usual implementation is modified Gram–Schmidt GMRES (MGS-GMRES). Here we show that MGS-GMRES is backward stable. The result depends on a more general result on the backward stability of a variant of the MGS algorithm applied to solving a linear least squares problem, and uses other new results on MGS and its loss of orthogonality, together with an important but neglected condition number, and a relation between residual norms and certain singular values.

Key words. rounding error analysis, backward stability, linear equations, condition numbers, large sparse matrices, iterative solution, Krylov subspace methods, Arnoldi method, generalized minimum residual method, modified Gram–Schmidt, QR factorization, loss of orthogonality, least squares, singular values

AMS subject classifications. 65F10, 65F20, 65F25, 65F35, 65F50, 65G50, 15A12, 15A42

DOI. 10.1137/050630416

1. Introduction.

Consider a system of linear algebraic equations $Ax = b$, where A is a given $n \times n$ (unsymmetric) nonsingular matrix and b a nonzero n-dimensional vector. Given an initial approximation x_0, one approach to finding x is to first compute the initial residual $r_0 = b - Ax_0$. Using this, derive a sequence of Krylov subspaces $\mathcal{K}_k(A, r_0) \equiv \text{span}\{r_0, Ar_0, \ldots, A^{k-1}r_0\}$, $k = 1, 2, \ldots$, in some way, and look for approximate solutions $x_k \in x_0 + \mathcal{K}_k(A, r_0)$. Various principles are used for constructing x_k, which determine various Krylov subspace methods for solving $Ax = b$. Similarly, Krylov subspaces for A can be used to obtain eigenvalue approximations or to solve other problems involving A.

Krylov subspace methods are useful for solving problems involving very large sparse matrices, since these methods use these matrices only for multiplying vectors, and the resulting Krylov subspaces frequently exhibit good approximation properties. The Arnoldi method [2] is a Krylov subspace method designed for solving the eigenproblem of unsymmetric matrices. The generalized minimum residual method (GMRES) [20] uses the Arnoldi iteration and adapts it for solving the linear system $Ax = b$. GMRES can be computationally more expensive per step than some other methods; see, for example, Bi-CGSTAB [24] and QMR [9] for unsymmetric A, and LSQR [16] for unsymmetric or rectangular A. However, GMRES is widely used for solving linear systems arising from discretization of partial differential equations, and

*Received by the editors May 2, 2005; accepted for publication (in revised form) by M. Benzi October 28, 2005; published electronically March 17, 2006.
http://www.siam.org/journals/simax/28-1/63041.html

†School of Computer Science, McGill University, Montreal, Quebec, Canada, H3A 2A7 (paige@cs.mcgill.ca). This author's work was supported by NSERC of Canada grant OGP0009236.

‡Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic (miro@cs.cas.cz, strakos@cs.cas.cz). The work of these authors was supported by the project 1ET400300415 within the National Program of Research "Information Society" and by the Institutional Research Plan AV0Z10300504 "Computer Science for the Information Society: Models, Algorithms, Applications."

as we will show, it is backward stable and it does effectively minimize the 2-norm of the residual at each step.

The most usual way of applying the Arnoldi method for large, sparse unsymmetric A is to use modified Gram–Schmidt orthogonalization (MGS). Unfortunately in finite precision computations this leads to loss of orthogonality among the MGS Arnoldi vectors. If these vectors are used in GMRES we have MGS-GMRES. Fortunately, experience suggests that MGS-GMRES succeeds despite this loss of orthogonality; see [12]. For this reason we examine the MGS version of Arnoldi's algorithm and use this to show that the MGS-GMRES method does eventually produce a backward stable approximate solution when applied to any member of the following class of linear systems with floating point arithmetic unit roundoff ϵ (σ means singular value):

$$Ax = b \neq 0, \quad A \in \mathbf{R}^{n\times n}, \quad b \in \mathbf{R}^n, \quad \sigma_{min}(A) \gg n^2\epsilon\|A\|_F; \tag{1.1}$$

see also the appendix. The restriction here is deliberately imprecise; see below. Moreover we show that MGS-GMRES gives backward stable solutions for its least squares problems at all iteration steps, thus answering important open questions. The proofs depend on new results on the loss of orthogonality and backward stability of the MGS algorithm, as well as the application of the MGS algorithm to least squares problems, and a lot of this paper is devoted to first obtaining these results.

While the kth step of MGS produces the kth orthonormal vector v_k, it is usual to say v_k is produced by step $k-1$ in the Arnoldi and MGS-GMRES algorithms. We will attempt to give a consistent development while avoiding this confusion. Thus step $k-1$ of MGS-GMRES is essentially the kth step of MGS applied to $[b, AV_{k-1}]$ to produce v_k in $[b, AV_{k-1}] = V_kR_k$, where $V_k \equiv [v_1, \ldots, v_k]$ and R_k is upper triangular. In practice, if we reach a solution at step $m-1$ of MGS-GMRES, then numerically b must lie in the range of AV_{m-1}, so that $B_m \equiv [b, AV_{m-1}]$ is numerically rank deficient. But this means we have to show that our rounding error analysis of MGS holds for rank deficient B_m—and this requires an extension of some results in [5].

In section 2 we describe our notation and present some of the tools we need which may be of more general use. For example we show the importance of the condition number $\tilde{\kappa}_F(A)$ in (2.1), prove the existence of a nearby vector in Lemma 2.3, and provide a variant of the singular value–residual norm relations of [17] in Theorem 2.4. In sections 3.1–3.2 we review MGS applied to $n\times m$ B of rank m, and its numerical equivalence to the Householder QR reduction of B augmented by an $m \times m$ matrix of zeros. In section 3.3 we show how the MGS rounding error results extend to the case of $m > n$, while in section 4 we show how these results apply to the Arnoldi algorithm. In section 5 we analyze the loss of orthogonality in MGS and the Arnoldi algorithm and how it is related to the near rank deficiency of the columns of B or its Arnoldi equivalent, refining a nice result of Giraud and Langou [10] and Langou [14]. Section 6 introduces the key step used to prove convergence of these iterations. In section 7.1 we prove the backward stability of the MGS algorithm applied to solving linear least squares problems of the form required by the MGS-GMRES algorithm, and in section 7.2 we show how loss of orthogonality is directly related to new normwise relative backward errors of a sequence of [illegible] least squares problems, supporting a conjecture on the convergence of MGS-GMRES and its loss of orthogonality; see [18]. In section 8.1 we show that at every step MGS-GMRES computes a backward stable solution for that step's linear least squares problem, and in section 8.2 we show that one of these solutions is also a backward stable solution for (1.1) in at most $n+1$ MGS steps.

The restriction on A in (1.1) is essentially a warning to be prepared for difficulties in using the basic MGS-GMRES method on singular systems; see, for example, [6, 23]. The imprecise nature of the condition (using $\gg$ instead of $>$ with some constant) was chosen to make the presentation easier. A constant could be provided (perhaps closer to 100 than 10), but since the long bounding sequence used was so loose, it would be meaningless. The appendix suggests that the form $n^2\epsilon\|A\|_F$ might be optimal, but since for large n rounding errors tend to combine in a semirandom fashion, it is reasonable to replace n^2 by n, and a more practical requirement than (1.1) might be

$$\text{For large } n, \qquad n\epsilon\|A\|_F/\sigma_{min}(A) \le 0.1. \tag{1.2}$$

2. Notation and mathematical basics. We describe the notation we will use, together with some generally useful results. We use "$\equiv$" to mean "is defined as" in the first occurrence of an expression, but in any later occurrences of this expression it means "is equivalent to (by earlier definition)." A bar above a symbol will denote a computed quantity, so if V_k is an ideal mathematical quantity, $\bar{V}_k$ will denote its actual computed value. The floating point arithmetic unit roundoff will be denoted by ϵ (half the machine epsilon; see [13, pp. 37–38]), I_n denotes the $n\times n$ unit matrix, e_j will be the jth column of a unit matrix I, so Be_j is the jth column of B, and $B_{i:j} \equiv [Be_i, \ldots, Be_j]$. We will denote the absolute value of a matrix B by $|B|$, its Moore–Penrose generalized inverse by $B^\dagger$, $\|\cdot\|_F$ will denote the Frobenius norm, $\sigma(\cdot)$ will denote a singular value, and $\kappa_2(B) \equiv \sigma_{max}(B)/\sigma_{min}(B)$; see (2.1) for $\tilde{\kappa}_F(\cdot)$. Matrices and vectors whose first symbol is Δ, such as ΔV_k, will denote rounding error terms. For the rounding error analyses we will use Higham's notation [13, pp. 63–68]: $\tilde{c}$ will denote a small integer ≥ 1 whose exact value is unimportant ($\tilde{c}$ might have a different value at each appearance) and $\gamma_n \equiv n\epsilon/(1-n\epsilon)$, $\tilde{\gamma}_n \equiv \tilde{c}n\epsilon/(1-\tilde{c}n\epsilon)$. Without mentioning it again, we will always assume the conditions are such that the denominators in objects like this (usually bounds) are positive; see, for example, [13, (19.6)]. We see $\tilde{\gamma}_n/(1-\tilde{\gamma}_n) = \tilde{c}n\epsilon/(1-2\tilde{c}n\epsilon)$, and might write $\tilde{\gamma}_n/(1-\tilde{\gamma}_n) = \tilde{\gamma}'_n$ for mathematical correctness, but will refer to the right-hand side as $\tilde{\gamma}_n$ thereafter. E_m, $\hat{E}_m$, $\tilde{E}_m$ will denote matrices of rounding errors (see just before Theorem 3.3), and $\|E_m e_j\|_2 \le \gamma\|Be_j\|_2$ implies this holds for $j = 1, \ldots, m$ unless otherwise stated.

[illegible] 2.1 (see also the appendix). An important idea used throughout this paper is that column bounds of the above form lead to several results which are independent of column scaling, and we take advantage of this by using the following condition number. Throughout the paper, D will represent [illegible] positive definite diagonal matrix.

The choice of norms is key to making error analyses readable, and fortunately there is a compact column-scaling-independent result with many uses. Define

$$\tilde{\kappa}_F(A) \equiv \min_{\text{diagonal } D>0} \|AD\|_F/\sigma_{min}(AD). \tag{2.1}$$

This condition number leads to some useful new results.

LEMMA 2.1. [illegible] E [illegible] B [illegible] m [illegible] [illegible] D $\|Ee_j\|_2 \le \gamma\|Be_j\|_2,\ j = 1, \ldots, m, \ \Rightarrow\ \|ED\|_F \le \gamma\|BD\|_F$.

$\|Ee_j\|_2 \le \gamma\|Be_j\|_2$ [illegible] $j = 1, \ldots, m$ *and* $\operatorname{rank}(B) = m \Rightarrow \|EB^\dagger\|_F \le \gamma\tilde{\kappa}_F(B)$.

[illegible] $B = Q_1 R$ [illegible] $\|ER^{-1}\|_F \le \gamma\tilde{\kappa}_F(B) = \gamma\tilde{\kappa}_F(R)$

[illegible] $\|Ee_j\|_2 \le \gamma\|Be_j\|_2$ implies $\|EDe_j\|_2 \le \gamma\|BDe_j\|_2$ so $\|ED\|_F \le \gamma\|BD\|_F$. For B of rank m, $(BD)^\dagger = D^{-1}B^\dagger$, $\|(BD)^\dagger\|_2 = \sigma_{min}^{-1}(BD)$, and so

$$\|EB^\dagger\|_F = \|ED(BD)^\dagger\|_F \le \|ED\|_F\|(BD)^\dagger\|_2 \le \gamma\|BD\|_F/\sigma_{min}(BD).$$

Since this is true for all such D, we can take the minimum, proving our results. ∎

LEMMA 2.2. $m \times m$ $\bar{R}$ $P_1^T P_1 = I$ $P_1 \bar{R} = B + E$ $\gamma \tilde{\kappa}_F(B) < 1$

$$\|Ee_j\|_2 \le \gamma \|Be_j\|_2, \quad j = 1, \ldots, m, \Rightarrow \|E\bar{R}^{-1}\|_F \le \gamma \tilde{\kappa}_F(B)/(1 - \gamma \tilde{\kappa}_F(B)).$$

. For any D in (2.1), $\|Ee_j\|_2 \le \gamma \|Be_j\|_2 \Rightarrow \|ED\|_F \le \gamma \|BD\|_F$, and then $\sigma_{min}(\bar{R}D) \ge \sigma_{min}(BD) - \gamma \|BD\|_F$, so $\|E\bar{R}^{-1}\|_F = \|ED(\bar{R}D)^{-1}\|_F$ is bounded by

$$\|ED\|_F \|(\bar{R}D)^{-1}\|_2 \le \frac{\gamma \|BD\|_F}{\sigma_{min}(BD) - \gamma \|BD\|_F} = \frac{\gamma \|BD\|_F / \sigma_{min}(BD)}{1 - \gamma \|BD\|_F / \sigma_{min}(BD)}.$$

Taking the minimum over D proves the result. ∎

Suppose $\bar{V}_m \equiv [\bar{v}_1, \ldots, \bar{v}_m]$ is an $n \times m$ matrix whose columns have been to have 2-norms of 1, and so have norms in $[1-\tilde{\gamma}_n, 1+\tilde{\gamma}_n]$. Now define $\tilde{V}_m \equiv [\tilde{v}_1, \ldots, \tilde{v}_m]$ where $\tilde{v}_j$ is just the correctly normalized version of $\bar{v}_j$, so

$$\begin{aligned} &\bar{V}_m = \tilde{V}_m(I + \Delta_m), \quad \Delta_m \equiv \mathrm{diag}(\nu_j), \quad \text{where } |\nu_j| \le \tilde{\gamma}_n, \quad j = 1, \ldots, m; \\ &\bar{V}_m^T \bar{V}_m = \tilde{V}_m^T \tilde{V}_m + \tilde{V}_m^T \tilde{V}_m . \Delta_m + \Delta_m . \tilde{V}_m^T \tilde{V}_m + \Delta_m . \tilde{V}_m^T \tilde{V}_m . \Delta_m, \\ &\|\bar{V}_m^T \bar{V}_m - \tilde{V}_m^T \tilde{V}_m\|_F / \|\tilde{V}_m^T \tilde{V}_m\|_F \le \tilde{\gamma}_n(2 + \tilde{\gamma}_n) \equiv \tilde{\gamma}_n'. \end{aligned} \tag{2.2}$$

From now on we will not document the analogues of the last step $\tilde{\gamma}_n(2+\tilde{\gamma}_n) \equiv \tilde{\gamma}_n'$ but finish with $\le \tilde{\gamma}_n$. In general it will be as effective to consider $\tilde{V}_m$ as $\bar{V}_m$, and we will develop our results in terms of $\tilde{V}_m$ rather than $\bar{V}_m$. The following will be useful here:

$$\|[\tilde{V}_m, I_n]\|_2^2 = \|I_n + \tilde{V}_m \tilde{V}_m^H\|_2 = 1 + \|\tilde{V}_m \tilde{V}_m^H\|_2 = 1 + \|\tilde{V}_m\|_2^2 \le 1 + \|\tilde{V}_m\|_F^2 = 1 + m. \tag{2.3}$$

Lemma 2.3 deals with the problem: Suppose we have $d \in \mathbf{R}^n$ and we know for some unknown perturbation $f \in \mathbf{R}^{(m+n)}$ that $\left\| \begin{bmatrix} 0 \\ d \end{bmatrix} + f \right\|_2 = \rho$. Is there a perturbation g d, and having a similar norm to that of f, such that $\|d + g\|_2 = \rho$ also? Here we show such a g exists in the form $g = Nf$, $\|N\|_2 \le \sqrt{2}$.

LEMMA 2.3. $d \in \mathbf{R}^n$ $f \in \mathbf{R}^{(m+n)}$

$$\begin{bmatrix} f_1 \\ d + f_2 \end{bmatrix} \equiv \begin{bmatrix} 0 \\ d \end{bmatrix} + f = p\rho \equiv \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} \rho, \quad \|p\|_2 = 1,$$

$0 \le \sigma \le 1$ $v \in \mathbf{R}^n$ $\|v\|_2 = 1$ $n \times (m+n)$ N

$$N \equiv [v(1+\sigma)^{-1} p_1^T, I_n], \tag{2.4}$$

$$d + Nf = v\rho. \tag{2.5}$$

$$\left\| \begin{bmatrix} 0 \\ d \end{bmatrix} + f \right\|_2 = \|d + Nf\|_2 = \rho, \quad 1 \le \|N\|_2 \le \sqrt{2}. \tag{2.6}$$

. Define $\sigma \equiv \|p_2\|_2$. If $\sigma = 0$ take any $v \in \mathbf{R}^n$ with $\|v\|_2 = 1$. Otherwise define $v \equiv p_2/\sigma$ so $\|v\|_2 = 1$. In either case $p_2 = v\sigma$ and $p_1^T p_1 = 1 - \sigma^2$. Now define N as in (2.4), so

$$\begin{aligned} &d + Nf = d + v(1+\sigma)^{-1} \|p_1\|_2^2 \rho + f_2 = p_2\rho + v(1-\sigma)\rho = v\rho, \\ &NN^T = I + v(1+\sigma)^{-2}(1-\sigma^2)v^T, \\ &1 \le \|N\|_2^2 = \|NN^T\|_2 = 1 + (1-\sigma)/(1+\sigma) \le 2, \end{aligned}$$

proving (2.5) and (2.6). ∎

This is a refinement of a special case of [5, Lem. 3.1]; see also [13, Ex. 19.12]. The fact that the perturbation g in d has the form of N times the perturbation f is important, as we shall see in section 7.1.

Finally we give a general result on the relation between least squares residual norms and singular values. The bounds below were given in [17, Thm. 4.1] but subject to the condition [17, (1.4)] that we cannot be sure will hold here. To prove that our results here hold subject to the different condition (1.1), we need to prove a related result. In order not to be too repetitive, we will prove a slightly more general result than we considered before, or need here, and make the theorem and proof brief.

THEOREM 2.4. $B \in \mathbf{R}^{n\times k}$ $\sigma_1 \geq \cdots \geq \sigma_s > 0$ $0 \neq c \in \mathbf{R}^n$ $\phi \geq 0$ $\hat{y} \equiv B^\dagger c$ $\hat{r} \equiv c - B\hat{y}$ $\sigma(\phi) \equiv \sigma_{s+1}([c\phi, B])$ $\delta(\phi) \equiv \sigma(\phi)/\sigma_s$ $\hat{r}\phi \neq 0$ $\sigma(\phi) > 0$ $\phi_0 \equiv \sigma_s/\|c\|$ $\phi \in [0, \phi_0)$ $0 \leq \delta(\phi) < 1$ $\phi > 0$ $\delta(\phi) < 1$

$$\sigma^2(\phi)[\phi^{-2} + \|\hat{y}\|_2^2] \;\leq\; \|\hat{r}\|_2^2 \;\leq\; \sigma^2(\phi)\left[\phi^{-2} + \frac{\|\hat{y}\|_2^2}{1-\delta^2(\phi)}\right].$$

$\hat{r}$ is the least squares residual for $By \approx c$, so $\hat{r}\phi \neq 0$ means $[c\phi, B]$ has rank $s+1$ and $\sigma(\phi) > 0$. If $0 \leq \phi < \phi_0$, then $\|c\phi\| < \|c\phi_0\| = \sigma_s$, so via Cauchy's interlacing theorem, $0 \leq \sigma(\phi) \equiv \sigma_{s+1}([c\phi, B]) < \sigma_s$, giving $0 \leq \delta(\phi) < 1$. Using the singular value decomposition $B = W\,\mathrm{diag}(\Sigma, 0)Z^T$, $W^T = W^{-1}$, $Z^T = Z^{-1}$, write

$$W^T[c, BZ] = \begin{bmatrix} a_1 & \Sigma & 0 \\ a_2 & 0 & 0 \end{bmatrix}, \quad \Sigma \equiv \begin{bmatrix} \sigma_1 & & \\ & \cdot & \\ & & \sigma_s \end{bmatrix}, \quad a_1 \equiv \begin{bmatrix} \alpha_1 \\ \cdot \\ \alpha_s \end{bmatrix}, \quad \hat{y} = Z\begin{bmatrix} \Sigma^{-1}a_1 \\ 0 \end{bmatrix}.$$

Then it can be shown (see, for example, [26, (39.4)], [17, (2.6)], [15, pp. 1508–10],) that for all ϕ such that $\phi > 0$ and $\delta(\phi) < 1$, $\sigma(\phi)$ is the smallest root of

$$\|\hat{r}\|^2 = \sigma(\phi)^2\left[\phi^{-2} + \sum_{i=1}^{s} \frac{\alpha_i^2/\sigma_i^2}{1-\sigma(\phi)^2/\sigma_i^2}\right].$$

But
$$\|\hat{y}\|_2^2 = \sum_{i=1}^{s} \frac{\alpha_i^2}{\sigma_i^2} \leq \sum_{i=1}^{s} \frac{\alpha_i^2/\sigma_i^2}{1-\sigma(\phi)^2/\sigma_i^2} \leq \sum_{i=1}^{s} \frac{\alpha_i^2/\sigma_i^2}{1-\sigma(\phi)^2/\sigma_s^2} = \frac{\|\hat{y}\|_2^2}{1-\delta^2(\phi)}$$

while $\delta(\phi) \equiv \sigma(\phi)/\sigma_s < 1$, and the result follows. ∎

We introduced ϕ_0 to show $\delta(\phi) < 1$ for some $\phi > 0$. For results related to Theorem 2.4 we refer to [15, pp. 1508–1510], which introduced this useful value ϕ_0.

3. The modified Gram–Schmidt (MGS) algorithm. In order to understand the numerical behavior of the MGS-GMRES algorithm, we first need a very deep understanding of the MGS algorithm. Here this is obtained by a further study of the numerical equivalence between MGS and the Householder QR factorization of an augmented matrix; see [5] and also, for example, [13, section 19.8].

We do not give exact bounds but work with terms of the form $\tilde{\gamma}_n$ instead; see [13, pp. 63–68] and our section 2. The exact bounds will not even be approached for the large n we are interested in, so there is little reason to include such fine detail. In sections 3.1–3.3 we will review the MGS-Householder equivalence and extend some of the analysis that was given in [5] and [13, section 19.8].

3.1. The basic MGS algorithm. Given a matrix $B \in \mathbf{R}^{n\times m}$ with rank $m \le n$, MGS in theory produces V_m and nonsingular R_m in the QR factorization

$$B = V_m R_m, \qquad V_m^T V_m = I_m, \qquad R_m \text{ upper triangular}, \tag{3.1}$$

where $V_m \equiv [v_1, \ldots, v_m]$, and $m \times m$ $R_m \equiv (\rho_{ij})$. The version of the MGS algorithm which immediately updates all columns computes a sequence of matrices $B = B^{(1)}, B^{(2)}, \ldots, B^{(m+1)} = V_m \in \mathbf{R}^{n\times m}$, where $B^{(i)} = [v_1, \ldots, v_{i-1}, b_i^{(i)}, \ldots, b_m^{(i)}]$. Here the first $(i-1)$ columns are final columns in V_m, and $b_i^{(i)}, \ldots, b_m^{(i)}$ have been made orthogonal to $v_1, \ldots, v_{i-1}$. In the ith step we take

$$\rho_{ii} := \|b_i^{(i)}\|_2 \neq 0 \ \text{ since } \operatorname{rank}(B) = m, \quad v_i := b_i^{(i)}/\rho_{ii}, \tag{3.2}$$

and orthogonalize $b_{i+1}^{(i)}, \ldots, b_m^{(i)}$ against v_i using the orthogonal projector $I - v_i v_i^T$,

$$\rho_{ij} := v_i^T b_j^{(i)}, \qquad b_j^{(i+1)} := b_j^{(i)} - v_i \rho_{ij}, \qquad j = i+1, \ldots, m. \tag{3.3}$$

We see $B^{(i)} = B^{(i+1)} R^{(i)}$, where $R^{(i)}$ has the same ith row as R_m but is the unit matrix otherwise. Note that in the mth step no computation is performed in (3.3), so that after m steps we have obtained the factorization

$$B = B^{(1)} = B^{(2)} R^{(1)} = B^{(3)} R^{(2)} R^{(1)} = B^{(m+1)} R^{(m)} \cdots R^{(1)} = V_m R_m, \tag{3.4}$$

where in exact arithmetic the columns of V_m are orthonormal by construction.

This formed R_m a row at a time. If the jth column of B is only available after v_{j-1} is formed, as in MGS-GMRES, then we usually form R_m a column at a time. This does not alter the numerical values if we produce $\rho_{1,j}$, $b_j^{(2)}$; $\rho_{2,j}$, $b_j^{(3)}$; etc.

It was shown in [3] that for the computed $\bar{R}_m$ and $\bar{V}_m$ in MGS

$$B + E = \bar{V}_m \bar{R}_m, \quad \|E\|_2 \le c_1(m,n)\epsilon \|B\|_2, \quad \|I - \bar{V}_m^T \bar{V}_m\|_2 \le c_2(m,n)\epsilon \kappa_2(B), \tag{3.5}$$

where $c_i(m,n)$ denoted a scalar depending on m, n and the details of the arithmetic. We get a deeper understanding by examining the MGS-Householder QR relationship.

3.2. MGS as a householder method. The MGS algorithm for the QR factorization of B can be interpreted as an orthogonal transformation applied to the matrix B augmented with a square matrix of zero elements on top. This is true in theory for [illegible] method of QR factorization, but for Householder's method [illegible]. This observation was made by Charles Sheffield and relayed to the authors of [5] by Gene Golub.

First we look at the theoretical result. Let $B \in \mathbf{R}^{n\times m}$ have rank m, and let $O_m \in \mathbf{R}^{m\times m}$ be a zero matrix. Consider the QR factorization

$$\tilde{B} \equiv \begin{bmatrix} O_m \\ B \end{bmatrix} = P_m \begin{bmatrix} R \\ 0 \end{bmatrix} \equiv \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix}, \quad P_m^T = P_m^{-1}. \tag{3.6}$$

Since B has rank m, P_{11} is zero, P_{21} is an $n \times m$ matrix of orthonormal columns, and, see (3.1), $B = V_m R_m = P_{21} R$. If upper triangular R_m and R are both chosen to have positive diagonal elements in $B^T B = R_m^T R_m = R^T R$, then $R_m = R$ by uniqueness, so $P_{21} = V_m$ can be found from any QR factorization of the augmented matrix $\tilde{B}$.

The last n columns of P_m are then arbitrary up to an $n \times n$ orthogonal multiplier, but in theory the Householder reduction produces, see [5, (2.7)–(2.8)], the (surprisingly symmetric) orthogonal matrix

$$P_m = \begin{bmatrix} O_m & V_m^T \\ V_m & I - V_m V_m^T \end{bmatrix}, \tag{3.7}$$

showing that in this case P_m is fully defined by V_m.

A crucial result for this paper is that the Householder QR factorization giving (3.6) is also [illegible] equivalent to MGS applied to B. A close look at this Householder reduction, see, for example, [5, (2.6)–(2.7)], shows that for the computed version

$$\bar{P}_m^T \equiv \bar{P}^{(m)} \cdots \bar{P}^{(1)}, \qquad \bar{P}^{(j)} = I - \bar{p}_j \bar{p}_j^T, \qquad \bar{p}_j = \begin{bmatrix} -e_j \\ \bar{v}_j \end{bmatrix}, \qquad j = 1, \ldots, m, \tag{3.8}$$

where the $\bar{v}_j$ are [illegible] to the computed $\bar{v}_j$ in (3.2), so for example after the first two Householder transformations, our computed equivalent of $\bar{P}^{(2)}\bar{P}^{(1)}\tilde{B}$ is

$$\begin{bmatrix} \bar{\rho}_{11} & \bar{\rho}_{12} & \bar{\rho}_{13} & \cdots & \bar{\rho}_{1m} \\ 0 & \bar{\rho}_{22} & \bar{\rho}_{23} & \cdots & \bar{\rho}_{2m} \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \bar{b}_3^{(3)} & \cdots & \bar{b}_m^{(3)} \end{bmatrix}, \tag{3.9}$$

where the $\bar{\rho}_{jk}$ and $\bar{b}_k^{(j)}$ are also [illegible] to the corresponding computed values in (3.2) and (3.3). That is, in practical computations, the $\bar{v}_j$, $\bar{\rho}_{jk}$, and $\bar{b}_k^{(j)}$ are [illegible] in both algorithms; see [5, p. 179]. Note that the jth row of $\bar{R}_m$ is completely formed in the jth step and not touched again, while $\bar{b}_j^{(j)}$ is eliminated.

3.3. MGS applied to $n \times m$ B with $m > n$. The paper [5] was written assuming that $m \le n$ and $n \times m$ B in (3.1) had rank m, but it was mentioned in [5, p. 181] that the rank condition was not necessary for proving the equivalence mentioned in the last paragraph of section 3.2 above. For computations involving $n \times m$ B with $m > n$, Householder QR on B will stop in at most $n-1$ steps, but both MGS on B, and Householder QR on $\tilde{B}$ in (3.6), can nearly always be carried on for the full m steps. The MGS-Householder QR equivalence also holds for $m > n$, since the MGS and augmented Householder methods, being identical theoretically and numerically, either both stop with some $\bar{\rho}_{kk} = 0$, $k < m$, see (3.2), or both carry on to step m. It is this $m > n$ case we need here, and we extend the results of [5] to handle this. Because of this numerical equivalence, the backward error analysis for the Householder QR factorization of the augmented matrix in (3.6) can also be applied to the MGS algorithm on B. Two basic lemmas contribute to Theorem 3.3 below.

LEMMA 3.1. [illegible] (3.8) [illegible] [26, [illegible] 4.2] [illegible] $P = I - pp^T$ [illegible] p [illegible] P [illegible] p [illegible] (3.8) [illegible] p [illegible] $n+1$ [illegible]

LEMMA 3.2 (see [13, Lem. 19.3]). [illegible] j [illegible] $b \in \mathbf{R}^n$ [illegible] $\bar{c}$ [illegible]

$$\bar{c} = P_j \cdots P_2 P_1 (b + \Delta b), \qquad \|\Delta b\|_2 \le j\tilde{\gamma}_n \|b\|_2.$$

In Theorem 3.3, E_m will refer to rounding errors in the basic MGS algorithm, while later $\hat{E}_m$ will refer to errors in the basic MGS algorithm applied to solving the equivalent of the MGS-GMRES least squares problem, and $\tilde{E}_m$ will refer to errors in the MGS-GMRES algorithm. All these matrices will be of the following form:

$$E_m \in \mathbf{R}^{(m+n)\times m}, \quad E_m \equiv \begin{bmatrix} E'_m \\ E''_m \end{bmatrix} \begin{matrix} \}m \\ \}n \end{matrix}. \tag{3.10}$$

THEOREM 3.3. [illegible] $\bar{R}_m$ [illegible] $\bar{V}_m = [\bar{v}_1, \ldots, \bar{v}_m]$ [illegible] $B \in \mathbf{R}^{n\times m}$ [illegible] (3.1)–(3.4) [illegible] $m > n$ [illegible] $j = 1, \ldots, m$ [illegible] j [illegible] $\bar{v}_j$ [illegible] j [illegible] $\bar{R}_m$ [illegible] $\bar{b}_{j+1}^{(j+1)}, \ldots, \bar{b}_m^{(j+1)}$ [illegible] (3.9) [illegible]

$$\bar{p}_j = \begin{bmatrix} -e_j \\ \bar{v}_j \end{bmatrix}, \quad \bar{P}^{(j)} = I - \bar{p}_j \bar{p}_j^T, \quad \bar{P}_m = \bar{P}^{(1)} \bar{P}^{(2)} \cdots \bar{P}^{(m)}, \tag{3.11}$$

$$\tilde{v}_j = \bar{v}_j / \|\bar{v}_j\|_2, \quad \tilde{p}_j = \begin{bmatrix} -e_j \\ \tilde{v}_j \end{bmatrix}, \quad \tilde{P}^{(j)} = I - \tilde{p}_j \tilde{p}_j^T, \quad \tilde{P}_m = \tilde{P}^{(1)} \tilde{P}^{(2)} \cdots \tilde{P}^{(m)}.$$

$\tilde{P}^{(j)}$ [illegible] $\bar{P}^{(j)}$ [illegible] j [illegible] $\tilde{B}$ [illegible]
(3.6) [illegible] $\tilde{P}_m^T \tilde{P}_m = I$ [illegible] $\bar{R}_m$ [illegible] $R = R_m$ [illegible] (3.6) [illegible] D [illegible] 2.1 [illegible] $j = 1, \ldots, m$

$$\tilde{P}_m \begin{bmatrix} \bar{R}_m \\ 0 \end{bmatrix} = \begin{bmatrix} E'_m \\ B + E''_m \end{bmatrix}; \quad \tilde{P}_m \text{ [illegible] } \bar{R}_m, E'_m \in \mathbf{R}^{m\times m}; \tag{3.12}$$

$$E_m \equiv \begin{bmatrix} E'_m \\ E''_m \end{bmatrix}; \quad \|E_m e_j\|_2 \le j\tilde{\gamma}_n \|Be_j\|_2, \|E_m D\|_F \le m\tilde{\gamma}_n \|BD\|_F;$$

$$\|\bar{R}_m e_j\|_2 \le \|Be_j\|_2 + \|E_m e_j\|_2 \le (1 + j\tilde{\gamma}_n)\|Be_j\|_2; \tag{3.13}$$

$$E'_m e_1 = 0, \quad \|E'_m e_j\|_2 \le j^{\frac{1}{2}} \tilde{\gamma}_n \|Be_j\|_2, \quad j = 2, \ldots, m; \tag{3.14}$$

$$\|E'_m D\|_F \le m^{\frac{1}{2}} \tilde{\gamma}_n \|(BD)_{2:m}\|_F;$$

$$\tilde{P}_m = \begin{bmatrix} \tilde{S}_m & (I - \tilde{S}_m)\tilde{V}_m^T \\ \tilde{V}_m(I - \tilde{S}_m) & I - \tilde{V}_m(I - \tilde{S}_m)\tilde{V}_m^T \end{bmatrix}, \qquad \tilde{P}_m \tilde{P}_m^T = I, \tag{3.15}$$

$m \times m$ E'_m [illegible] $\tilde{S}_m$ [illegible] j [illegible] E'_m [illegible] j [illegible] $\bar{R}_m$ [illegible] j [illegible] $\tilde{S}_m$ [illegible] j [illegible] $\bar{\rho}_{kk} = 0$ [illegible] (3.2) [illegible] $k, \ldots, m$ [illegible] $\bar{R}_m$ [illegible] E'_m [illegible] $k, \ldots, m$ [illegible] $\bar{V}_m$ [illegible] $\tilde{S}_m$ [illegible] m [illegible] k [illegible]. The MGS-augmented Householder QR equivalence for the case of $m \le n$ was proven in [5], and that this extends to $m > n$ is proven in the first paragraph of section 3.3. As a result we can apply Lemmas 3.1 and 3.2 to give (3.12)–(3.13). The ideal P in (3.6) has the structure in (3.7), but it was shown in [5, Thm. 4.1, and (4.5)] (which did not require $n \ge m$ in our notation) that $\tilde{P}_m$ in (3.11) and (3.12) has the extremely important structure of (3.15) for some strictly upper triangular $m \times m$ $\tilde{S}_m$. Since $E'_m = \tilde{S}_m \bar{R}_m$, this is strictly upper triangular too.

The rest follow with Lemmas 3.1 and 3.2. We have used $\tilde{\gamma}_n = \tilde{\gamma}'_{n+1}$ rather than $\tilde{\gamma}_{m+n}$ because in each step, $\bar{p}_j$ in (3.11) has only $n+1$ elements; see (3.9) and Lemma 3.1. Row j in $\bar{R}_m$ is not touched again after it is formed in step j, see (3.9), and so the same is true for row j in E'_m in (3.12); see Lemma 3.1. Since $E'_m = \tilde{S}_m \bar{R}_m$, the jth column of $\tilde{S}_m$ is not defined until $\bar{\rho}_{jj}$ is computed in step j, and since these three matrices are all upper triangular, it is not altered in later steps. Finally we obtain new bounds in (3.14). The element $\bar{\rho}_{ij}$ is formed by the ith transformation $\bar{P}^{(i)}$ in (3.11) applied to $\bar{b}_j^{(i)}$ in (3.9), and so from Lemma 3.2 we can say (remember $(E'_m)_{ii} = 0$)

$$|(E'_m)_{ij}| \le \tilde{\gamma}_n \|\bar{b}_j^{(i)}\|_2 \le \tilde{\gamma}'_n \|Be_j\|_2, \quad j = i+1, \ldots, m,$$

which is quite loose but leads to the bounds in (3.14). ∎

Note that (3.14) involves $j^{\frac{1}{2}}$, rather than the j in previous publications.

Lemma 3.1. It is counterintuitive that E'_m is strictly upper triangular, so we will explain it. We need only consider the first augmented Householder-MGS transformation of the first vector to form $\bar{\rho}_{11}$ in (3.9). We can rewrite the relevant part of the first transformation ideally as, see (3.11) and Lemma 3.1,

$$P \begin{bmatrix} 0 \\ b \end{bmatrix} = \begin{bmatrix} \rho \\ 0 \end{bmatrix}, \qquad P = \begin{bmatrix} 0 & v^T \\ v & I - vv^T \end{bmatrix}, \qquad b = v\rho, \qquad \|v\|_2 = 1.$$

From b we compute $\bar{\rho}$ and $\bar{v}$ and then define $\tilde{v} \equiv \bar{v}/\|\bar{v}\|_2$ so $\|\tilde{v}\|_2 = 1$. In order for $E'_m e_1 = 0$ in (3.12), there must exist a backward error term Δb such that

$$\begin{bmatrix} 0 & \tilde{v}^T \\ \tilde{v} & I - \tilde{v}\tilde{v}^T \end{bmatrix} \begin{bmatrix} 0 \\ b + \Delta b \end{bmatrix} = \begin{bmatrix} \bar{\rho} \\ 0 \end{bmatrix},$$

which looks like $n+1$ conditions on the n-vector Δb. But multiplying throughout by P shows there is a solution $\Delta b = \tilde{v}\bar{\rho} - b$. The element above Δb is forced to be zero, so that there are actually $n+1$ conditions on $n+1$ unknowns. An error analysis (see Lemma 3.2) then bounds $\|\Delta b\|_2 \le \tilde{\gamma}_n \|b\|_2$.

4. The Arnoldi algorithm *as* MGS. The Arnoldi algorithm [2] is the basis of MGS-GMRES. We assume that the initial estimate of x in (1.1) is $x_0 = 0$, so that the initial residual $r_0 = b$, and use the Arnoldi algorithm with $\rho \equiv \|b\|_2$, $v_1 \equiv b/\rho$, to sequentially generate the columns of $V_{k+1} \equiv [v_1, \ldots, v_{k+1}]$ via the ideal process:

$$AV_k = V_k H_{k,k} + v_{k+1} h_{k+1,k} e_k^T = V_{k+1} H_{k+1,k}, \qquad V_{k+1}^T V_{k+1} = I_{k+1}. \tag{4.1}$$

Here $k \times k$ $H_{k,k} = (h_{ij})$ is upper Hessenberg, and we stop at the first $h_{k+1,k} = 0$. Because of the orthogonality, this ideal algorithm must stop for some $k \le n$. Then $AV_k = V_k H_{k,k}$, where $H_{k,k}$ has rank at least $k-1$. If $h_{k+1,k} = 0$ and $H_{k,k}$ has rank $k-1$, there exists a nonzero z such that $AV_k z = V_k H_{k,k} z = 0$, so that A must be singular. Thus when A is nonsingular so is $H_{k,k}$, and so in MGS-GMRES, solving $H_{k,k} y = e_1 \rho$ and setting $x = V_k y$ solves (1.1). But if A is singular, this might not provide a solution even to consistent $Ax = b$:

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad x = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad v_1 = b = Ax = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad AV_1 = V_1 H_{1,1}, \quad H_{1,1} = 0.$$

Thus it is no surprise that we will require a restriction of the form (1.1) to ensure that the MGS-GMRES algorithm always obtains a meaningful solution.

To relate the Arnoldi and MGS-GMRES algorithms to the MGS algorithm, we now replace $k+1$ by m and say that in the mth MGS step these produce v_m, and MGS-GMRES also produces the approximation $x_{m-1} = V_{m-1}y_{m-1}$ to the solution x of (1.1). Then apart from forming the Av_j, the algorithm we use to give (4.1) is identical to (3.2)–(3.3) with the same vectors v_j, and

$$b_1 \equiv b,\ \rho_{11} \equiv \rho; \quad \text{and for } j=1,\dots,m-1,\ b_{j+1} \equiv Av_j,\ \rho_{i,j+1} \equiv h_{i,j}\ i=1,\dots,j+1,$$

except that Av_j cannot be formed and orthogonalized against $v_1,\dots,v_j$ until v_j is available. This does not alter the numerical values. Thus with upper triangular R_m,

$$B_m \equiv A\,[x, V_{m-1}] = [b, AV_{m-1}] = V_m\,[e_1\rho, H_{m,m-1}] \equiv V_m R_m, \quad V_m^T V_m = I. \tag{4.2}$$

So in theory [illegible] $B_m \equiv [b, AV_{m-1}]$ [illegible] B_m. Computationally we can see that we have applied MGS to $\bar{B}_m \equiv [b, fl(A\bar{V}_{m-1})]$, where $\bar{V}_{m-1} \equiv [\bar{v}_1,\dots,\bar{v}_{m-1}]$ is the matrix of supposedly orthonormal vectors computed by MGS, and see, for example, [13, section 3.5],

$$fl(A\bar{v}_j) = (A+\Delta A_j)\bar{v}_j, \quad |\Delta A_j| \le \gamma_n|A|, \ \text{ so } \ fl(A\bar{V}_{m-1}) = A\tilde{V}_{m-1}+\Delta V_{m-1},$$
$$|\Delta V_{m-1}| \le \gamma_n|A|.|\bar{V}_{m-1}|, \quad \|\Delta V_{m-1}\|_F \le m^{\frac{1}{2}}\gamma_n\||A|\|_2 \le m^{\frac{1}{2}}\gamma_n\|A\|_F, \tag{4.3}$$

gives the computed version of $A\bar{V}_{m-1}$. We could replace n by the maximum number of nonzeros per row, while users of preconditioners, or less simple multiplications, could insert their own bounds on ΔV_{m-1} here.

4.1. The bounds in (4.3) are not column-scaling independent. Also any scaling applies to the columns of $A\bar{V}_{m-1}$, not to A, and so would not be of such an advantage for MGS-GMRES as for ordinary MGS. Therefore it would seem important to ensure the columns of A are reasonably scaled for MGS-GMRES—e.g., to approach the minimum over positive diagonal D of $\|AD\|_F/\sigma_{min}(AD)$; see the appendix.

The rounding error behavior of the Arnoldi algorithm is as follows.

THEOREM 4.1. [illegible] (4.1) [illegible] $m \equiv k+1$ [illegible] ϵ [illegible] $\bar{V}_m$ [illegible] $\bar{R}_m \equiv [e_1\bar{\rho}, \bar{H}_{m,m-1}]$ [illegible] (4.2) [illegible] $n+m$ [illegible] $\tilde{P}_m$ [illegible] (3.15) [illegible] $\tilde{V}_m$ [illegible] $\bar{V}_m$ [illegible]

$$\bar{B}_m \equiv [b, fl(A\bar{V}_{m-1})] = [b, A\tilde{V}_{m-1}] + [0, \Delta V_{m-1}], \tag{4.4}$$

[illegible] ΔV_{m-1} [illegible] (4.3) [illegible] 3.3 [illegible] B [illegible] $\bar{B}_m$ [illegible]

Thus whatever we say for MGS will hold for the Arnoldi algorithm if we simply replace B by $\bar{B}_m \equiv [b, fl(A\bar{V}_{m-1})] = [b, A\tilde{V}_{m-1}]+[0, \Delta V_{m-1}]$. The key idea of viewing the Arnoldi algorithm as MGS applied to $[b, AV_n]$ appeared in [25]. It was used in [8] and [1], and in particular in [18], in which we outlined another possible approach to backward stability analysis of MGS-GMRES. Here we have chosen a different way of proving the backward stability result, and this follows the spirit of [5] and [10].

5. Loss of orthogonality of $\bar{V}_m$ from MGS and the Arnoldi algorithm. The analysis here is applicable to both the MGS and Arnoldi algorithms. B will denote the given matrix in MGS, or $\bar{B}_m \equiv [b, fl(A\bar{V}_{m-1})]$ in the Arnoldi algorithm. Unlike [10, 14], we do not base the theory on [5, Lem. 3.1], since a direct approach is cleaner and gives nicer results. It is important to be aware that our bounds will

be of a different nature to those in [10, 14]. Even though the rounding error analysis of MGS in [10, 14] is based on the ideas in [5], the bounds obtained in [10] and [14, pp. 32–38] are unexpectedly strong compared with our results based on [5]. This is because [10, (18)–(19)] and [14, (1.68)–(1.69)] leading to [10, Thm. 3.1] and [14, Thm. 1.4.1] follow from [26, p. 160, (45.3)]. But in Wilkinson [26], (45.3) follows from his (45.2), (45.1), and (44.6), where this last is clearly for fl_2 arithmetic (double precision accumulation of inner products). Since double precision is used in [10, 14], their analysis is essentially assuming what could be called fl_4—quadruple precision accumulation of inner products. This is not stated in [10, 14], and the result is that their bounds appear to be much better (tighter) and the conditions much easier (less strict) than those that would have been obtained using standard floating point arithmetic. We will now obtain refined bounds based on our standard floating point arithmetic analysis and attempt to correct this misunderstanding.

5.1. The $\tilde\gamma_n$ in each expression in (3.12)–(3.14) is essentially the same $\tilde\gamma_n$, that from Lemma 3.2, so we will call it $\hat\gamma_n$. We could legitimately absorb various small constants into a series of new $\tilde\gamma_n$, but that would be less transparent, so we will develop a sequence of loose bounds based on this fixed $\hat\gamma_n$.

To simplify our bounds, we use "$\{\le\}$" to mean "$\le$" under the assumption that $m\hat\gamma_n\tilde\kappa_F(B) \le 1/8$. Note that this has the following consequences:

$$m\hat\gamma_n\tilde\kappa_F(B) \le 1/8 \quad \Rightarrow \quad \{(1-m\hat\gamma_n\tilde\kappa_F(B))^{-1} \le 8/7 \quad \& \tag{5.1}$$
$$\mu \equiv m^{\frac12}\hat\gamma_n\tilde\kappa_F(B)8/7 \le 1/7 \qquad \& \quad (1+\mu)/(1-\mu) \le 4/3\}.$$

The basic bound is for $\tilde S_m = E_m'\bar R_m^{-1}$; see (3.12), (3.15). This is part of an orthogonal matrix so $\|\tilde S_m\|_2 \le 1$. From (3.12) and (3.14) for any $m\times m$ diagonal matrix $D>0$,

$$\|\tilde S_m\|_F = \|E_m'D(\bar R_mD)^{-1}\|_F \le \|E_m'D\|_F\|(\bar R_mD)^{-1}\|_2 = \|E_m'D\|_F/\sigma_{min}(\bar R_mD)$$
$$\le \frac{\|E_m'D\|_F}{\sigma_{min}(BD)-\|E_mD\|_2} \le \frac{m^{\frac12}\hat\gamma_n\|(BD)_{2:m}\|_F}{\sigma_{min}(BD)-m\hat\gamma_n\|BD\|_F}, \tag{5.2}$$
$$\|\tilde S_m\|_F \le m^{\frac12}\hat\gamma_n\tilde\kappa_F(B)/(1-m\hat\gamma_n\tilde\kappa_F(B)) \ \{\le\} \ \frac87 m^{\frac12}\hat\gamma_n\tilde\kappa_F(B) \ \{\le\} \ \frac17, \tag{5.3}$$

with obvious restrictions. The bounds (5.3) took a minimum over D.

$\bar V_m \equiv [\bar v_1,\dots,\bar v_m]$ is the $n\times m$ matrix of vectors computed by m steps of MGS, $\tilde V_m \equiv [\tilde v_1,\dots,\tilde v_m]$ is the correctly normalized version of $\bar V_m$, so $\tilde V_m$ satisfies (2.2)–(2.3). Since $I-\tilde S_m$ is nonsingular upper triangular, the first m rows of $\tilde P_m$ in (3.15) give

$$\begin{aligned}(I-\tilde S_m)\tilde V_m^T\tilde V_m(I-\tilde S_m)^T &= I-\tilde S_m\tilde S_m^T\\ &= (I-\tilde S_m)(I-\tilde S_m)^T+(I-\tilde S_m)\tilde S_m^T+\tilde S_m(I-\tilde S_m)^T,\end{aligned}$$
$$\tilde V_m^T\tilde V_m = I+\tilde S_m^T(I-\tilde S_m)^{-T}+(I-\tilde S_m)^{-1}\tilde S_m, \tag{5.4}$$
$$\begin{aligned}(I-\tilde S_m)^{-1}\tilde S_m &= \tilde S_m(I-\tilde S_m)^{-1}\\ &= \text{strictly upper triangular part}(\tilde V_m^T\tilde V_m).\end{aligned} \tag{5.5}$$

Since $\tilde V_{m-1}^T\tilde v_m$ is the above diagonal part of the last column of symmetric $\tilde V_m^T\tilde V_m - I$, (5.5) and (5.3) give the key bound (at first using $2m\hat\gamma_n\tilde\kappa_F(B)<1$; see (5.1)),

$$\sqrt2\|\tilde V_{m-1}^T\tilde v_m\|_2 \le \|I-\tilde V_m^T\tilde V_m\|_F = \sqrt2\|(I-\tilde S_m)^{-1}\tilde S_m\|_F \tag{5.6}$$
$$\le \sqrt2\|\tilde S_m\|_F/(1-\|\tilde S_m\|_2) \le (2m)^{\frac12}\hat\gamma_n\tilde\kappa_F(B)/[1-(m+m^{\frac12})\hat\gamma_n\tilde\kappa_F(B)],$$
$$\{\le\} \ \frac43(2m)^{\frac12}\hat\gamma_n\tilde\kappa_F(B) \qquad \text{(cf. [3, 5, (5.3)])},$$

and similarly for $\bar{V}_m$; see (2.2). This is superior to the bound in [5], but the scaling idea is not new. Higham [13, p. 373] (and in the 1996 first edition) argued that $\kappa_2(B)$ in [5, 3], see (3.5), might be replaced by the minimum over positive diagonal matrices D of $\kappa_2(BD)$, which is almost what we have proven using $\tilde{\kappa}_F(B)$ in (2.1).

One measure of the extent of loss of orthogonality of $\tilde{V}_m$ is $\kappa_2(\tilde{V}_m)$.

LEMMA 5.1. $\tilde{V}_m^T\tilde{V}_m = I + \tilde{F}_m + \tilde{F}_m^T$ [illegible] $\tilde{F}_m$ [illegible] $\tilde{S}_m$ [illegible] $\tilde{F}_m \equiv \tilde{S}_m(I-\tilde{S}_m)^{-1}$ [illegible] (5.4) [illegible] $\sigma_i(\tilde{V}_m)$

$$\frac{1-\|\tilde{S}_m\|_2}{1+\|\tilde{S}_m\|_2} \le \sigma_i^2(\tilde{V}_m) \le \frac{1+\|\tilde{S}_m\|_2}{1-\|\tilde{S}_m\|_2}, \qquad \kappa_2(\tilde{V}_m) \le \frac{1+\|\tilde{S}_m\|_2}{1-\|\tilde{S}_m\|_2}.$$

[illegible]. Obviously $\|\tilde{F}_m\|_2 \le \|\tilde{S}_m\|_2/(1-\|\tilde{S}_m\|_2)$. For any $y \in \mathbf{R}^k$ such that $\|y\|_2 = 1$, $\|\tilde{V}_m y\|_2^2 = 1+2y^T\tilde{F}_m y \le 1+2\|\tilde{F}_m\|_2 \le (1+\|\tilde{S}_m\|_2)/(1-\|\tilde{S}_m\|_2)$, which gives the upper bound on every $\sigma_i^2(\tilde{V}_m)$. From (5.4) $(I-\tilde{S}_m)\tilde{V}_m^T\tilde{V}_m(I-\tilde{S}_m)^T = I-\tilde{S}_m\tilde{S}_m^T$, so for any $y \in \mathbf{R}^k$ such that $\|y\|_2 = 1$, define $z \equiv (I-\tilde{S}_m)^T y$ so $\|z\|_2 \le 1+\|\tilde{S}_m\|_2$ and then

$$\frac{z^T\tilde{V}_m^T\tilde{V}_m z}{z^T z} = \frac{1-y^T\tilde{S}_m\tilde{S}_m^T y}{z^T z} \ge \frac{1-\|\tilde{S}_m\|_2^2}{(1+\|\tilde{S}_m\|_2)^2} = \frac{1-\|\tilde{S}_m\|_2}{1+\|\tilde{S}_m\|_2},$$

giving the lower bound on every $\sigma_i^2(\tilde{V}_m)$. The bound on $\kappa_2(\tilde{V}_m)$ follows. ∎

Combining Lemma 5.1 with (5.1) and (5.3) gives the major result

$$\text{(5.7)} \quad \text{for } j=1,\dots,m, \;\; j\hat{\gamma}_n\tilde{\kappa}_F(B_j) \le 1/8 \;\; \Rightarrow \;\; \|\tilde{S}_j\|_F \le 1/7 \;\; \Rightarrow \;\; \kappa_2(\tilde{V}_j),\ \sigma_{min}^{-2}(\tilde{V}_j),\ \sigma_{max}^2(\tilde{V}_j) \;\le 4/3.$$

At this level the distinction between $\kappa_2(\bar{V}_m)$ and $\kappa_2(\tilde{V}_m)$ is miniscule, see (2.2), and by setting $j = m$ we can compare this with the elegant result which was the main theorem of Giraud and Langou [10]; see [14, Thm. 1.4.1].

THEOREM 5.2 (see [10, Thm. 3.1; 14, Thm. 1.4.1]). [illegible] $B \in \mathbf{R}^{n\times m}$ [illegible] $m \le n$ [illegible] $\kappa_2(B)$ [illegible]

$$\text{(5.8)} \qquad 2.12(m+1)\epsilon < 0.01 \quad [\text{illegible}] \quad 18.53m^{\frac{3}{2}}\epsilon\kappa_2(B) \le 0.1.$$

[illegible] (present comment in 2005: actually fl_2, or fl_4 if we use double precision) [illegible] $\bar{V}_m \in \mathbf{R}^{n\times m}$ [illegible]

$$\kappa_2(\bar{V}_m) \le 1.3. \qquad \blacksquare$$

Note that the conditions (5.8) do not involve the dimension n of each column of $\bar{V}_m$, and this is the result of their analysis using fl_2. We can assume m satisfying the second condition in (5.8) will also satisfy the first.

To compare Theorem 5.2 with $j = m$ in (5.7), note that $m\tilde{\gamma}_n$ essentially means $\tilde{c}mn\epsilon$ for some constant $\tilde{c} > 1$, probably less than the 18.53 in Theorem 5.2. We assumed standard (IEEE) floating point arithmetic, but if we had assumed fl_2 arithmetic, that would have eliminated the n from our condition in (5.7). We used (2.1), which involves $\|BD\|_F \le m^{\frac{1}{2}}\|BD\|_2$. If we inserted this upper bound, that would mean our condition would be like that in Theorem 5.2, except we have the optimal result over column scaling; see (2.1). So if the same arithmetic is used, (5.7) is more revealing than Theorem 5.2. It is worth noting that with the introduction of XBLAS [7], the fl_2 and fl_4 options may become available in the near future.

6. A critical step in the Arnoldi and MGS-GMRES iterations. It will simplify the analysis if we use (5.7) to define a distinct value $\hat{m}$ of m. This value will depend on the problem and the constants we have chosen, but it will be sufficient for us to prove convergence and backward stability of MGS-GMRES in $\hat{m}-1 \le n$ steps. For the ordinary MGS algorithm remember $\bar{B}_m = B_m$, and think of m as increasing.

$$\text{Let } \hat{m} \text{ be the smallest integer such that } \kappa_2(\tilde{V}_{\hat{m}}) > 4/3 \tag{6.1}$$

then we know from (5.7) that for $\bar{B}_{\hat{m}}$ in the Arnoldi algorithm, see (4.4) and (2.1),

$$\hat{m}\hat{\gamma}_n\tilde{\kappa}_F(\bar{B}_{\hat{m}}) > 1/8, \text{ so } \sigma_{min}(\bar{B}_{\hat{m}}D) < 8\hat{m}\hat{\gamma}_n\|\bar{B}_{\hat{m}}D\|_F \ \forall \text{ diagonal } D > 0. \tag{6.2}$$

But since $\sigma_{min}(\tilde{V}_j) \le \sigma_1(\tilde{v}_1) = \|\tilde{v}_1\|_2 = 1 \le \sigma_{max}(\tilde{V}_j)$, (6.1) also tells us that

$$\kappa_2(\tilde{V}_j),\ \sigma_{min}^{-1}(\tilde{V}_j),\ \sigma_{max}(\tilde{V}_j) \ \le 4/3, \quad j = 1, \ldots, \hat{m}-1. \tag{6.3}$$

The above reveals the philosophy of the present approach to proving backward stability of MGS-GMRES. Other approaches have been tried. Here all is based on $\tilde{\kappa}_F(\bar{B}_m)$ rather than the backward error or residual norm. In [12, Thm. 3.2, p. 713] a different approach was taken—the assumption was directly related to the norm of the residual. The present approach leads to very compact and elegant formulations, and it is hard to say now whether the earlier approaches (see [18]) would have succeeded.

7. Least squares solutions via MGS. The linear least squares problem

$$\hat{y} \equiv \arg\min_y \|b - Cy\|_2, \qquad \hat{r} \equiv b - C\hat{y}, \qquad C \in \mathbf{R}^{n\times(m-1)}, \tag{7.1}$$

may be solved via MGS in different ways. Here we discuss two of these ways, but first we remind the reader how this problem appears in MGS-GMRES with $C = AV_{m-1}$.

After carrying out step $m-1$ of the Arnoldi algorithm as in section 4 to produce $[b, AV_{m-1}] = V_m R_m$, see (4.2), the MGS-GMRES algorithm in theory minimizes the 2-norm of the residual $\|r_{m-1}\|_2 = \|b - Ax_{m-1}\|_2$ over $x_{m-1} \in x_0 + \mathcal{K}_{m-1}(A, r_0)$, where for simplicity we are assuming $x_0 = 0$ here. It does this by using V_{m-1} from (4.1) to provide an approximation $x_{m-1} \equiv V_{m-1}y_{m-1}$ to the solution x of (1.1). Then the corresponding residual is

$$r_{m-1} \equiv b - Ax_{m-1} = [b, AV_{m-1}] \begin{bmatrix} 1 \\ -y_{m-1} \end{bmatrix} = V_m R_m \begin{bmatrix} 1 \\ -y_{m-1} \end{bmatrix}, \tag{7.2}$$

where $R_m \equiv [e_1\rho, H_{m,m-1}]$. The ideal least squares problem is

$$y_{m-1} = \arg\min_y \| [b, AV_{m-1}] \begin{bmatrix} 1 \\ -y \end{bmatrix} \|_2, \tag{7.3}$$

but (in theory) the MGS-GMRES least squares solution is found by solving

$$y_{m-1} \equiv \arg\min_y \| R_m \begin{bmatrix} 1 \\ -y \end{bmatrix} \|_2. \tag{7.4}$$

7.1. The MGS least squares solution used in MGS-GMRES. If $B = [C, b]$ in (3.1)–(3.4), and C has rank $m-1$, then it was shown in [5, (6.3)], see also [13, section 20.3], that MGS can be used to compute $\hat{y}$ in (7.1) in a backward stable way. Here we need to show that we can solve (7.1) in a stable way with MGS applied

to $B = [b, C]$ (note the reversal of C and b) in order to prove the backward stability of MGS-GMRES. Just remember $B = [b, C] \equiv \bar{B}_m$ in (4.4) for MGS-GMRES. The analysis could be based directly on [5, Lem. 3.1], but the following is more precise.

Let MGS on B in (3.1) lead to the computed $\bar{R}_m$ (we can assume $\bar{R}_m$ is nonsingular; see later) satisfying (3.12), where $B = [b, C]$. Then (3.12) and (7.1) give

$$\tilde{P}_m \begin{bmatrix} \bar{R}_m \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ [b, C] \end{bmatrix} + E_m; \quad \|E_m e_j\|_2 \le j\tilde{\gamma}_n \|[b, C]e_j\|_2, \quad j = 1, \ldots, m, \tag{7.5}$$

$$\hat{y} \equiv \arg\min_y \|B \begin{bmatrix} 1 \\ -y \end{bmatrix} \|_2, \qquad \hat{r} = B \begin{bmatrix} 1 \\ -\hat{y} \end{bmatrix}. \tag{7.6}$$

To solve the latter computationally, having applied MGS to B to give $\bar{R}_m$, we

$$\text{carry out a backward stable solution of } \min_y \|\bar{R}_m \begin{bmatrix} 1 \\ -y \end{bmatrix} \|_2 \tag{7.7}$$

by orthogonal reduction followed by the solution of a triangular system. With (3.13) we will see this leads to

$$\hat{Q}^T(\bar{R}_m + \Delta R_m) = \begin{bmatrix} \bar{t} & \bar{U} + \Delta U \\ \bar{\tau} & 0 \end{bmatrix}, \qquad (\bar{U} + \Delta U)\bar{y} = \bar{t}, \tag{7.8}$$

$$\|\Delta R_m e_j\|_2 \le \tilde{\gamma}'_m \|\bar{R} e_j\|_2 \le \tilde{\gamma}_m \|B e_j\|_2 = \tilde{\gamma}_m \|[b, C]e_j\|_2, \quad j = 1, \ldots, m,$$

where $\hat{Q}$ is an orthogonal matrix while $\bar{\tau}$, $\bar{t}$, nonsingular upper triangular $\bar{U}$, and $\bar{y}$ are computed quantities. Here ΔU is the backward rounding error in the solution of the upper triangular system to give $\bar{y}$, see, for example, [13, Thm. 8.3], and ΔR_m was obtained by combining ΔU with the backward rounding error in the QR factorization that produced $\bar{\tau}$, $\bar{t}$ and $\bar{U}$; see, for example, [13, Thm. 19.10] (where here there are $m-1$ stages, each of one rotation). Clearly $\bar{y}$ satisfies

$$\bar{y} = \arg\min_y \left\| (\bar{R}_m + \Delta R_m) \begin{bmatrix} 1 \\ -y \end{bmatrix} \right\|_2. \tag{7.9}$$

In order to relate this least squares solution back to the MGS factorization of B, we add the error term ΔR_m to (7.5) to give (replacing $j\tilde{\gamma}_n + \tilde{\gamma}_m$ by $j\tilde{\gamma}_n$)

$$\tilde{P}_m \begin{bmatrix} (\bar{R}_m + \Delta R_m) \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ [b, C] \end{bmatrix} + \hat{E}_m, \quad \hat{E}_m \equiv E_m + \tilde{P}_m \begin{bmatrix} \Delta R_m \\ 0 \end{bmatrix}, \tag{7.10}$$

$$\|\hat{E}_m e_j\|_2 \le j\tilde{\gamma}_n \|[b, C]e_j\|_2, \quad j = 1, \ldots, m.$$

Now we can write for any $y \in \mathbf{R}^{m-1}$

$$r = r(y) \equiv b - Cy, \quad p = p(y) \equiv \tilde{P}_m \begin{bmatrix} (\bar{R}_m + \Delta R_m) \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ -y \end{bmatrix} = \begin{bmatrix} 0 \\ r \end{bmatrix} + \hat{E}_m \begin{bmatrix} 1 \\ -y \end{bmatrix}, \tag{7.11}$$

and we see from (2.6) in Lemma 2.3 that for any $y \in \mathbf{R}^{m-1}$ there exists $N(y)$ so that

$$\|p(y)\|_2 = \left\| (\bar{R}_m + \Delta R_m) \begin{bmatrix} 1 \\ -y \end{bmatrix} \right\|_2 = \left\| b - Cy + N(y)\hat{E}_m \begin{bmatrix} 1 \\ -y \end{bmatrix} \right\|_2, \quad \|N(y)\|_2 \le \sqrt{2}.$$

Defining $[\Delta b(y), \Delta C(y)] \equiv N(y)\hat{E}_m$ shows that for all $y \in \mathbf{R}^{m-1}$

$$\left\| (\bar{R}_m + \Delta R_m) \begin{bmatrix} 1 \\ -y \end{bmatrix} \right\|_2 = \|b + \Delta b(y) - [C + \Delta C(y)]y\|_2. \tag{7.12}$$

Thus $\bar{y}$ in (7.9) also satisfies

$$\bar{y} = \arg\min_y \|b+\Delta b(y)-[C+\Delta C(y)]y\|_2, \tag{7.13}$$
$$\|[\Delta b(y), \Delta C(y)]e_j\|_2 \le j\tilde{\gamma}_n\|[b,C]e_j\|_2, \qquad j=1,\dots,m,$$

where the bounds are independent of y, so that $\bar{y}$ is a backward stable solution for (7.1). That is, MGS applied to $B=[b,C]$ followed by (7.7) is backward stable as long as the computed $\bar{R}_m$ from MGS is nonsingular (we can stop early to ensure this). The almost identical analysis and result applies wherever b is in B, but we just gave the $B=[b,C]$ case for clarity.

Since we have a backward stable solution $\bar{y}$, we expect various related quantities to have reliable values, and we now quickly show two cases of this. If $\|E\|_F \le \gamma\|B\|_F$, then $\|Ey\|_2^2 = \sum_i \|e_i^T Ey\|_2^2 \le \sum_i \|e_i^T E\|_2^2\|y\|_2^2 = \|E\|_F^2\|y\|_2^2 \le \gamma^2\|B\|_F^2\|y\|_2^2$. So from the bounds in (7.10) we have for any $y \in \mathbf{R}^{m-1}$ the useful basic bound

$$\left\|\hat{E}_m\begin{bmatrix}1\\-y\end{bmatrix}\right\|_2 \le \tilde{\gamma}_{mn}\psi_m(y), \quad \psi_m(y) \equiv \|b\|_2+\|C\|_F\|y\|_2. \tag{7.14}$$

Multiplying (7.8) and (7.10) on the right by $\begin{bmatrix}1\\-\bar{y}\end{bmatrix}$ shows that the residual $\bar{r}$ satisfies

$$\bar{r} \equiv b - C\bar{y}, \quad \tilde{P}_m\begin{bmatrix}\hat{Q}e_m\bar{\tau}\\0\end{bmatrix} = \begin{bmatrix}0\\\bar{r}\end{bmatrix} + \hat{E}_m\begin{bmatrix}1\\-\bar{y}\end{bmatrix}, \quad |\,\|\bar{r}\|_2 - |\bar{\tau}|\,| \le \tilde{\gamma}_{mn}\psi_m(\bar{y}), \tag{7.15}$$

so that $|\tau|$ approximates $\|\bar{r}\|_2$ with a good relative error bound. Multiplying the last equality in this on the left by $[\tilde{V}_m, I_n]$, and using (3.15), (3.12), (7.10), (7.8), (3.14), and (2.3) with the argument leading to (7.14), we see that

$$\tilde{V}_m\hat{Q}e_m\bar{\tau} = \bar{r} + [\tilde{V}_m, I_n]\hat{E}_m\begin{bmatrix}1\\-\bar{y}\end{bmatrix} = \bar{r} + [\tilde{V}_m(E'_m+\Delta R_m)+E''_m]\begin{bmatrix}1\\-\bar{y}\end{bmatrix}, \tag{7.16}$$
$$\|\bar{r} - \tilde{V}_m\hat{Q}e_m\bar{\tau}\|_2 \le \tilde{\gamma}_{mn}\psi_m(\bar{y}) \text{ for } m < \hat{m} \text{ in (6.1)}.$$

Thus $\bar{V}_m\hat{Q}e_m\bar{\tau}$ also approximates $\bar{r} \equiv b - C\bar{y}$ with a good relative error bound; see (2.2) and its following sentence.

7.2. Least squares solutions and loss of orthogonality in MGS. An apparently strong relationship was noticed between convergence of finite precision MGS-GMRES and loss of orthogonality among the Arnoldi vectors; see [12, 19]. It was thought that if this relationship was fully understood, we might use it to prove that finite precision MGS-GMRES would necessarily converge; see, for example, [18]. A similar relationship certainly exist—it is the relationship between the loss of orthogonality in ordinary MGS applied to B, and the residual norms for what we will call the last vector least squares (LVLS) problems involving B, and we will derive this here. It adds to our understanding, but it is not necessary for our other proofs and could initially be skipped.

Because this is a theoretical tool, we will only consider rounding errors in the MGS part of the computation. We will do the analysis for MGS applied to any matrix $B=[b_1,\dots,b_m]$. After step j we have $n\times j$ $\bar{V}_j$ and $j\times j$ $\bar{R}_j$, so that

$$\bar{R}_j \equiv \begin{bmatrix}\bar{U}_j & \bar{t}_j\\ & \bar{\tau}_j\end{bmatrix}, \quad \bar{U}_j\bar{y}_j = \bar{t}_j, \quad \bar{y}_j = \arg\min_y\left\|\bar{R}_j\begin{bmatrix}-y\\1\end{bmatrix}\right\|_2, \quad |\bar{\tau}_j| = \left\|\bar{R}_j\begin{bmatrix}-\bar{y}_j\\1\end{bmatrix}\right\|_2. \tag{7.17}$$

In theory $\bar{y}_j$ minimizes $\|b_j - B_{j-1}y\|_2$, but we would like to know that loss of orthogonality caused by rounding errors in MGS does not prevent this. One indicator of loss of orthogonality is $\tilde{V}_{j-1}^T \tilde{v}_j$. From (7.17) we see that

$$\bar{R}_j^{-1} = \begin{bmatrix} \bar{U}_j^{-1} & -\bar{U}_j^{-1}\bar{t}_j\bar{\tau}_j^{-1} \\ & \bar{\tau}_j^{-1} \end{bmatrix} = \begin{bmatrix} \bar{U}_j^{-1} & \begin{bmatrix} -\bar{y}_j \\ 1 \end{bmatrix} \bar{\tau}_j^{-1} \end{bmatrix}, \quad \bar{R}_j^{-1} e_j \bar{\tau}_j = \begin{bmatrix} -\bar{y}_j \\ 1 \end{bmatrix}, \tag{7.18}$$

so that with (5.5) we have with $\bar{r}_j \equiv b_j - B_{j-1}\bar{y}_j$ (see (7.14) and (7.15) but now using E'_j and its bound in (3.14) rather than $\hat{E}_j$ and its bound in (7.10))

$$(I - \tilde{S}_j)\begin{bmatrix} \tilde{V}_{j-1}^T \tilde{v}_j \\ 0 \end{bmatrix} = \tilde{S}_j e_j = E'_j \bar{R}_j^{-1} e_j = E'_j \begin{bmatrix} -\bar{y}_j \\ 1 \end{bmatrix} \bar{\tau}_j^{-1}, \ |\|\bar{r}_j\|_2 - |\bar{\tau}_j|| \le j^{\frac{1}{2}} \tilde{\gamma}_n \psi_m(\bar{y}_j). \tag{7.19}$$

Now define a [illegible] (in the terminology of [13, Thm. 7.1])

$$\beta_F(b, A, y) \equiv \beta_F^{A,b}(b, A, y), \quad \text{where} \quad \beta_F^{G,f}(b, A, y) \equiv \frac{\|b - Ay\|_2}{\|f\|_2 + \|G\|_F \|y\|_2}. \tag{7.20}$$

7.1. The theory in [13, Thm. 7.1] assumes a vector norm with its subordinate matrix norm, but with the Frobenius norm in the denominator Rigal and Gaches' theory still works, so this is a possibly new, useful (and usually smaller) construct that is easier to compute than the usual one. A proof similar to that in [13, Thm. 7.1] shows that

$$\beta_F^{G,f}(b, A, y) = \min_{\delta A, \delta b} \{\eta : (A + \delta A)y = b + \delta b, \ \|\delta A\|_F \le \eta \|G\|_F, \ \|\delta b\|_2 \le \eta \|f\|_2\}.$$

Using (7.20) with the bounds in (3.14), (5.6), (7.19), and the definition in (7.14) (see also (5.3)) shows that

$$|\bar{\tau}_j| . \|\tilde{V}_{j-1}^T \tilde{v}_j\|_2 = \left\| (I - \tilde{S}_j)^{-1} E'_j \begin{bmatrix} -\bar{y}_j \\ 1 \end{bmatrix} \right\|_2 \le j^{\frac{1}{2}} \tilde{\gamma}_n \psi_m(\bar{y}_j) / (1 - \|\tilde{S}_j\|_2),$$

$$\beta_F(b_j, B_{j-1}, \bar{y}_j) \|\tilde{V}_{j-1}^T \tilde{v}_j\|_2 \le \frac{j^{\frac{1}{2}} \tilde{\gamma}_n}{1 - \|\tilde{S}_j\|_2}. \tag{7.21}$$

7.2. The product of the loss of orthogonality $\|\tilde{V}_{j-1}^T \tilde{v}_j\|_2$ at step j and the normwise relative backward error $\beta_F(b_j, B_{j-1}, \bar{y}_j)$ of the LVLS problem is bounded by $O(\epsilon)$ until $\|\tilde{S}_j\|_2 \approx 1$, that is, until orthogonality of the $\tilde{v}_1, \ldots, \tilde{v}_j$ is totally lost; see (5.5) and Lemma 5.1.

This is another nice result, as it again reveals how MGS applied to B_m loses orthogonality at [illegible] step—see the related section 5. These bounds on the individual $\|\tilde{V}_{j-1}^T \tilde{v}_j\|_2$ complement the bounds in (5.6), since they are essentially in terms of the individual normwise relative backward errors $\beta_F(b_j, B_{j-1}, \bar{y}_j)$, rather than $\tilde{\kappa}_F(B_j)$. However it is important to note that the LVLS problem considered in this section (see the line after (7.17)) is [illegible] the least squares problem solved for MGS-GMRES, which has the form of (7.6) instead. The two can give very different results in the general case, but in the problems we have solved via MGS-GMRES, these normwise relative backward errors seem to be of similar magnitudes for both problems, and this led to the conjecture in the first place. The similarity in behavior of the two problems is apparently related to the fact that B_m in MGS-GMRES is a Krylov basis. In this case it appears that the normwise relative backward errors of both least squares problems will converge (numerically) as the columns of B_j approach numerical linear dependence; see [17, 18]. Thus we have neither proven nor disproven the conjecture, but we have added weight to it.

8. Numerical behavior of the MGS-GMRES algorithm. We now only consider MGS-GMRES and use k instead of $m-1$ to avoid many indices of the form $m-1$. In section 4 we saw that k steps of the Arnoldi algorithm is in theory just $k+1$ steps of the MGS algorithm applied to $B_{k+1} \equiv [b, AV_k]$ to give $[b, AV_k] = V_{k+1}R_{k+1} = V_{k+1}[e_1\rho, H_{k+1,k}]$. And in practice the only difference in the rounding error analysis is that we apply ordinary MGS to $\bar{B}_{k+1} \equiv [b, fl(A\bar{V}_k)] = [b, A\tilde{V}_k] + [0, \Delta V_k]$; see (4.3). In section 8.1 we combine this fact with the results of section 7.1 to prove backward stability of the MGS-GMRES least squares solution $\bar{y}_k$ [illegible].

In theory MGS-GMRES [illegible] solve $Ax = b$ for nonsingular $n \times n$ A in n steps since we cannot have more than n orthonormal vectors in $\mathbf{R}^n$. But in practice the vectors in MGS-GMRES lose orthogonality, so we need another way to prove that we reach a solution to (1.1). In section 8.2 we will show that the MGS-GMRES algorithm for any problem satisfying (1.1) must, for some k, produce $\bar{V}_{k+1}$ so that numerically b lies in the range of $A\bar{V}_k$, and that MGS-GMRES must give a backward stable solution to (1.1). This k is $\hat{m} - 1$, which is $\le n$; see (6.1).

8.1. Backward stability of the MGS-GMRES least squares solutions. The equivalent of the MGS result (7.13) for MGS-GMRES is obtained by replacing $[b, C]$ by $\bar{B}_{k+1} \equiv [b, A\tilde{V}_k + \Delta V_k]$ throughout (7.13); see Theorem 4.1. Thus the computed $\bar{y}_k$ at step k in MGS-GMRES satisfies (with (4.3) and section 6)

$$\bar{y}_k = \arg\min_y \|\tilde{r}_k(y)\|_2, \quad \tilde{r}_k(y) \equiv b + \Delta b_k(y) - [A\tilde{V}_k + \Delta V_k + \Delta C_k(y)]y \tag{8.1}$$

$$\|[\Delta b_k(y), \Delta C_k(y)]e_j\|_2 \le \tilde{\gamma}_{kn}\|\bar{B}_{k+1}e_j\|_2, \quad j = 1, \ldots, k+1; \quad \|\Delta V_k\|_F \le k^{\frac{1}{2}}\gamma_n\|A\|_F,$$
$$\|\Delta b_k(y)\|_2 \le \tilde{\gamma}_{kn}\|b\|_2, \ \|\Delta V_k + \Delta C_k(y)\|_F \le \tilde{\gamma}_{kn}[\|A\|_F + \|A\tilde{V}_k\|_F] \le \tilde{\gamma}'_{kn}\|A\|_F \text{ if } k < \hat{m}.$$

This has proven the MGS-GMRES least squares solution $\bar{y}_k$ is backward stable for

$$\min_y \|b - A\tilde{V}_k y\|_2 \quad \forall\ k < \hat{m},$$

which is all we need for this least squares problem. But even if $k \ge \hat{m}$, it is straightforward to show that it still gives a backward stable least squares solution.

8.2. Backward stability of MGS-GMRES for $Ax = b$ in (1.1). Even though MGS-GMRES always computes a backward stable solution $\bar{y}_k$ for the least squares problem (7.3), see section 8.1, we still have to prove that $\bar{V}_k\bar{y}_k$ will be a backward stable solution for the original system (1.1) for some k (we take this k to be $\hat{m}-1$ in (6.1)), and this is exceptionally difficult. Usually we want to show we have a backward stable solution when we [illegible] we have a small residual. The analysis here is different in that we will first prove that $\bar{B}_{\hat{m}}$ is numerically rank deficient, see (8.4), but to prove backward stability, we will then have to [illegible] that our residual will be small, amongst other things, and this is far from obvious. Fortunately two little known researchers have studied this arcane area, and we will take ideas from [17]; see Theorem 2.4. To simplify the development and expressions we will absorb all small constants into the $\tilde{\gamma}_{kn}$ terms below.

In (8.1) set $k \equiv \hat{m}-1 \le n$ from (6.1) and write

$$\tilde{r}_k(\bar{y}_k) = b_k - A_k\bar{y}_k, \quad b_k \equiv b + \Delta b_k(\bar{y}_k), \quad A_k \equiv A\tilde{V}_k + \Delta\tilde{V}_k(\bar{y}_k), \tag{8.2}$$
$$\|\Delta b_k(\bar{y}_k)\|_2 \le \tilde{\gamma}_{kn}\|b\|_2, \quad \Delta\tilde{V}_k(y) \equiv \Delta V_k + \Delta C_k(y), \quad \|\Delta\tilde{V}_k(y)\|_F \le \tilde{\gamma}_{kn}\|A\|_F.$$

We need to take advantage of the scaling invariance of MGS in order to obtain our results. Here we need only scale b, so write $D \equiv \mathrm{diag}(\phi, I_k)$ for any scalar $\phi > 0$. Since

$\bar{B}_{k+1} \equiv [b, fl(A\bar{V}_k)] = [b, A\tilde{V}_k + \Delta V_k]$, from (8.2) with the bounds in (8.1) we have

$$
\begin{aligned}
(8.3)\qquad & [b_k\phi, A_k] = \bar{B}_{k+1}D + \Delta B_k D, \quad \Delta B_k \equiv [\Delta b_k(\bar{y}_k), \Delta C_k(\bar{y}_k)],\\
& \|\Delta B_k D\|_F \le \tilde{\gamma}_{kn}\|\bar{B}_{k+1}D\|_F \le \tilde{\gamma}'_{kn}\|[b_k\phi, A_k]\|_F,\\
& \|\bar{B}_{k+1}D\|_F \le (1-\tilde{\gamma}_{kn})^{-1}\|[b_k\phi, A_k]\|_F, \quad \|b_k\|_2 \le (1+\tilde{\gamma}_{kn})\|b\|_2.
\end{aligned}
$$

In addition, $k+1$ is the first integer such that $\kappa_2(\tilde{V}_{k+1}) > 4/3$, so section 6 gives

$$
\begin{aligned}
(8.4)\qquad & \sigma_{min}(\bar{B}_{k+1}D) < 8(k+1)\hat{\gamma}_n\|\bar{B}_{k+1}D\|_F \le \tilde{\gamma}_{kn}\|[b_k\phi, A_k]\|_F \ \forall\ \phi > 0;\\
& \kappa_2(\tilde{V}_k),\ \sigma^{-1}_{min}(\tilde{V}_k),\ \sigma_{max}(\tilde{V}_k) \le 4/3;\\
& \text{and similarly} \quad \|A_k\|_F \le \|A\tilde{V}_k\|_F + \tilde{\gamma}_{kn}\|A\|_F \le (4/3+\tilde{\gamma}_{kn})\|A\|_F.
\end{aligned}
$$

We can combine (8.2), (8.3), and (8.4) to give under the condition in (1.1)

$$
\begin{aligned}
(8.5)\qquad & \sigma_{min}(A_k) \ge \sigma_{min}(A\tilde{V}_k) - \|\Delta\tilde{V}_k(\bar{y}_k)\|_2 \ge 3\sigma_{min}(A)/4 - \tilde{\gamma}_{kn}\|A\|_F > 0,\\
& \sigma_{min}([b_k\phi, A_k]) \le \sigma_{min}(\bar{B}_{k+1}D) + \|\Delta B_k D\|_2 \le \tilde{\gamma}_{kn}\|[b_k\phi, A_k]\|_F.
\end{aligned}
$$

The above allows us to define and analyze an important scalar, see Theorem 2.4,

$$
(8.6)\qquad \delta_k(\phi) \equiv \frac{\sigma_{min}([b_k\phi, A_k])}{\sigma_{min}(A_k)} \le 1,
$$

where from (8.5) A_k has full column rank. Now $\bar{y}_k$ and $\tilde{r}_k(\bar{y}_k)$ solve the linear least squares problem $A_k y \approx b_k$ in (8.2); see (8.1). If $[b_k, A_k]$ does not have full column rank, then $\tilde{r}_k(\bar{y}_k) = 0$, so $\tilde{x}_k \equiv \tilde{V}_k\bar{y}_k$ is a backward stable solution for (1.1), which we wanted to show. Next suppose $[b_k, A_k]$ has full column rank. We will not seek to minimize with respect to ϕ the upper bound on $\|\hat{r}\|_2^2$ in Theorem 2.4, which would be unnecessarily complicated, but instead prove that there exists a value $\hat{\phi}$ of ϕ satisfying (8.7) below, and use this value:

$$
(8.7)\qquad \hat{\phi} > 0, \qquad \sigma^2_{min}(A_k) - \sigma^2_{min}([b_k\hat{\phi}, A_k]) = \sigma^2_{min}(A_k)\|\bar{y}_k\hat{\phi}\|_2^2.
$$

Writing LHS $\equiv \sigma^2_{min}(A_k) - \sigma^2_{min}([b_k\phi, A_k])$, RHS $\equiv \sigma^2_{min}(A_k)\|\bar{y}_k\phi\|_2^2$ we want to find ϕ so that LHS=RHS. But $\phi{=}0 \Rightarrow$ LHS $>$ RHS, while $\phi{=}\|\bar{y}_k\|_2^{-1} \Rightarrow$ LHS $<$ RHS, so from continuity $\exists\ \hat{\phi} \in (0, \|\bar{y}_k\|_2^{-1})$ satisfying (8.7). With (8.6) this shows that

$$
(8.8)\qquad \delta_k(\hat{\phi}) < 1, \quad \hat{\phi}^{-2} = \|\bar{y}_k\|_2^2/[1{-}\delta_k(\hat{\phi})^2], \quad 0 < \hat{\phi} < \|\bar{y}_k\|_2^{-1}.
$$

It then follows from Theorem 2.4 that with (8.5), (8.8), and (8.4),

$$
\begin{aligned}
(8.9)\qquad \|\tilde{r}_k(\bar{y}_k)\|_2^2 &\le \sigma^2_{min}([b_k\hat{\phi}, A_k])(\hat{\phi}^{-2} + \|\bar{y}_k\|_2^2/[1{-}\delta_k(\hat{\phi})^2])\\
&\le \tilde{\gamma}^2_{kn}(\|b_k\hat{\phi}\|_2^2 + \|A_k\|_F^2)2\hat{\phi}^{-2}.
\end{aligned}
$$

But from (8.1) and (8.2) since $\tilde{r}_k(\bar{y}_k) = b_k - A_k\bar{y}_k$, $A_k^T\tilde{r}_k(\bar{y}_k) = 0$, and from (8.8),

$$
\begin{aligned}
\|b_k\hat{\phi}\|_2^2 &= \|\tilde{r}_k(\bar{y}_k)\hat{\phi}\|_2^2 + \|A_k\bar{y}_k\hat{\phi}\|_2^2,\\
&\le 2\tilde{\gamma}^2_{kn}(\|b_k\hat{\phi}\|_2^2 + \|A_k\|_F^2) + \|A_k\|_2^2(1{-}\delta_k(\hat{\phi})^2)\\
&\le 2\tilde{\gamma}^2_{kn}\|b_k\hat{\phi}\|_2^2 + (1 + 2\tilde{\gamma}^2_{kn})\|A_k\|_F^2,\\
(8.10)\qquad \|b_k\hat{\phi}\|_2^2 &\le \frac{1+2\tilde{\gamma}^2_{kn}}{1-2\tilde{\gamma}^2_{kn}}\|A_k\|_F^2.
\end{aligned}
$$

This with (8.4) and (8.5) shows that

$$\delta_k(\hat{\phi}) \equiv \frac{\sigma_{min}([b_k\hat{\phi}, A_k])}{\sigma_{min}(A_k)} \le \frac{\tilde{\gamma}'_{kn}\|[b_k\hat{\phi}, A_k]\|_F}{\sigma_{min}(A)-\tilde{\gamma}_{kn}\|A\|_F} \le \frac{\tilde{\gamma}''_{kn}\|A_k\|_F}{\sigma_{min}(A)-\tilde{\gamma}_{kn}\|A\|_F} \le \frac{\tilde{\gamma}'''_{kn}\|A\|_F}{\sigma_{min}(A)-\tilde{\gamma}_{kn}\|A\|_F} \le \frac{1}{2} \text{ under (1.1)}, \tag{8.11}$$

since this last bound can be rewritten as $\sigma_{min}(A) \ge (2\tilde{\gamma}'''_{kn}+\tilde{\gamma}_{kn})\|A\|_F$, which we see will hold if A satisfies (1.1). This bound on $\delta_k(\hat{\phi})$ shows that $\hat{\phi}^{-2} \le 4\|\bar{y}_k\|_2^2/3$ in (8.8), and using this in (8.9) gives the desired bound

$$\|\tilde{r}_k(\bar{y}_k)\|_2 \le \tilde{\gamma}_{kn}(\|b\|_2^2 + \|A\|_F^2\|\bar{y}_k\|_2^2)^{\frac{1}{2}} \le \tilde{\gamma}_{kn}(\|b\|_2 + \|A\|_F\|\bar{y}_k\|_2). \tag{8.12}$$

But we compute $\bar{x}_j = fl(\bar{V}_j\bar{y}_j)$, not $\tilde{V}_j\bar{y}_j$, so to complete this analysis, we have to show that $\bar{x}_k$ is a backward stable solution for (1.1). Now, see (4.3), $\bar{x}_k = fl(\bar{V}_k\bar{y}_k) = (\bar{V}_k + \Delta V'_k)\bar{y}_k$ with $|\Delta V'_k| \le \gamma_k|\bar{V}_k|$. With $\Delta\tilde{V}_k(y)$ in (8.2) define

$$\Delta A_k \equiv [\Delta\tilde{V}_k(\bar{y}_k) - A(\Delta V'_k + \bar{V}_k - \tilde{V}_k)]\bar{y}_k\|\bar{x}_k\|_2^{-2}\bar{x}_k^T,$$

so that $(A + \Delta A_k)\bar{x}_k = (A\tilde{V}_k + \Delta\tilde{V}_k(\bar{y}_k))\bar{y}_k$, and, see (8.1), (8.2), and (2.2),

$$\|b+\Delta b_k(\bar{y}_k)-(A + \Delta A_k)\bar{x}_k\|_2 = \min_y \|b+\Delta b_k(y)-[A\tilde{V}_k+\Delta\tilde{V}_k(y)]y\|_2, \tag{8.13}$$

$$\|\Delta b_k(\bar{y}_k)\|_2 \le \tilde{\gamma}_{kn}\|b\|_2,$$

$$\|\Delta A_k\|_F \le [\|\Delta\tilde{V}_k(\bar{y}_k)\|_F+\|A(\Delta V'_k+\tilde{V}_k\Delta_k)\|_F]\|\bar{y}_k\|_2/\|\bar{x}_k\|_2, \tag{8.14}$$

where we know from (8.12) that (8.13) is bounded by $\tilde{\gamma}_{kn}(\|b\|_2 + \|A\|_F\|\bar{y}_k\|_2)$. But $\|\Delta V'_k\|_F \le k^{\frac{1}{2}}\gamma_k$, so from (2.2) $\|A(\Delta V'_k+\tilde{V}_k\Delta_k)\|_F \le k^{\frac{1}{2}}\tilde{\gamma}_n\|A\|_2$, and from (8.2) $\|\Delta\tilde{V}_k(\bar{y}_k)\|_F \le \tilde{\gamma}_{kn}\|A\|_F$, so with (2.2) and (8.4)

$$\|\bar{x}_k\|_2 = \|(\bar{V}_k + \Delta V'_k)\bar{y}_k\|_2 \ge \|\bar{V}_k\bar{y}_k\|_2 - \|\Delta V'_k\|_F\|\bar{y}_k\|_2 \ge \|\bar{y}_k\|_2(3/4 - k^{\frac{1}{2}}\gamma_n).$$

Combining these with (8.1) shows that $\|\Delta A_k\|_F \le \tilde{\gamma}_{kn}\|A\|_F$ in (8.14). Summarizing,

$$\tilde{r}_k(\bar{y}_k) = b+\Delta b_k(\bar{y}_k)-(A + \Delta A_k)\bar{x}_k, \quad \|\tilde{r}_k(\bar{y}_k)\|_2 \le \tilde{\gamma}_{kn}(\|b\|_2 + \|A\|_F\|\bar{x}_k\|_2), \tag{8.15}$$
$$\|\Delta b_k(\bar{y}_k)\|_2 \le \tilde{\gamma}_{kn}\|b\|_2, \qquad \|\Delta A_k\|_F \le \tilde{\gamma}_{kn}\|A\|_F.$$

Using the usual approach of combining (8.15) with the definitions

$$\Delta b'_k \equiv -\frac{\|b\|_2}{\|b\|_2 + \|A\|_F\|\bar{x}_k\|_2}\tilde{r}_k(\bar{y}_k), \quad \Delta A'_k \equiv \frac{\|A\|_F\|\bar{x}_k\|_2}{\|b\|_2 + \|A\|_F\|\bar{x}_k\|_2}\frac{\tilde{r}_k(\bar{y}_k)\bar{x}_k^T}{\|\bar{x}_k\|_2^2},$$

shows $(A + \Delta A_k + \Delta A'_k)\bar{x}_k = b + \Delta b_k(\bar{y}_k) + \Delta b'_k$,

$$\|\Delta A_k + \Delta A'_k\|_F \le \tilde{\gamma}_{kn}\|A\|_F, \quad \|\Delta b_k(\bar{y}_k) + \Delta b'_k\|_2 \le \tilde{\gamma}_{kn}\|b\|_2,$$

proving that the MGS-GMRES solution $\bar{x}_k$ is backward stable for (1.1).

9. Comments and conclusions. The form of the restriction in (1.1) suggests that we might be able to ease this restriction somewhat by using $\tilde{\kappa}_F(A)$ as defined in (2.1), instead of $\|A\|_F/\sigma_{min}(A)$ in (1.1). However, $\tilde{\kappa}_F(B_j)$ was useful when we applied MGS to B_j, see, for example, (5.7), while in MGS-GMRES we apply MGS to $[b, AV_{j-1}]$, so it looks like we cannot get an a priori restriction involving $\tilde{\kappa}_F(A)$ this way; see also Remark 4.1. The appendix discusses a possibly superior way of meeting the restriction in (1.1) for difficult problems.

Now to conclude this. Among many other things, we showed that MGS-GMRES

- gives a backward stable least squares solution at every step (section 8.1);
- obtains a backward stable solution to the problem (1.1) (section 8.2);
- and up until this point $\kappa_2(\tilde{V}_m) \leq 4/3$ (section 6).

Thus we can say that the MGS-GMRES method is backward stable for computing the solution x to $Ax = b$ for sufficiently nonsingular A, answering an important open question. Despite loss of orthogonality, it provides an acceptable solution within $n+1$ MGS steps (n steps of MGS-GMRES). The loss of orthogonality is usually inversely proportional to the level of convergence. Complete loss of orthogonality implies a solution exists, and MGS-GMRES necessarily finds this under reasonable restrictions (1.1) (or more practically but less rigorously (1.2)) on the problem. From this we see that the numerical behavior is far better than was often thought. This means we do not have to do anything special to ameliorate the effect of rounding errors—we certainly do not need reorthogonalization—and need only concentrate on finding solutions more quickly, mainly by seeking better preconditioning techniques.

The final proof was seen to require an instance of a more general result on the backward stability of a variant of the MGS algorithm applied to a matrix B in order to solve a linear least squares problem; see section 7.1. In section 5 we showed more precisely than before how orthogonality could be lost in the MGS algorithm, in particular by using the condition number $\tilde{\kappa}_F(B)$ defined in (2.1).

Appendix. Condition numbers. If $\kappa_F(A) \equiv \|A\|_F/\sigma_{min}(A)$, then (2.1) is

$$\tilde{\kappa}_F(A) \equiv \min_{\text{diagonal } D>0} \kappa_F(AD).$$

For $m \times n$ A, if positive diagonal $\widetilde{D}$ is such that in $A\widetilde{D}$ all columns have equal 2-norm, then van der Sluis [21, Thm. 3.5, (b)] showed that $\kappa_F(A\widetilde{D})$ is no more than a factor $\sqrt{n}$ away from its minimum (here $\tilde{\kappa}_F(A)$), and this is the first mention of the condition number $\kappa_F(A)$ (and, at least by implication, of $\tilde{\kappa}_F(A)$) that we have seen so far. He also stated in [22, section 3.9] that if $\|\delta A e_j\| < \|Ae_j\|/\kappa_F(A)$ for $j = 1, \ldots, n \leq m$, then $A+\delta A$ has full rank n. This is easy to see since it ensures that $\|\delta A\|_F < \sigma_{min}(A)$. He also points out that this is in some sense tight, in that if $\|\delta A e_j\| = \|Ae_j\|/\kappa_F(A)$ for $j = 1, \ldots, n \leq m$ is allowed, then for any prescribed value of $\kappa_F(A) \geq \sqrt{n}$ there exist A and δA such that $A + \delta A$ is rank deficient. Since the backward error bounds in this paper were obtained column by column, see Lemma 3.2 and, for example, the column bounds in (8.1), this suggests that the form of the restriction in (1.1) is optimal, even down to the factor $n^2\epsilon$. See also the first paragraph of section 4.

Moreover, instead of solving (1.1) we can solve $(AD)y = b$ for some positive diagonal D and then form $x = Dy$. By taking $D = \widetilde{D}$ above we see from van der Sluis's theory that we can approach the value of $\tilde{\kappa}_F(A)$ with $\kappa_F(A\widetilde{D})$ and perhaps alter a problem with an ill-conditioned A so that it meets the restriction (1.1). This is another justification for using such a $\widetilde{D}$ as a basic simple preconditioner when MGS-GMRES is applied to ill-conditioned problems.

Acknowledgments. The main approach here was to base the analysis on the surprising relationship between MGS and the Householder reduction of an augmented matrix that was discovered by Charles Sheffield and proven and developed by Björck and Paige in [5], and combine this with the elegant result discovered by Giraud and Langou in [10] (responding to a request by Mario Arioli). Once we had made that choice the task was still extremely difficult, and we had to draw on many other works as well—among these the excellent book by Higham [13] facilitated our work greatly.

This paper is the end result of a long term collaboration of its three authors aimed at producing a rounding error analysis of the MGS-GMRES method. And although this is unusual, the second and third authors (alphabetically) would like to thank the first author for carrying this project to such a satisfying conclusion.

Two referees' comments added nicely to the history and precision of the paper.

REFERENCES

[1] M. Arioli and C. Fassino, *Roundoff error analysis of algorithms based on Krylov subspace methods*, BIT, 36 (1996), pp. 189–206.

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

[3] Å. Björck, *Solving linear least squares problems by Gram-Schmidt orthogonalization*, BIT, 7 (1967), pp. 1–21.

[4] Å. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996.

[5] Å. Björck and C. C. Paige, *Loss and recapture of orthogonality in the modified Gram-Schmidt algorithm,* SIAM J. Matrix Anal. Appl., 13 (1992), pp. 176–190.

[6] P. Brown and H. Walker, *GMRES on (nearly) singular systems*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 37–51.

[7] J. Demmel, Y. Hida, W. Kahan, X. S. Li, S. Mukherjee, and E. J. Riedy, *Error bounds from extra precise iterative refinement*, ACM Trans. Math. Software, to appear.

[8] J. Drkošová, A. Greenbaum, M. Rozložník, and Z. Strakoš, *Numerical stability of GMRES*, BIT, 35 (1995), pp. 309–330.

[9] R. W. Freund and N. M. Nachtigal, *QMR: A quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.

[10] L. Giraud and J. Langou, *When modified Gram-Schmidt generates a well-conditioned set of vectors*, IMA J. Numer. Anal., 22 (2002), pp. 521–528.

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

[12] A. Greenbaum, M. Rozložník, and Z. Strakoš, *Numerical behavior of the modified Gram-Schmidt GMRES implementation*, BIT, 37 (1997), pp. 706–719.

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

[14] J. Langou, *Iterative Methods for Solving Linear Systems with Multiple Right-Hand Sides*, Ph.D. thesis, Institut Nationales des Sciences Appliquées de Toulouse, Toulouse, France, 2003.

[15] J. Liesen, M. Rozložník, and Z. Strakoš, *Least squares residuals and minimal residual methods*, SIAM J. Sci. Comput., 23 (2002), pp. 1503–1525.

[16] C. C. Paige and M. A. Saunders, *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Software, 8 (1982), pp. 43–71.

[17] C. C. Paige and Z. Strakoš, *Bounds for the least squares distance using scaled total least squares*, Numer. Math., 91 (2002), pp. 93–115.

[18] C. C. Paige and Z. Strakoš, *Residual and backward error bounds in minimum residual Krylov subspace methods*, SIAM J. Sci. Comput., 23 (2002), pp. 1898–1923.

[19] M. Rozložník, *Numerical Stability of the GMRES Method*, Ph.D. thesis, Institute of Computer Science, Academy of Sciences, Prague, Czech Republic, 1997.

[20] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[21] A. van der Sluis, *Condition numbers and equilibration matrices*, Numer. Math., 14 (1969), pp. 14–23.

[22] A. van der Sluis, *Stability of the solutions of linear least squares problems*, Numer. Math., 23 (1975), pp. 241–254.

[23] L. Smoch, *Some results about GMRES in the singular case*, Numer. Algorithms, 22 (1999), pp. 193–212.

[24] H. A. van der Vorst, *Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 631–644.

[25] H. F. Walker, *Implementation of the GMRES method*, J. Comput. Phys., 53 (1989), pp. 311–320.

[26] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 1, pp. 285–300

DIRECT EIGENVALUE REORDERING IN A PRODUCT OF MATRICES IN PERIODIC SCHUR FORM*

ROBERT GRANAT† AND BO KÅGSTRÖM†

Abstract. A direct method for eigenvalue reordering in a product of a K-periodic matrix sequence in periodic or extended periodic real Schur form is presented and analyzed. Each reordering of two adjacent sequences of diagonal blocks is performed tentatively to guarantee backward stability and involves solving a K-periodic Sylvester equation (PSE) and constructing a K-periodic sequence of orthogonal transformation matrices. An error analysis of the direct reordering method is presented, and results from computational experiments confirm the stability and accuracy of the method for well-conditioned as well as ill-conditioned problems. These include matrix sequences with fixed and time-varying dimensions, and sequences of small and large periodicity.

Key words. product of K-periodic matrix sequence, extended periodic real Schur form, eigenvalue reordering, K-periodic Sylvester equation, periodic eigenvalue problem

AMS subject classifications. 65F15, 15A18, 93B60

DOI. 10.1137/05062490X

1. Introduction. Given a K-periodic real matrix sequence, $A_0,\ A_1, \ldots, A_{K-1}$ with $A_{i+K} = A_i$, the [illegible] (PRSF) is defined as follows [5, 13]: given the real matrix sequence $A_k \in R^{n\times n}$, for $k = 0, 1, \ldots, K-1$, there exists an orthogonal matrix sequence $Z_k \in R^{n\times n}$ such that the real sequence

$$Z_{k+1}^T A_k Z_k = T_k,\ k = 0, 1, \ldots, K-1, \tag{1.1}$$

with $Z_K = Z_0$, consists of $K-1$ upper triangular matrices and one upper quasi-triangular matrix. The products of conforming 1×1 and 2×2 diagonal blocks of the matrix sequence T_k contain the real and complex conjugate pairs of eigenvalues of the matrix product $A_{K-1}\cdots A_1 A_0$. Similar to the standard case ($K = 1$; e.g., see [10, 25]), the periodic real Schur form is computed by means of a reduction to periodic Hessenberg form followed by applying a periodic QR-algorithm to the resulting sequence [5, 13]. The PRSF is an important tool in several applications, including solving periodic Sylvester-type and Riccati matrix equations [13, 22, 27, 30]. The quasi-triangular matrix in the PRSF can occur anywhere in the sequence but is usually chosen to be T_0 or T_{K-1}.

The [illegible] (EPRSF) generalizes PRSF to the case when the dimensions of the matrices are time-variant [28]: given the real matrix sequence $A_k \in R^{n_{k+1}\times n_k},\ k = 0, 1, \ldots, K-1$, with $n_K = n_0$, there exists an orthogonal matrix sequence $Z_k \in R^{n_k\times n_k},\ k = 0, 1, \ldots, K-1$, such that the real sequence

$$Z_{k+1}^T A_k Z_k = T_k \equiv \begin{bmatrix} T_{11}^{(k)} & T_{12}^{(k)} \\ 0 & T_{22}^{(k)} \end{bmatrix} \in R^{n_{k+1}\times n_k}, \tag{1.2}$$

*Received by the editors February 21, 2005; accepted for publication (in revised form) by P. Van Dooren January 12, 2006; published electronically April 7, 2006. This research was conducted using the resources of the High Performance Computing Center North (HPC2N). Financial support was provided by the Swedish Research Council under grant VR 621-2001-3284 and by the Swedish Foundation for Strategic Research under the frame program grant A3 02:128.

http://www.siam.org/journals/simax/28-1/62490.html

†Department of Computing Science and HPC2N, Umeå University, SE-901 87 Umeå, Sweden (granat@cs.umu.se, bokg@cs.umu.se).

for $k = 0, 1, \ldots, K-1$, with $Z_K = Z_0$, is block upper triangular and $T_{11}^{(k)} \in R^{\min_k(n_k) \times \min_k(n_k)}$, $T_{22}^{(k)} \in R^{(n_{k+1} - \min_k(n_k)) \times (n_k - \min_k(n_k))}$. Moreover, the subsequence $T_{11}^{(k)}$, $k = 0, 1, \ldots, K-1$, is in PRSF (1.1) with eigenvalues called the [illegible] of the sequence A_k, and the matrices in the subsequence $T_{22}^{(k)}$, $k = 0, 1, \ldots, K-1$, are upper trapezoidal. For EPRSF, the quasi-triangular matrix can occur at any position in the sequence T_k. However, to simplify the reduction to extended periodic Hessenberg form it is normally placed at position j, where $n_{j+1} = \min_k(n_k)$, i.e., in the matrix T_j which has the smallest row dimension in the sequence [28]. For T_j, $j \in [0, K-1]$, to have a trapezoidal block $T_{22}^{(j)}$, it must hold that $n_j, n_{j+1} > \min_k(n_k)$. The EPRSF is motivated by the increasing interest in [illegible] of the form

$$\begin{aligned} x_{k+1} &= A_k x_k + B_k u_k, \\ y_k &= C_k x_k + D_k u_k, \end{aligned} \tag{1.3}$$

where the matrices $A_k \in R^{n_{k+1} \times n_k}$, $B_k \in R^{n_{k+1} \times m}$, $C_k \in R^{r \times n_k}$, and $D_k \in R^{r \times m}$ are periodic with periodicity $K \geq 1$. The [illegible] of the system (1.3) is defined as the $n_j \times n_i$ matrix $\Phi_A(j, i) = A_{j-1} A_{j-2} \ldots A_i$, where $\Phi_A(i, i) = I_{n_i}$. The state transition matrix over one whole period $\Phi_A(j+K, j) \in R^{n_j \times n_j}$ is called the [illegible] of (1.3) at time j, and its eigenvalues are called the [illegible] at time j. All t nonzero together with $(\min_k(n_k) - t)$ zero characteristic multipliers belong to the set of core characteristic values. One important issue is how to reorder the eigenvalues of the monodromy matrix without evaluating the corresponding product. Evaluating the product is costly and may lead to a significant loss of accuracy [5], especially when computing eigenvalues of small magnitude.

Direct eigenvalue reordering in the real Schur form was investigated in [2, 8, 7] and in the [illegible] of a regular matrix pencil $A - \lambda B$ in [16, 18]. Iterative QR-based reordering methods have also been proposed [23, 26], but they may fail to converge (e.g., see [16, 18]). Reordering of eigenvalues in PRSF and related problems have also been considered; see, e.g., [5], where the approach is based on applying Givens rotations on explicitly formed products of small (2×2, 3×3, or 4×4) matrix sequences, and [6] for a discussion on swapping 1×1 blocks by propagating orthogonal transformations through 2×2 sequences. In this paper, we present a direct swapping algorithm for performing eigenvalue reordering in a product of a K-periodic matrix sequence in (E)PRSF for $K \geq 2$ without evaluating any part of the matrix product. Our direct algorithm relies on orthogonal transformations only and extends earlier work on direct eigenvalue reordering of matrices to products of matrices [11, 19].

The rest of this paper is organized as follows. In section 2, we settle some important notation and definitions. In section 3, we discuss reordering of two diagonal blocks (leaving the eigenvalues invariant) by cyclic orthogonal transformations, and in section 4, we present our direct periodic reordering algorithm. Next, we discuss the numerical solution of the associated periodic Sylvester equation (PSE) in section 5. An error analysis of the direct periodic swapping algorithm is presented in section 6. Some numerical examples are presented and discussed in section 7, and finally, we outline some future work in section 8.

2. Notation and definitions. We introduce some notation to simplify the presentation that follows. Let I_n denote the identity matrix of order n. Let M^+ denote the pseudoinverse (see, e.g., [10]) of a matrix M. Let $\sigma(M)$ and $\lambda(M)$ denote the sets of the singular values and the eigenvalues of the matrix M, respectively. Let

$A \otimes B$ denote the Kronecker product of two matrices, defined as the matrix with its (i,j)-block element as $a_{ij}B$. Let $\mathrm{vec}(A)$ denote a vector representation of an $m \times n$ matrix A with the columns of A stacked on top of each other in the order $1, 2, \ldots, n$. Let $\|A\|_F$ denote the Frobenius matrix norm defined as $\|A\|_F = \sqrt{\mathrm{trace}(A^T A)}$. We define the [illegible] $\oplus$ such that $a \oplus b = (a+b) \bmod K$, where K denotes the periodicity. We use the product operator $\prod_{k=i}^{j} b_k$ to denote a product $b_i b_{i-1} \cdots b_{j+1} b_j$ of scalars, with the convention that $\prod_{k=i}^{j} b_k = 1$ for $i < j$.

Each K-periodic matrix sequence A_k is associated with a [illegible] $\bar{A} = (A_{K-1}, A_{K-2}, \ldots, A_1, A_0)$ [4]. The vector tuple $\bar{u} = (u_{K-1}, u_{K-2}, \ldots, u_1, u_0)$, with $u_k \neq 0$, is called a [illegible] of the tuple $\bar{A}$ corresponding to the eigenvalue λ if there exist scalars α_k, possibly complex, such that the relations

$$\begin{aligned} &A_k u_k = \alpha_k u_{k\oplus 1}, \ k = 0, 1, \ldots, K-1, \\ &\lambda := \prod_{k=K-1}^{0} \alpha_k \end{aligned} \tag{2.1}$$

hold with $u_K = u_0$. A [illegible] $\bar{v}$ of the tuple $\bar{A}$ corresponding to λ is defined similarly:

$$\begin{aligned} &v_{k\oplus 1}^H A_k = \beta_k v_k^H, \ k = 0, 1, \ldots, K-1, \\ &\lambda := \prod_{k=K-1}^{0} \beta_k, \end{aligned} \tag{2.2}$$

where $v_k \neq 0$, and β_k are (possibly complex) scalars for $k = 0, 1, \ldots, K-1$.

Without loss of generality, we assume that $p < \min_k (n_k)$ is specified such that no 2×2 block corresponding to a complex conjugate pair of eigenvalues is positioned at rows (and columns) p and $p+1$ of $\Phi_T(K, 0)$. Given such a p and with Z_k and T_k from (1.2), the leading p columns of each Z_k span an invariant subspace for $\Phi_T(K+k, k)$ for $k = 0, 1, \ldots, K-1$. As a whole, the space spanned by the first p columns of each matrix in the matrix tuple $\bar{Z}$ is called a [illegible] of the tuple $\bar{A}$ corresponding to the p eigenvalues located in the upper-leftmost part of $\Phi_T(K, 0)$. In general, $\Phi_T(K, 0)_{ij}$ denotes the (i,j) block of the matrix product $\Phi_T(K, 0)$.

3. Reordering diagonal blocks in a product of matrices in EPRSF by orthogonal transformations. Consider the K-periodic (or K-cyclic) matrix sequences $A_k \in R^{n_{k\oplus 1} \times n_k}$, $T_k \in R^{n_{k\oplus 1} \times n_k}$, and $Z_k \in R^{n_k \times n_k}$, $k = 0, 1, \ldots, K-1$, such that A_k is general, T_k is in EPRSF, and Z_k is the corresponding orthogonal transformation, as in (1.2). The eigenvalues of the product $\Phi_T(K, 0) = T_{K-1} T_{K-2} \ldots T_1 T_0 \in R^{n_0 \times n_0}$ are contained in the diagonal blocks of size 1×1 (real) and 2×2 (complex conjugate pairs) of $\Phi_T(K, 0)$.

Assume that each T_k, $k = 0, 1, \ldots, K-1$, is partitioned as

$$T_k = \left[\begin{array}{c|c|c|c} T_{11}^{(k)} & \star & \star & \star \\ \hline 0 & T_{22}^{(k)} & \star & \star \\ 0 & 0 & T_{33}^{(k)} & \star \\ \hline 0 & 0 & 0 & T_{44}^{(k)} \end{array}\right], \tag{3.1}$$

where $T_{11}^{(k)} \in R^{p_1 \times p_1}$, $T_{22}^{(k)} \in R^{p_2 \times p_2}$, $T_{33}^{(k)} \in R^{p_3 \times p_3}$, $T_{44}^{(k)} \in R^{(n_{k\oplus 1}-p) \times (n_k - p)}$, $k = 0, 1, \ldots, K-1$, and $p = p_1 + p_2 + p_3$. Assume there exists a K [illegible]

$Q_k,\ k = 0, 1, \ldots, K-1$, such that the

$$Q_{k\oplus 1}^T \begin{bmatrix} T_{22}^{(k)} & \star \\ 0 & T_{33}^{(k)} \end{bmatrix} Q_k = \begin{bmatrix} \hat{T}_{22}^{(k)} & \star \\ 0 & \hat{T}_{33}^{(k)} \end{bmatrix} \tag{3.2}$$

results in $\lambda(\Phi_{\hat{T}}(K,0)_{22}) = \lambda(\Phi_T(K,0)_{33}),\ \lambda(\Phi_{\hat{T}}(K,0)_{33}) = \lambda(\Phi_T(K,0)_{22})$. Then the reordered EPRSF of the sequence A_k is the sequence $\hat{T}_k$, where

$$\begin{aligned} \hat{T}_k &= \underbrace{\left[\begin{array}{c|c|c} I_{p_1} & 0 & 0 \\ \hline 0 & Q_{k\oplus 1}^T & 0 \\ \hline 0 & 0 & I_{p_4} \end{array}\right]}_{\hat{Q}_{k\oplus 1}^T} \left[\begin{array}{c|c|c|c} T_{11}^{(k)} & \star & \star & \star \\ \hline 0 & T_{22}^{(k)} & \star & \star \\ \hline 0 & 0 & T_{33}^{(k)} & \star \\ \hline 0 & 0 & 0 & T_{44}^{(k)} \end{array}\right] \underbrace{\left[\begin{array}{c|c|c} I_{p_1} & 0 & 0 \\ \hline 0 & Q_k & 0 \\ \hline 0 & 0 & I_{p_4} \end{array}\right]}_{\hat{Q}_k} \\ &= \hat{Q}_{k\oplus 1}^T T_k \hat{Q}_k = \hat{Q}_{k\oplus 1}^T Z_{k\oplus 1}^T A_k Z_k \hat{Q}_k = \hat{Z}_{k\oplus 1}^T A_k \hat{Z}_k, \end{aligned} \tag{3.3}$$

with the associated K-cyclic orthogonal sequence $\hat{Z}_k = Z_k \hat{Q}_k,\ k = 0, 1, \ldots, K-1$. The first $p_1 + p_3$ columns of $\hat{Z}_0$ span an orthonormal basis for the invariant subspace of $\Phi_A(K,0)$ associated with the first $p_1 + p_3$ eigenvalues in the upper left part of the product $\Phi_{\hat{T}}(K,0)$. In addition, the first $p_1 + p_3$ columns of each transformation matrix $\hat{Z}_k$ in the tuple $(\hat{Z}_{K-1}, \hat{Z}_{K-2}, \ldots, \hat{Z}_1, \hat{Z}_0)$ span an orthonormal basis for the periodic invariant subspace of the tuple $\bar{A}$ associated with the same $p_1 + p_3$ eigenvalues in $\Phi_{\hat{T}}(K,0)$.

4. A direct algorithm for periodic diagonal block reordering. In this section, we focus on the K-cyclic swapping in (3.3). Without loss of generality, we assume that T_k in (3.1) is square, i.e., the sequence T_k is in PRSF, and partitioned as

$$T_k = \begin{bmatrix} T_{11}^{(k)} & T_{12}^{(k)} \\ 0 & T_{22}^{(k)} \end{bmatrix},\ k = 0, 1, \ldots, K-1, \tag{4.1}$$

and that we want to swap the blocks $T_{11}^{(k)} \in R^{p_1 \times p_1}$ and $T_{22}^{(k)} \in R^{p_2 \times p_2}$. Throughout the paper we assume that $\Phi_T(K,0)_{11}$ and $\Phi_T(K,0)_{22}$ are of size 2×2 or 1×1 and have no eigenvalues in common; otherwise, the diagonal blocks need not be swapped. Define the K-cyclic matrix sequence $\boldsymbol{X}_k$ as

$$\boldsymbol{X}_k \equiv \begin{bmatrix} I_{p_1} & X_k \\ 0 & I_{p_2} \end{bmatrix}, \tag{4.2}$$

where $X_k \in R^{p_1 \times p_2},\ k = 0, 1, \ldots, K-1$. The key observation is that the cyclic transformation

$$\boldsymbol{X}_{k\oplus 1}^{-1} \begin{bmatrix} T_{11}^{(k)} & T_{12}^{(k)} \\ 0 & T_{22}^{(k)} \end{bmatrix} \boldsymbol{X}_k = \begin{bmatrix} T_{11}^{(k)} & T_{12}^{(k)} + T_{11}^{(k)} X_k - X_{k\oplus 1} T_{22}^{(k)} \\ 0 & T_{22}^{(k)} \end{bmatrix} \tag{4.3}$$

block-diagonalizes $T_k,\ k = 0, 1, \ldots, K-1$, if and only if the sequence X_k satisfies the (PSE)

$$T_{11}^{(k)} X_k - X_{k\oplus 1} T_{22}^{(k)} = -T_{12}^{(k)},\ k = 0, 1, \ldots, K-1. \tag{4.4}$$

Replacing I_{p_2} in $\boldsymbol{X}_0$ (4.2) by a $p_2 \times p_2$ zero block results in a spectral projector (e.g., see [25]) associated with the matrix product $\Phi_T(K,0)$ that projects onto the spectrum of $\Phi_T(K,0)_{11}$. We refer to the matrix X_0 as the [illegible] for the periodic reordering of the product $\Phi_T(K,0)$.

The similarity transformation

$$S_0^{-1}T_{K-1}S_{K-1}S_{K-1}^{-1}T_{K-2}S_{K-2}\dots S_2^{-1}T_1S_1S_1^{-1}T_0S_0$$

$$= \begin{bmatrix} T_{22}^{(K-1)} & 0 \\ 0 & T_{11}^{(K-1)} \end{bmatrix} \cdots \begin{bmatrix} T_{22}^{(1)} & 0 \\ 0 & T_{11}^{(1)} \end{bmatrix} \begin{bmatrix} T_{22}^{(0)} & 0 \\ 0 & T_{11}^{(0)} \end{bmatrix}$$

performs the wanted swapping of the diagonal blocks by the nonorthogonal sequence

$$S_k = \boldsymbol{X}_k \begin{bmatrix} 0 & I_{p_1} \\ I_{p_2} & 0 \end{bmatrix} = \begin{bmatrix} X_k & I_{p_1} \\ I_{p_2} & 0 \end{bmatrix}, \ k = 0, 1, \dots, K-1.$$

Since the first p_2 columns of each S_k are linearly independent there exist orthogonal matrices Q_k of order $p_1 + p_2$ such that

$$D_k \equiv \begin{bmatrix} X_k \\ I_{p_2} \end{bmatrix} = Q_k \begin{bmatrix} R_k \\ 0 \end{bmatrix}, \tag{4.5}$$

where R_k of size $p_2 \times p_2$ is upper triangular and nonsingular, $k = 0, 1, \dots, K-1$. By partitioning Q_k conformally with S_k, we observe that

$$Q_k^T S_k = \begin{bmatrix} R_k & Q_{11}^{(k)^T} \\ 0 & Q_{12}^{(k)^T} \end{bmatrix}, \qquad S_k^{-1}Q_k = \begin{bmatrix} R_k^{-1} & -R_k^{-1}Q_{11}^{(k)^T}Q_{12}^{(k)^{-T}} \\ 0 & Q_{12}^{(k)^{-T}} \end{bmatrix}.$$

An orthonormal similarity transformation of $\Phi_T(K,0)$ can now be written as

$$Q_0^T(T_{K-1}T_{K-2}\dots T_1T_0)Q_0 = Q_0^T T_{K-1}Q_{K-1}Q_{K-1}^T T_{K-2}Q_{K-2}\dots Q_2^T T_1 Q_1 Q_1^T T_0 Q_0$$

$$= Q_0^T S_0 \begin{bmatrix} T_{22}^{(K-1)} & 0 \\ 0 & T_{11}^{(K-1)} \end{bmatrix} S_{K-1}^{-1}Q_{K-1}Q_{K-1}^T S_{K-1} \begin{bmatrix} T_{22}^{(K-2)} & 0 \\ 0 & T_{11}^{(K-2)} \end{bmatrix} S_{K-2}^{-1}Q_{K-2}$$

$$\dots Q_2^T S_2 \begin{bmatrix} T_{22}^{(1)} & 0 \\ 0 & T_{11}^{(1)} \end{bmatrix} S_1^{-1}Q_1Q_1^T S_1 \begin{bmatrix} T_{22}^{(0)} & 0 \\ 0 & T_{11}^{(0)} \end{bmatrix} S_0^{-1}Q_0 = \hat{T}_{K-1}\hat{T}_{K-2}\dots\hat{T}_1\hat{T}_0,$$

where

$$\hat{T}_k = \begin{bmatrix} \hat{T}_{11}^{(k)} & \hat{T}_{12}^{(k)} \\ 0 & \hat{T}_{22}^{(k)} \end{bmatrix}$$

and

$$\begin{cases} \hat{T}_{11}^{(k)} &= R_{k\oplus 1}T_{22}^{(k)}R_k^{-1}, \\ \hat{T}_{22}^{(k)} &= Q_{12}^{(k\oplus 1)^T}T_{11}^{(k)}Q_{12}^{(k)^{-T}}, \\ \hat{T}_{12}^{(k)} &= -R_{k\oplus 1}T_{22}^{(k)}R_k^{-1}Q_{11}^{(k)^T}Q_{12}^{(k)^{-T}} + Q_{11}^{(k\oplus 1)^T}T_{11}^{(k)}Q_{12}^{(k)^{-T}} \end{cases} \tag{4.6}$$

for $k = 0, 1, \ldots, K-1$. Thus, the orthogonal sequence Q_k from (4.5) performs the required reordering of the diagonal blocks. Observe that the sequences $\hat{T}_{11}^{(k)}$ and $\hat{T}_{22}^{(k)}$ in (4.6) may not be in PRSF and might have to be further transformed after periodic reordering by additional orthogonal transformations to get the sequence $\hat{T}_k$ in PRSF.

We summarize our direct algorithm for periodic eigenvalue reordering as follows:

1. Solve for the sequence $X_k,\ k = 0, 1, \ldots, K-1$, in the PSE

$$T_{11}^{(k)} X_k - X_{k\oplus 1} T_{22}^{(k)} = -T_{12}^{(k)},\ k = 0, 1, \ldots, K-1.$$

2. Compute K orthogonal matrices Q_k such that

$$\begin{bmatrix} X_k \\ I_{p_2} \end{bmatrix} = Q_k \begin{bmatrix} R_k \\ 0 \end{bmatrix},\ k = 0, 1, \ldots, K-1.$$

3. Perform reordering by the cyclic transformations

$$\hat{T}_k = Q_{k\oplus 1}^T T_k Q_k,\ k = 0, 1, \ldots, K-1. \tag{4.7}$$

4. Restore the subsequences $\hat{T}_{11}^{(k)}$ and $\hat{T}_{22}^{(k)}$ to PRSF using K-cyclic orthogonal transformations.

Step 4 is conducted by computing PRSFs of the two K-periodic subsequences $\hat{T}_{11}^{(k)}$ and $\hat{T}_{22}^{(k)}$. Care must be taken to assure that each of the two quasi-triangular matrices in the PRSFs appear in the same position of the $\hat{T}_k$ sequence, say $\hat{T}_i$. However, for a K-periodic 2×2 sequence it is sufficient to compute a periodic Hessenberg form [5] specifying the position of the 2×2 Hessenberg matrix, given that the complex conjugate pair has not collapsed into two real eigenvalues because of round-off errors.

In the presence of rounding errors, the most critical step in the reordering process is to solve the PSE. In analogy to eigenvalue swapping in the real (generalized) Schur form, a small sep-function (defined in equation (5.3)) may ruin backward stability and thus forces us to perform the swapping tentatively to guarantee backward stability [2, 16, 18]. See also Kressner [19] for a brief discussion on direct swapping methods for PRSF.

The direct algorithm extends directly to EPRSF by considering reordering of the core characteristic values (see section 2) of the sequence T_k.

5. The periodic Sylvester equation. In analogy with solving the standard Sylvester equation (e.g., see [3]), we construct a matrix representation Z_{PSE} of the periodic Sylvester operator defined by the PSE (4.4) in terms of Kronecker products, where

(5.1)

$$Z_{\mathrm{PSE}} = \begin{bmatrix} -T_{22}^{(K-1)^T} \otimes I_{p_1} & & & & I_{p_2} \otimes T_{11}^{(K-1)} \\ I_{p_2} \otimes T_{11}^{(0)} & -T_{22}^{(0)^T} \otimes I_{p_1} & & & \\ & \ddots & \ddots & & \\ & & & I_{p_2} \otimes T_{11}^{(K-2)} & -T_{22}^{(K-2)^T} \otimes I_{p_1} \end{bmatrix}.$$

Only the nonzero blocks of Z_{PSE} are displayed explicitly in (5.1). Then we solve the resulting linear system of equations $Z_{\mathrm{PSE}} x = c$, with x and c as stacked vector

representations of the matrix sequences X_k, for $k = 0, 1, \ldots, K-1$, and $-T_{12}^{(k)}$, $k = K-1, 0, 1, \ldots, K-2$, respectively:

$$x = \begin{bmatrix} \mathrm{vec}(X_0) \\ \mathrm{vec}(X_1) \\ \cdots \\ \mathrm{vec}(X_{K-1}) \end{bmatrix}, \qquad c = \begin{bmatrix} \mathrm{vec}(-T_{12}^{(K-1)}) \\ \mathrm{vec}(-T_{12}^{(0)}) \\ \cdots \\ \mathrm{vec}(-T_{12}^{(K-2)}) \end{bmatrix}. \tag{5.2}$$

To exploit the structure of the matrix Z_{PSE}, Gaussian elimination with partial pivoting (GEPP) is used at the cost of $O(K(p_1^2 p_2 + p_1 p_2^2))$ flops, possibly combined with fixed precision iterative refinement for improved accuracy on badly scaled problems. By storing only the block main diagonal, the block subdiagonal, and the rightmost block column vector, the storage requirement for Z_{PSE} can be kept at $3Kp_1^2p_2^2$.

Linear systems with this kind of sparsity structure, [illegible] (BABD) linear systems, were studied extensively in [9, 32]. It appears that there exists no general-purpose numerically stable method designed specifically for BABD systems, and it is not clear under what conditions (if any) GEPP is stable for solving PSEs of the form (5.1). As an alternative, it is possible to consider QR-factorizations [9] for solving (5.1). However, by introducing explicit stability tests (see section 7) the resulting periodic reordering algorithm is conditionally backward stable by rejecting swaps that appear unstable by some given criterion.

One could employ Gaussian elimination with complete pivoting (GECP) to solve this linear system (see, e.g., LAPACK's `DTGSYL` [18]), but that would make it difficult, if not impossible, to exploit the sparsity structure of the problem. The complete pivoting process causes fill-in elements, requires explicit storage of the whole matrix Z_{PSE}, and increases the number of flops to $O((Kp_1p_2)^3)$.

Also in analogy with the standard Sylvester equation (e.g., see [14, 17]), the conditioning of the PSE is related to the sep-function

$$\begin{aligned} \mathrm{sep}[\mathrm{PSE}] &= \inf_{\|x\|_2=1} \|Z_{\mathrm{PSE}} x\|_2 = \|Z_{\mathrm{PSE}}^{-1}\|_2^{-1} = \sigma_{\min}(Z_{\mathrm{PSE}}) \\ &= \inf_{(\sum_{k=0}^{K-1} \|X_k\|_F^2)^{1/2}=1} \left(\sum_{k=0}^{K-1} \|T_{11}^{(k)} X_k - X_{k\oplus 1} T_{22}^{(k)}\|_F^2 \right)^{1/2}. \end{aligned} \tag{5.3}$$

The quantity sep[PSE] can be estimated at the cost of solving a few PSEs by exploiting the estimation technique for the 1-norm of the inverse of a matrix [12, 14, 17, 18].

6. Error analysis. In this section, we present an error analysis of the direct reordering method presented in section 4, where we extend the analysis from [2, 16] to the periodic case. For $K = 1$ we also get sharper error bounds compared to [2].

6.1. Perturbation of individual matrices under periodic reordering. If Householder reflections are used to compute the orthogonal sequence $\tilde{Q}_k$, $k = 0, 1, \ldots, K-1$, each matrix $\tilde{Q}_k$ is orthogonal up to machine precision [31], and the stability of the direct reordering method is mainly affected by the conditioning and accuracy of the solution to the associated PSE.

Without loss of generality, we assume that $p_1 = p_2 = 2$. Let $\tilde{X}_k$ be the computed solution sequence to the PSE (4.4), where $\tilde{X}_k = X_k + \Delta X_k$, X_k is the exact and unique solution sequence and ΔX_k is the corresponding error matrix for $k = 0, 1, \ldots, K-1$. We let

$$Y_k \equiv T_{11}^{(k)} \tilde{X}_k - \tilde{X}_{k\oplus 1} T_{22}^{(k)} + T_{12}^{(k)} = T_{11}^{(k)} \Delta X_k - \Delta X_{k\oplus 1} T_{22}^{(k)} \tag{6.1}$$

denote the residual sequence associated with the computed PSE solution sequence.

Under mild conditions (such as $\|D_k^+\|_2\|\Delta X_k\|_F < 1$, where D_k is defined in (4.5)) the K QR-factorizations of $(\tilde{X}_k,\ I)^T$ can be written as

$$\begin{bmatrix} X_k + \Delta X_k \\ I \end{bmatrix} = D_k + \begin{bmatrix} \Delta X_k \\ 0 \end{bmatrix} = \tilde{Q}_k \begin{bmatrix} \tilde{R}_k \\ 0 \end{bmatrix} = (Q_k + \Delta Q_k) \begin{bmatrix} R_k + \Delta R_k \\ 0 \end{bmatrix},$$

where ΔQ_k and ΔR_k are perturbations of the orthogonal matrices Q_k and the triangular matrices R_k, and $\tilde{Q}_k = Q_k + \Delta Q_k$ is orthogonal [24]. Here $\|\Delta Q_k\|_F$ and $\|\Delta R_k\|_F$ are essentially bounded by $\|D_k^+\|_2\|\Delta X_k\|_F$, $k = 0, 1, \ldots, K-1$ [24, 2]. We do not assume anything about the structure of these perturbation matrices.

Given the computed sequences $\tilde{X}_k$ and $\tilde{Q}_k$, the following theorem shows how the errors in these quantities propagate to the results of the direct method for reordering two adjacent sequences of diagonal blocks in the periodic Schur form.

THEOREM 6.1. [illegible] $\tilde{X}_k = X_k + \Delta X_k$ [illegible] $\Delta X_k \neq 0$ [illegible] $\tilde{Q}_k$ [illegible] Y_k (6.1) [illegible] $k = 0, 1, \ldots, K-1$ [illegible] $\tilde{Q}_k$ [illegible] $(1,1)$ [illegible] $(2,2)$ [illegible] T_k (4.1) [illegible]

$$\tilde{T}_k \equiv \tilde{Q}_{k\oplus 1}^T \begin{bmatrix} T_{11}^{(k)} & T_{12}^{(k)} \\ 0 & T_{22}^{(k)} \end{bmatrix} \tilde{Q}_k = \hat{T}_k + E_k, \tag{6.2}$$

[illegible]

$$\hat{T}_k = \begin{bmatrix} \hat{T}_{11}^{(k)} & \hat{T}_{12}^{(k)} \\ 0 & \hat{T}_{22}^{(k)} \end{bmatrix}, \qquad E_k = \begin{bmatrix} E_{11}^{(k)} & E_{12}^{(k)} \\ E_{21}^{(k)} & E_{22}^{(k)} \end{bmatrix} \tag{6.3}$$

[illegible] $k = 0, 1, \ldots, K-1$ [illegible] E_k [illegible]

$$\|E_{11}^{(k)}\|_2 \leq \frac{\sigma_{\max}(X_{k\oplus 1})}{(1+\sigma_{\max}^2(X_{k\oplus 1}))^{1/2}} \cdot \frac{1}{(1+\sigma_{\min}^2(X_k))^{1/2}} \|Y_k\|_F + 2\|\hat{T}_{11}^{(k)}\|_2(\|D_k^+\|_2\|\Delta X_k\|_F + \|D_{k\oplus 1}^+\|_2\|\Delta X_{k\oplus 1}\|_F), \tag{6.4}$$

$$\|E_{21}^{(k)}\|_2 \leq \frac{1}{(1+\sigma_{\min}^2(X_{k\oplus 1}))^{1/2}} \cdot \frac{1}{(1+\sigma_{\min}^2(X_k))^{1/2}} \|Y_k\|_F, \tag{6.5}$$

$$\|E_{22}^{(k)}\|_2 \leq \frac{1}{(1+\sigma_{\min}^2(X_{k\oplus 1}))^{1/2}} \cdot \frac{\sigma_{\max}(X_k)}{(1+\sigma_{\max}^2(X_k))^{1/2}} \|Y_k\|_F. \tag{6.6}$$

[illegible] Transform the sequence T_k with $\tilde{Q}_k$ in a cyclic transformation:

$$\tilde{Q}_{k\oplus 1}^T T_k \tilde{Q}_k = \underbrace{Q_{k\oplus 1}^T T_k Q_k}_{\hat{T}_k} + \Delta Q_{k\oplus 1}^T T_k Q_k + Q_{k\oplus 1}^T T_k \Delta Q_k + \Delta Q_{k\oplus 1}^T T_k \Delta Q_k.$$

Let $Z_k = Q_k^T \Delta Q_k$. From $(Q_k + \Delta Q_k)^T (Q_k + \Delta Q_k) = I$ we have that $Q_k^T \Delta Q_k = -\Delta Q_k^T Q_k$ up to first order, and by dropping the second order term, we get

$$\tilde{Q}_{k\oplus 1}^T T_k \tilde{Q}_k = \hat{T}_k + \hat{T}_k Z_k - Z_{k\oplus 1} \hat{T}_k$$

for $k = 0, 1, \ldots, K-1$.

Let E_k denote the error matrix corresponding to the kth cyclic transformation (4.7), i.e., $\tilde{T}_k = \hat{T}_k + E_k$. Partition Z_k, $k = 0, 1, \ldots, K-1$ conformally with $\hat{T}_k$ and observe that

$$\tilde{Q}_{k\oplus 1}^T T_k \tilde{Q}_k = \hat{T}_k + E_k = \begin{bmatrix} \hat{T}_{11}^{(k)} & \hat{T}_{12}^{(k)} \\ 0 & \hat{T}_{22}^{(k)} \end{bmatrix} + \begin{bmatrix} E_{11}^{(k)} & E_{12}^{(k)} \\ E_{21}^{(k)} & E_{22}^{(k)} \end{bmatrix},$$

where

$$E_k = \begin{bmatrix} E_{11}^{(k)} & E_{12}^{(k)} \\ E_{21}^{(k)} & E_{22}^{(k)} \end{bmatrix} = \hat{T}_k Z_k - Z_{k\oplus 1} \hat{T}_k,$$

i.e.,

$$\text{(6.7)} \qquad \begin{cases} E_{11}^{(k)} = \hat{T}_{11}^{(k)} Z_{11}^{(k)} + \hat{T}_{12}^{(k)} Z_{21}^{(k)} - Z_{11}^{(k\oplus 1)} \hat{T}_{11}^{(k)}, \\ E_{12}^{(k)} = \hat{T}_{11}^{(k)} Z_{12}^{(k)} + \hat{T}_{12}^{(k)} Z_{22}^{(k)} - Z_{11}^{(k\oplus 1)} \hat{T}_{12}^{(k)} - Z_{12}^{(k\oplus 1)} \hat{T}_{22}^{(k)}, \\ E_{21}^{(k)} = \hat{T}_{22}^{(k)} Z_{21}^{(k)} - Z_{21}^{(k\oplus 1)} \hat{T}_{11}^{(k)}, \\ E_{22}^{(k)} = \hat{T}_{22}^{(k)} Z_{22}^{(k)} - Z_{22}^{(k\oplus 1)} \hat{T}_{22}^{(k)} - Z_{21}^{(k\oplus 1)} \hat{T}_{12}^{(k)}. \end{cases}$$

As we will show below, $E_{22}^{(k)}$ and $E_{11}^{(k)}$ perturb the eigenvalues of the matrix product $\Phi_A(K, 0)$ directly but do not affect stability. $E_{21}^{(k)}$ is critical since it affects both the stability of the reordering and the eigenvalues. $E_{12}^{(k)}$ is of minor interest since it does not perturb the eigenvalues explicitly nor does it affect the stability. The task is now to derive norm bounds for the error matrix blocks $E_{11}^{(k)}$, $E_{21}^{(k)}$, and $E_{22}^{(k)}$.

By assuming that ΔX_k, $k = 0, 1, \ldots, K-1$, are nonsingular and applying the analysis of the QR-factorization from [2] to each of our K independent QR-factorizations, we get

$$\text{(6.8)} \qquad Z_{11}^{(k)} = {Q_{11}^{(k)}}^T \Delta X_k R_k^{-1} - \Delta R_k R_k^{-1},$$

$$\text{(6.9)} \qquad Z_{21}^{(k)} = {Q_{12}^{(k)}}^T \Delta X_k R_k^{-1},$$

$$\text{(6.10)} \qquad Z_{22}^{(k)} = -{Q_{12}^{(k)}}^T \Delta X_k R_k^{-1} {Q_{11}^{(k)}}^T {Q_{12}^{(k)}}^{-T}.$$

Using (6.8), (6.9), (6.10), (4.6), and (6.1), the error matrix blocks $E_{11}^{(k)}$, $E_{21}^{(k)}$, and $E_{22}^{(k)}$ in (6.7) boil down to

$$\text{(6.11)} \qquad \begin{cases} E_{11}^{(k)} = {Q_{11}^{(k\oplus 1)}}^T Y_k R_k^{-1} - \hat{T}_{11}^{(k)} \Delta R_k R_k^{-1} + \Delta R_{k\oplus 1} R_{k\oplus 1}^{-1} \hat{T}_{11}^{(k)}, \\ E_{21}^{(k)} = {Q_{12}^{(k\oplus 1)}}^T Y_k R_k^{-1}, \\ E_{22}^{(k)} = -{Q_{12}^{(k\oplus 1)}}^T Y_k R_k^{-1} {Q_{11}^{(k)}}^T {Q_{12}^{(k)}}^{-T} \end{cases}$$

as first order results. We see that $E_{22}^{(k)}$, $E_{21}^{(k)}$, and $E_{11}^{(k)}$ are essentially related to the K residual matrices Y_k of the associated PSE and the blocks R_k, $Q_{11}^{(k)}$, and $Q_{12}^{(k)}$ from the K QR-factorizations. From (4.5) we have that

$$Q_{21}^{(k)} = R_k^{-1}, \qquad R_k^T R_k = I + X_k^T X_k,$$

which gives

$$\sigma^2(R_k) = \lambda(R_k^T R_k) = \lambda(I + X_k^T X_k) = 1 + \lambda(X_k^T X_k) = 1 + \sigma^2(X_k).$$

By the above argument we get

$$\|Q_{21}^{(k)}\|_2 = \|R_k^{-1}\|_2 = \frac{1}{\sigma_{\min}(R_k)} = \frac{1}{(1+\sigma_{\min}^2(X_k))^{1/2}}.$$

Further, from [24] we have

$$\|\Delta R_k R_k^{-1}\|_F \leq 2\|D_k^+\|_2 \|\Delta X_k\|_F,$$

and by the CS decomposition of Q (see, e.g., [10, 25]) we get the following norm relations:

$$\|Q_{21}^{(k)}\|_2 = \|Q_{12}^{(k)}\|_2, \ \|Q_{11}^{(k)}\|_2 = \|Q_{22}^{(k)}\|_2.$$

Now by combining these facts with (6.11) and applying the product and triangle inequalities for norms, we obtain the bounds of the theorem. □

1. For $K = 1$ and by inequality $(1+\sigma_{\min}^2(X_k))^{-1/2} \geq (1+\sigma_{\max}^2(X_k))^{-1/2}$, the norm bounds of Theorem 6.1 can be further bounded from above to achieve

$$\|E_{11}\|_2 \leq \frac{\sigma_{\max}(X)}{(1+\sigma_{\min}^2(X))}\|Y\|_F + 4\|\hat{T}_{11}\|_2\|D^+\|_2\|\Delta X\|_F, \tag{6.12}$$

$$\|E_{21}\|_2 \leq \frac{1}{(1+\sigma_{\min}^2(X))}\|Y\|_F, \tag{6.13}$$

$$\|E_{22}\|_2 \leq \frac{\sigma_{\max}(X)}{(1+\sigma_{\min}^2(X))}\|Y\|_F, \tag{6.14}$$

which are the norm bounds from the main theorem of [2] on the perturbation of the eigenvalues under standard eigenvalue reordering in the real Schur form.

2. Numerical experiments show that iterative refinement may improve on the computed solution X_k, especially for badly scaled problems, but may not improve on the residual sequence Y_k or on the computed eigenvalues. See also [2] for a similar observation.

6.2. Perturbation of matrix products under periodic reordering. In this section, we investigate how the errors in the individual matrices after a periodic reordering of two adjacent sequences of diagonal blocks in T_k propagate into the matrix product $\Phi_T(K,0) = T_{K-1}T_{K-2}\dots T_1T_0$.

We present a general result in the following theorem.

THEOREM 6.2. T_k K

$$T_k = \begin{bmatrix} T_{11}^{(k)} & T_{12}^{(k)} \\ 0 & T_{22}^{(k)} \end{bmatrix}.$$

$\tilde{Q}_k$ $k = 0, 1, \dots, K-1$ $\Phi_T(K,0)$ (6.2) $\tilde{T}_k$ $\hat{T}_k$ E_k (6.2)–(6.3) 6.1

$$\Phi_{\tilde{T}}(K,0) = \prod_{k=K-1}^{0} \tilde{Q}_{k\oplus 1}^T T_k \tilde{Q}_k = \Phi_{\hat{T}}(K,0) + \boldsymbol{E}, \tag{6.15}$$

$\Phi_{\hat{T}}(K,0) = Q_0^T \Phi_T(K,0) Q_0$ [illegible] $\boldsymbol{E}$ [illegible] $\boldsymbol{E}$ [illegible] T_k [illegible]

$$
\begin{cases}
\|\boldsymbol{E}_{11}\|_2 \leq \sum_{k=0}^{K-1} ((\prod_{j=K-1}^{k+1} \|\hat{T}_{11}^{(j)}\|_2) \|E_{11}^{(k)}\|_2 \\
\qquad + (\sum_{j=K-1}^{k+1} \|\varphi_1^{(k,j)}\|_2) \|E_{21}^{(k)}\|_2) (\prod_{j=k-1}^{0} \|\hat{T}_{11}^{(j)}\|_2), \\
\|\boldsymbol{E}_{21}\|_2 \leq \sum_{k=0}^{K-1} (\prod_{j=K-1}^{k+1} \|\hat{T}_{22}^{(j)}\|_2) \|E_{21}^{(k)}\|_2 (\prod_{j=k-1}^{0} \|\hat{T}_{11}^{(j)}\|), \\
\|\boldsymbol{E}_{22}\|_2 \leq \sum_{k=0}^{K-1} (\prod_{j=K-1}^{k+1} \|\hat{T}_{22}^{(j)}\|_2) (\|E_{21}^{(k)}\|_2 \sum_{j=k-1}^{0} \|\varphi_2^{(k,j)}\|_2 \\
\qquad + \|E_{22}^{(k)}\|_2 (\prod_{j=k-1}^{0} \|\hat{T}_{22}^{(j)}\|_2)),
\end{cases} \tag{6.16}
$$

[illegible]

$$
\|\varphi_1^{(k,j)}\|_2 \leq \|\hat{T}_{12}^{(j)}\|_2 \prod_{l=K-1}^{j+1} \|\hat{T}_{11}^{(l)}\|_2 \prod_{l=j-1}^{k+1} \|\hat{T}_{22}^{(l)})\|_2, \tag{6.17}
$$

$$
\|\varphi_2^{(k,j)}\|_2 \leq \|\hat{T}_{12}^{(j)}\|_2 \prod_{l=k-1}^{j+1} \|\hat{T}_{11}^{(l)}\|_2 \prod_{l=j-1}^{0} \|\hat{T}_{22}^{(l)}\|_2 \tag{6.18}
$$

[illegible] Up to first order perturbations, we have

$$
\begin{aligned}
\Phi_{\tilde{T}}(K,0) &= \prod_{k=K-1}^{0} \tilde{Q}_{k\oplus 1}^T T_k \tilde{Q}_k \\
&= \Phi_{\hat{T}}(K,0) + \sum_{k=0}^{K-1} \Phi_{\hat{T}}(K,k+1) E_k \Phi_{\hat{T}}(k,0) = \Phi_{\hat{T}}(K,0) + \boldsymbol{E}.
\end{aligned} \tag{6.19}
$$

The bounds follow by applying the triangle inequality and the submultiplicativity of norms to the error matrix $\boldsymbol{E}$ in block partitioned form. For details see [11]. ∎

For illustration, we display the explicit results of Theorem 6.2 for two simple cases in the following corollary.

COROLLARY 6.3. [illegible] 6.2 [illegible] $K = 2$ [illegible] $\boldsymbol{E}$ (6.15) [illegible]

$$
\begin{aligned}
\|\boldsymbol{E}_{11}\|_2 &\leq \|\hat{T}_{11}^{(1)}\|_2 \|E_{11}^{(0)}\|_2 + \|\hat{T}_{12}^{(1)}\|_2 \|E_{21}^{(0)}\|_2 + \|\hat{T}_{11}^{(0)}\|_2 \|E_{11}^{(1)}\|_2, \\
\|\boldsymbol{E}_{21}\|_2 &\leq \|\hat{T}_{22}^{(1)}\|_2 \|E_{21}^{(0)}\|_2 + \|\hat{T}_{11}^{(0)}\|_2 \|E_{21}^{(1)}\|_2, \\
\|\boldsymbol{E}_{22}\|_2 &\leq \|\hat{T}_{22}^{(1)}\|_2 \|E_{22}^{(0)}\|_2 + \|\hat{T}_{12}^{(0)}\|_2 \|E_{21}^{(1)}\|_2 + \|\hat{T}_{22}^{(0)}\|_2 \|E_{22}^{(1)}\|_2.
\end{aligned}
$$

[illegible] $K = 3$ [illegible]

$$
\begin{aligned}
\|\boldsymbol{E}_{11}\|_2 \leq\ & \|\hat{T}_{11}^{(2)}\|_2 \|\hat{T}_{11}^{(1)}\|_2 \|E_{11}^{(0)}\|_2 + (\|\hat{T}_{11}^{(2)}\|_2 \|\hat{T}_{12}^{(1)}\|_2 + \|\hat{T}_{12}^{(2)}\|_2 \|\hat{T}_{22}^{(1)}\|_2) \|E_{21}^{(0)}\|_2 \\
& + \|\hat{T}_{11}^{(2)}\|_2 \|\hat{T}_{11}^{(0)}\|_2 \|E_{11}^{(1)}\|_2 + \|\hat{T}_{12}^{(2)}\|_2 \|\hat{T}_{11}^{(0)}\|_2 \|E_{21}^{(1)}\|_2 + \|\hat{T}_{11}^{(1)}\|_2 \|\hat{T}_{11}^{(0)}\|_2 \|E_{11}^{(2)}\|_2, \\
\|\boldsymbol{E}_{21}\|_2 \leq\ & \|\hat{T}_{22}^{(2)}\|_2 \|\hat{T}_{22}^{(1)}\|_2 \|E_{21}^{(0)}\|_2 + \|\hat{T}_{22}^{(2)}\|_2 \|\hat{T}_{11}^{(0)}\|_2 \|E_{21}^{(1)}\|_2 + \|\hat{T}_{11}^{(1)}\|_2 \|\hat{T}_{11}^{(0)}\|_2 \|E_{21}^{(2)}\|_2, \\
\|\boldsymbol{E}_{22}\|_2 \leq\ & \|\hat{T}_{22}^{(2)}\|_2 \|\hat{T}_{22}^{(1)}\|_2 \|E_{22}^{(0)}\|_2 + \|\hat{T}_{22}^{(2)}\|_2 \|\hat{T}_{12}^{(0)}\|_2 \|E_{21}^{(1)}\|_2 + \|\hat{T}_{22}^{(2)}\|_2 \|\hat{T}_{22}^{(0)}\|_2 \|E_{22}^{(1)}\|_2 \\
& + (\|\hat{T}_{11}^{(1)}\|_2 \|\hat{T}_{12}^{(0)}\|_2 + \|\hat{T}_{12}^{(1)}\|_2 \|\hat{T}_{22}^{(0)}\|_2) \|E_{21}^{(2)}\|_2 + \|\hat{T}_{22}^{(1)}\|_2 \|\hat{T}_{22}^{(0)}\|_2 \|E_{22}^{(2)}\|_2
\end{aligned}
$$

We remark that the analysis in Theorem 6.2 and Corollary 6.3 assumes that the involved matrix products and sums are computed exactly. For a rounding error analysis regarding matrix products and sums, see, e.g., [15].

Theorems 6.1 and 6.2 can be combined to produce computable bounds for the perturbations of the diagonal blocks of $\Phi_{\tilde{T}}(K,0)$ under periodic eigenvalue reordering. We can also apply known perturbation results for the standard eigenvalue problem [25] and the periodic eigenvalue problem [20, 4] to the submatrix products $\Phi_{\tilde{T}}(K,0)_{11}$ and $\Phi_{\tilde{T}}(K,0)_{22}$. This is a matter of further investigation.

7. Computational experiments. We demonstrate the stability and reliability of the direct reordering method by considering some numerical examples. The test examples range from well-conditioned to ill-conditioned problems, including matrix sequences with fixed and time-varying dimensions, and sequences of small and large periodicity. In the following, we present results for a representative selection of problems, where, except for one example, two complex conjugate eigenvalue pairs of a periodic real sequence A_k are reordered ($p_1 = p_2 = 2$). The associated PSEs of our direct periodic reordering method are solved by applying GEPP to $Z_{\mathrm{PSE}}x = c$ and utilizing the structure of Z_{PSE} in (5.1). All experiments are carried out in double precision ($\epsilon_{\mathrm{mach}} \approx 2.2 \times 10^{-16}$) on an UltraSparc II (450 Mhz) workstation.

Examples 1 and 3 below are constructed as follows. First, we specify K, n_k, $k = 0, 1, \ldots, K-1$, and $\min_k(n_k)$ eigenvalues or $K \cdot \min_k(n_k)$ diagonal and $\min_k(n_k) - 1$ subdiagonal elements. Then a random sequence T_k as in (1.2) is generated with 1×1 and 2×2 diagonal blocks corresponding to specified eigenvalues or diagonal, subdiagonal, and superdiagonal entries. Finally, orthogonal matrices Z_k, $k = 0, 1, \ldots, K-1$, are constructed from QR-factorizing K uniformly distributed random $n_k \times n_k$ matrices, which are applied in a K-cyclic orthogonal transformation of T_k to get A_k. Examples 4 and 5 illustrate reordering of two periodic sequences already in PRSF. Finally, Example 2 is from a real application.

In Table 7.1, we display the periodicity K, problem dimensions n_k for $k = 0, 1, \ldots, K-1$, the computed value of sep[PSE], and a reciprocal condition number s for the eigenvalues of $\Phi_T(K,0)_{11}$,

$$s = 1/\sqrt{1 + \|X_0\|_F^2},$$

where X_0 is the generator matrix for the periodic reordering of $\Phi_T(K,0)$ (see section 4). The last two quantities signal the conditioning of the problems considered.

Results from periodic reordering using our direct method are presented in Table 7.2. We display the maximum relative change of the eigenvalues under the periodic reordering

$$e_\lambda = \max_k \frac{|\lambda_k - \tilde{\lambda}_k|}{|\lambda_k|}, \ \lambda_k \in \lambda(\Phi_T(K,0)).$$

In addition, we display five residual quantities for the computed results. These include two stability tests used in our method, namely a

$$R_{\mathrm{weak}} = \max_k \|\tilde{Q}_{11}^{(k)} - X_k \tilde{Q}_{21}^{(k)}\|_F,$$

and a

$$R_{\mathrm{strong}} = \max_k (\|T_k - \tilde{Q}_{k\oplus 1}\tilde{T}_k\tilde{Q}_k^T\|_F, \|\tilde{T}_k - \tilde{Q}_{k\oplus 1}^T T_k \tilde{Q}_k\|_F),$$

TABLE 7.1

Problem characteristics for the examples considered. 4a *and* 4b *refer to Example* 4 *with period* 2 *and* 100*, respectively.*

Example	K	n_k	sep[PSE]	s
1	3	$4+k$	6.9E-01	7.2E-01
2	120	4	4.7E-03	5.5E-01
3	10	2	9.9E+00	1.0E+00
4a	2	4	4.5E-15	1.1E-14
4b	100	4	1.3E-16	1.3E-16
5	2	4	6.2E+03	6.6E-01

TABLE 7.2

Computational results for periodic reordering. 4a *and* 4b *refer to Example* 4 *with period* $K = 2$ *and* 100*, respectively.* 5a *and* 5b *refer to Example* 5 *without scaling and with scaling.*

Example	e_λ	R_{weak}	R_{strong}	R_{eprsf}	R_{reord}	R_{orth}
1	4.6E-16	2.2E-16	1.6E-15	4.7E-15	5.6E-15	1.3E+01
2	1.6E-15	2.9E-16	1.8E-15	9.0E-15	9.8E-15	2.0E+01
3	1.4E-15	1.9E-16	8.4E-15	7.3E-15	1.0E-14	4.1E+00
4a	3.6E-16	2.5E-16	1.4E-15	0	1.2E-15	2.1E+00
4b	3.7E-16	2.3E-16	3.2E-18	0	1.9E-15	3.6E+00
5a	2.2E-01	1.2E-16	6.6E-12	0	5.8E-12	3.3E+00
5b	2.0E-09	2.3E-16	4.3E-12	0	5.6E-12	3.3E+00

which is the maximum residual norm associated with the cyclic transformations $\tilde{Q}_k$ used in the reordering. Tolerances for these tests can optionally be specified. Depending on the outcome of our stability test (weak or strong), we either reject the swap or perform a swapping with guaranteed backward stability. Rejecting a swap means that we avoid the risk that errors induced during the reordering computations may change the eigenvalues drastically. It is the sensitivity of the associated eigenspaces that matters most (see [18]). Since the extra cost for the strong stability test is marginal, it is recommended. The last three columns in Table 7.2 display the maximum residual norms of the (extended) periodic Schur decomposition (1.2) before and after reordering, computed as

$$R_{\text{eprsf}} = \max_k(\|A_k - Z_{k\oplus 1}T_kZ_k^T\|_F, \|T_k - Z_{k\oplus 1}^T A_k Z_k\|_F),$$

and

$$R_{\text{reord}} = \max_k(\|A_k - \tilde{Z}_{k\oplus 1}\tilde{T}_k\tilde{Z}_k^T\|_F, \|\tilde{T}_k - \tilde{Z}_{k\oplus 1}^T A_k \tilde{Z}_k\|_F),$$

and a relative orthogonality check over the whole period K after periodic reordering:

$$R_{\text{orth}} = \frac{\max_k(\|I_{n_k} - \tilde{Z}_k^T\tilde{Z}_k\|_F, \|I_{n_k} - \tilde{Z}_k\tilde{Z}_k^T\|_F)}{\epsilon_{\text{mach}}}.$$

For these three residual norms, the K-cyclic transformations Z_k and $\tilde{Z}_k$ correspond to Z_k and $\hat{Z}_k$ in (3.3), respectively.

The computed eigenvalues before and after the periodic reordering are presented to full machine accuracy under each example.

1. We consider a time-varying sequence with $K = 3$ and $n_k = 4+k,\ k = 0, 1, 2$, and eigenvalues $1.0 \pm 2.0i, -7.0 \pm 0.5i$. The computed eigenvalues of the matrix product $\Phi_T(K, 0) = T_2T_1T_0$ are

$$\begin{aligned} \lambda_1 &= 1.000000000000000 \pm 2.000000000000000i, \\ \lambda_2 &= -7.000000000000001 \pm 5.000000000000001i. \end{aligned}$$

The spectrum is well separated. After the periodic reordering of the blocks we obtained $\tilde{\lambda}_1 = \lambda_2$ and $\tilde{\lambda}_2 = \lambda_1$ to full accuracy.

EXAMPLE 2 (satellite control [29]). We consider reordering in a 4×4 periodic matrix sequence that describes a control system of a satellite on orbit around the earth. The periodicity is $K = 120$. The computed eigenvalues of the sequence are

$$\begin{aligned} \lambda_1 &= 0.9941836588706161 \pm 0.1076979685723037i, \\ \lambda_2 &= 0.7625695885261465 \pm 0.6469061930874623i. \end{aligned}$$

The reordered eigenvalues are

$$\begin{aligned} \tilde{\lambda}_1 &= 0.7625695885261450 \pm 0.6469061930874582i, \\ \tilde{\lambda}_2 &= 0.9941836588706161 \pm 0.1076979685723021i. \end{aligned}$$

This application example shows that periodic reordering works fine for well-conditioned problems with large periods as well.

EXAMPLE 3. We consider reordering a sequence with $K = 10$, and $p_1 = p_2 = 1$, and the computed sequence in PRSF is

$$T_k = \begin{bmatrix} 10^1 & t_{12}^{(k)} \\ 0 & 10^{-1} \end{bmatrix}, \ k = 0, 1, \ldots, K-1,$$

where $|t_{12}^{(k)}| \leq 1$. The computed eigenvalues of the product $\Phi_T(K, 0)$ are

$$\begin{aligned} \lambda_1 &= 9.999999999999987 \times 10^9, \\ \lambda_2 &= 1.000000000000013 \times 10^{-10}. \end{aligned}$$

After the periodic reordering we obtain

$$\begin{aligned} \tilde{\lambda}_1 &= 1.000000000000015 \times 10^{-10}, \\ \tilde{\lambda}_2 &= 9.999999999999989 \times 10^9. \end{aligned}$$

Reordering of 1×1 blocks in PRSF can be carried out by propagating a Givens rotation through the matrix product [5], but this process is not forward stable. For this example, the rotation approach does not deliver one single correct digit in the reordered eigenvalues, whereas the direct reordering method delivers an acceptable error in the eigenvalues.

EXAMPLE 4. We consider a sequence already in PRSF with $K = 2$ and $n_k = 4$, $k = 0, 1$, and eigenvalues $0.2 \pm (1.2 + 10^{-14})i$, $0.2 \pm 1.2i$. The computed eigenvalues of the matrix $\Phi_T(K, 0) = T_1 T_0$ are

$$\begin{aligned} \lambda_1 &= 0.200000000000000 \pm 1.200000000000001i, \\ \lambda_2 &= 0.200000000000000 \pm 1.200000000000000i. \end{aligned}$$

The spectrum is not well separated. After the periodic reordering we obtained

$$\begin{aligned} \tilde{\lambda}_1 &= 0.200000000000000 \pm 1.200000000000000i, \\ \tilde{\lambda}_2 &= 0.200000000000000 \pm 1.200000000000001i, \end{aligned}$$

so the periodic reordering was perfect, even though the problem has very close eigenvalues. Indeed, we obtain reordered eigenvalues to full machine accuracy for periods up to 100.

5. First, we consider a problem already in PRSF with large separation and $K = 2$, $n_k = 4$, $k = 0, 1$, and the eigenvalues $\epsilon_{\text{mach}}^{1/2} \pm \epsilon_{\text{mach}}^{1/2} i$, $\epsilon_{\text{mach}}^{-1/2} \pm \epsilon_{\text{mach}}^{-1/2} i$. Moreover, the involved matrices have almost the same Frobenius norm ($\approx 1.8 \times 10^4$), but the matrices in the subsequences $T_{11}^{(k)}$ and $T_{22}^{(k)}$ have very different norms: $\|T_{11}^{(0)}\|_F \approx 1.4 \times 10^4$, $\|T_{11}^{(1)}\|_F \approx 1.4 \times 10^4$, $\|T_{22}^{(0)}\|_F \approx 7.0 \times 10^{-12}$, $\|T_{22}^{(1)}\|_F \approx 8.6 \times 10^3$. The computed eigenvalues of the product $\Phi_T(K, 0)$ are

$$\begin{aligned} \lambda_1 &= 6.710886400000000 \times 10^7 \quad \pm \quad 6.710886400000003 \times 10^7 i, \\ \lambda_2 &= 1.490116119384766 \times 10^{-8} \quad \pm \quad 1.490116119384766 \times 10^{-8} i. \end{aligned}$$

After the periodic reordering without diagonal scaling we obtain

$$\begin{aligned} \tilde{\lambda}_1 &= 1.168840447839719 \times 10^{-8} \quad \pm \quad 9.309493732240201 \times 10^{-9} i, \\ \tilde{\lambda}_2 &= 6.710886400000001 \times 10^7 \quad \pm \quad 6.710886400000000 \times 10^7 i. \end{aligned}$$

The problem is well-conditioned in the sense of sep[PSE], the norm of the generator matrix (see s in Table 7.1), and the reordering passes the stability tests, but since the eigenvalues differ almost 16 orders of magnitude the relative error in the smallest eigenvalues become very large due to the finite precision arithmetic.

Next, we consider the same problem as above, but now we perform diagonal scaling $T_1 T_0 = T_1 D_1 D_1^{-1} T_0$ before periodic reordering such that the blocks $T_{22}^{(0)}$ and $T_{22}^{(1)}$ have about the same norm. Now the periodic reordering gives

$$\begin{aligned} \tilde{\lambda}_1 &= 1.490116120748016 \times 10^{-8} \quad \pm \quad 1.490116125160257 \times 10^{-8} i, \\ \tilde{\lambda}_2 &= 6.710886400000000 \times 10^7 \quad \pm \quad 6.710886400000001 \times 10^7 i, \end{aligned}$$

which is quite an improvement (8 orders of magnitude) compared to the results without scaling. Not surprisingly, periodic reordering is sensitive to large differences in the norms within the subsequences $T_{11}^{(k)}$ and $T_{22}^{(k)}$.

8. Future work. Next, we will focus on computing periodic eigenspaces with specified eigenvalues and associated error bounds based on condition estimation (see, e.g., [18]), as well as producing library-standard (LAPACK [1], SLICOT [21]) software for the eigenvalue reordering algorithm presented in this paper.

Acknowledgments. The authors are grateful to Daniel Kressner for constructive comments on the subject and earlier versions of this manuscript and to Andras Varga for valuable comments on the subject and for providing us with software for computing the extended periodic Schur decomposition and data for Example 2.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. W. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, 3rd ed., SIAM, Philadelphia, 1999.

[2] Z. Bai and J. W. Demmel, *On swapping diagonal blocks in real Schur form*, Linear Algebra Appl., 186 (1993), pp. 73–95.

[3] R. H. Bartels and G. W. Stewart, *Algorithm* 432 : *The Solution of the Matrix Equation* $AX - BX = C$, Comm. ACM, 8 (1972), pp. 820–826.

[4] P. Benner, V. Mehrmann, and H. Xu, *Perturbation analysis for the eigenvalue problem of a formal product of matrices*, BIT, 42 (2002), pp. 1–43.

[5] A. Bojanczyk, G. H. Golub, and P. Van Dooren, *The periodic Schur decomposition: Algorithm and applications*, in Advanced Signal Processing Algorithms, Architectures, and Implementations III, Proc. SPIE Conference 1770, SPIE, Bellingham, WA, 1992, pp. 31–42.

[6] A. Bojanczyk and P. Van Dooren, *On propagating orthogonal transformations in a product of* 2×2 *triangular matrices*, in Numerical Linear Algebra, de Gruyter, New York, 1993, pp. 1–9.

[7] A. Bojanczyk and P. Van Dooren, *Reordering diagonal blocks in the real Schur form*, in Linear Algebra for Large Scale and Real-Time Applications, M. S. Moonen, G. H. Golub, and B. L. R. De Moor, eds., Kluwer Academic Publishers, Amsterdam, 1993, pp. 351–352.

[8] J. J. Dongarra, S. Hammarling, and J. H. Wilkinson, *Numerical considerations in computing invariant subspaces*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 145–161.

[9] G. Fairweather and I. Gladwell, *Algorithms for almost block diagonal linear systems*, SIAM Rev., 46 (2004), pp. 49–58.

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

[11] R. Granat and B. Kågström, *Direct Eigenvalue Reordering in a Product of Matrices in Extended Periodic Real Schur Form*, Report UMINF 05.05, Umeå University, Umeå, Sweden, 2005.

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

[13] J. J. Hench and A. J. Laub, *Numerical solution of the discrete-time periodic Riccati equation*, IEEE Trans. Automat. Control, 39 (1994), pp. 1197–1210.

[14] N. J. Higham, *Fortran codes for estimating the one-norm of a real or complex matrix, with applications to condition estimation*, ACM Trans. Math. Software, 14 (1988), pp. 381–396.

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

[16] B. Kågström, *A direct method for reordering eigenvalues in the generalized real Schur form of a regular matrix pair* (A, B), in Linear Algebra for Large Scale and Real-Time Applications, M. S. Moonen, G. H. Golub, and B. L. R. De Moor, eds., Kluwer Academic Publishers, Amsterdam, 1993, pp. 195–218.

[17] B. Kågström and P. Poromaa, *Distributed and shared memory block algorithms for the triangular Sylvester equation with* sep^{-1} *estimators*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 90–101.

[18] B. Kågström and P. Poromaa, *Computing eigenspaces with specified eigenvalues of a regular matrix pair* (A, B) *and condition estimation: Theory, algorithms, and software*, Numer. Algorithms, 12 (1996), pp. 369–407.

[19] D. Kressner, *Numerical Methods and Software for General and Structured Eigenvalue Problems*, Ph.D. thesis, TU Berlin, Institut für Mathematik, Berlin, Germany, 2004.

[20] W.-W. Lin and J.-G. Sun, *Perturbation analysis for the eigenproblem of periodic matrix pairs*, Linear Algebra Appl., 337 (2001), pp. 157–187.

[21] *SLICOT Library, The Numerics in Control Network (Niconet)*, http://www.win.tue.nl/niconet/index.html.

[22] J. Sreedhar and P. Van Dooren, *A Schur approach for solving some periodic matrix equations*, in Systems and Networks: Mathematical Theory and Applications, U. Helmke, R. Mennicken, and J. Saurer, eds., Akademie Verlag, Berlin, 1994, pp. 339–362.

[23] G. W. Stewart, *Algorithm* 407 : *HQR*3 *and EXCHNG: FORTRAN programs for calculating the eigenvalues of a real upper Hessenberg matrix in a prescribed order*, ACM Trans. Math. Software, 2 (1976), pp. 275–280.

[24] G. W. Stewart, *Perturbation bounds for the QR factorization of a matrix*, SIAM J. Numer. Anal., 14 (1977), pp. 509–518.

[25] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

[26] P. Van Dooren, *Algorithm* 590 : *DSUBSP and EXCHQZ: Fortran subroutines for computing deflating subspaces with specified spectrum*, ACM Trans. Math. Software, 8 (1982), pp. 376–382.

[27] A. Varga, *Periodic Lyapunov equations: Some applications and new algorithms*, Internat. J. Control, 67 (1997), pp. 69–87.

[28] A. Varga, *Balancing related methods for minimal realization of periodic systems*, Systems Control Lett., 36 (1999), pp. 339–349.

[29] A. Varga and S. Pieters, *Gradient-based approach to solve optimal periodic output feedback control problems*, Automatica, 45 (1998), pp. 477–481.

[30] A. Varga and P. Van Dooren, *Computational methods for periodic systems—an overview*, in Proceedings of the of IFAC Workshop on Periodic Control Systems, Como, Italy, 2001, pp. 171–176.

[31] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, UK, 1965.

[32] S. J. Wright, *A collection of problems for which Gaussian elimination with partial pivoting is unstable*, SIAM J. Sci. Comput., 14 (1993), pp. 231–238.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 2, pp. 301–305

DISTANCES FROM A HERMITIAN PAIR TO DIAGONALIZABLE AND NONDIAGONALIZABLE HERMITIAN PAIRS*

CHI-KWONG LI† AND ROY MATHIAS‡

Abstract. Let $W(T)$ and $r(T)$ denote the numerical range and numerical radius of an $n \times n$ complex matrix T. Let H_n^2 denote the space of pairs of $n \times n$ Hermitian matrices. Define a norm on H_n^2 by $\|(X,Y)\| = r(X+iY)$. Take $(A,B) \in H_n^2$. It is shown that if $0 \in W(A+iB)$, then $\inf\{|\mu| : \mu \notin W(A+iB)\}$ is an upper bound on the distance to the nearest pair that is simultaneously diagonalizable by congruence. If $0 \notin W(A+iB)$, then $\min\{|\mu| : \mu \in W(A+iB)\}$, which is the Crawford number of the pair (A,B), is equal to the distance to the nearest pair that is not simultaneously diagonalizable by congruence. The results are similar when the numerical radius is replaced by the spectral norm.

Key words. definite Hermitian pair, nondiagonalizable, numerical range, numerical radius, Crawford number

AMS subject classifications. 15A60, 15A18

DOI. 10.1137/040616346

1. Introduction.

Let M_n (respectively, H_n) be the set of $n \times n$ complex (respectively, Hermitian) matrices. Two Hermitian matrices $A, B \in H_n$ are said to be a definite pair if $|x^*(A+iB)x| \neq 0$ for every nonzero vector $x \in \mathbb{C}^n$.

Definite Hermitian pairs have useful algorithmic and theoretical properties. For example, it is known (see [3, Theorem 1.7.17]) that if (A,B) is a definite Hermitian pair, then it is diagonalizable by congruence, i.e., there is an invertible matrix $S \in M_n$ so that both S^*AS and S^*BS are diagonal matrices; equivalently, $S^*(A+iB)S$ is a diagonal matrix. This property is very useful in the analysis of the Hermitian generalized eigenvalue problem, $Ax = \lambda Bx$. If (A,B) is a definite pair, then the corresponding generalized eigenvalues are real and can be found by solving a related Hermitian eigenvalue problem [1, section 8.7.3].

Recall that the numerical range of $T \in M_n$ is

$$W(T) = \{x^*Tx : x \in \mathbb{C}^n,\ x^*x = 1\}$$

and that the numerical radius of T is

$$r(T) = \max\{|x^*Tx| : x \in \mathbb{C}^n, x^*x = 1\},$$

which is the maximum distance of a point in the numerical range to the origin.

It is known that $W(T)$ is always a compact convex set in $\mathbb{C}$, and that the numerical radius is a norm on M_n satisfying

$$r(T) \leq \|T\| \leq 2r(T) \quad \text{for all } T \in M_n, \tag{1.1}$$

*Received by the editors October 5, 2004; accepted for publication (in revised form) by I. C. F. Ipsen July 13, 2005; published electronically April 21, 2006. This work was supported in part by NSF grants DMS-9704534 and DMS-0071994.

http://www.siam.org/journals/simax/28-2/61634.html

†Department of Mathematics, College of William & Mary, Williamsburg, VA 23187 (ckli@math.wm.edu).

‡School of Mathematics, University of Birmingham, Birmingham B15 2TT, United Kingdom (mathias@maths.bham.ac.uk). This author was supported by an Engineering and Physical Sciences Research Council Visiting Fellowship under grant GR/T08739 at the University of Manchester, UK.

in comparison with the spectral norm $\|T\|$; for example, see [3, 4]. Also, it is known that (A, B) is a definite Hermitian pair if and only if $W(A+iB)$ does not contain the origin, which is equivalent to the existence of $a, b \in \mathbb{R}$ such that $aA + bB$ is positive definite; see [3, p. 72]. We define the Crawford number of (A, B) by

$$c(A, B) = \min\{|x^*(A + iB)x| : x \in \mathbb{C}^n, x^*x = 1\},$$

which is the shortest distance between a point in $W(A + iB)$ and the origin. The Crawford number often appears in the study of perturbation bounds in the study of problems involving definite Hermitian pairs; see [5, Chapter VI].

It is easily shown (Proposition 2.1) that $c(A, B)$ is the distance to the nearest nondefinite pair. The purpose of this paper, Theorem 2.3, is to show that $c(A, B)$ is also the distance from (A, B) to the set of nondiagonalizable pairs even though diagonalizability by congruence is not equivalent to definiteness. If $c(A, B) = 0$, i.e., $0 \in W(A + iB)$, then $A + iB$ may or may not be diagonalizable by congruence, but in Proposition 2.2 we give an upper bound for the distance between (A, B) to the set of diagonalizable pairs.

2. Results and proofs.

PROPOSITION 2.1. [illegible] (A, B) [illegible] $x \in \mathbb{C}^n$ [illegible] $|x(A + iB)x| = c(A, B)$ [illegible] $(E_0, F_0) = -(x^*AxI, x^*BxI)$ [illegible] $(A + E_0, B + F_0)$ [illegible]

$$c(A, B) = r(E_0+iF_0) = \min\{r(E+iF) : (A+E, B+F) \text{ [illegible] }\}. \tag{2.1}$$

[illegible] (2.1) [illegible] $r(\cdot)$ [illegible] $\|\cdot\|$ [illegible]. Let r_D denote the right-hand side of (2.1). Let $r_{D,\|\cdot\|}$ denote the right-hand side of (2.1) when $r(\cdot)$ is replaced by $\|\cdot\|$.

Suppose $x \in \mathbb{C}^n$ is a unit vector such that $|x^*(A+iB)x| = c(A, B)$ and $(E_0, F_0) = -(x^*AxI, x^*BxI)$. Then $0 \in W((A + E_0) + i(B + F_0))$ and hence $(A + E_0, B + F_0)$ is not definite. Since $E_0 + iF_0$ is a multiple of the identity, it follows that

$$\|E_0 + iF_0\| = |(x^*Ax) + i(x^*Bx)| = c(A, B).$$

Thus $r_{D,\|\cdot\|} \le c(A, B)$.

By (1.1), we have $r_D \le r_{D,\|\cdot\|}$. Let (E, F) be a Hermitian pair such that $(A + E, B + F)$ is not definite. Consider a unit vector $y \in \mathbb{C}^n$ such that $y^*(A + E)y = y^*(B + F)y = 0$ or, equivalently, $y^*Ay = -y^*Ey$ and $y^*By = -y^*Fy$. Therefore

$$c(A, B) \le |y^*(A + iB)y| = |y^*(E + iF)y| \le r(E + iF). \tag{2.2}$$

Thus $c(A, B) \le r_D$. Combining this with the conclusion of the previous paragraph we have $c(A, B) = r_D = r_{D,\|\cdot\|}$. □

PROPOSITION 2.2. [illegible] (A, B) [illegible] $0 \in W(A + iB)$ [illegible]

$$\begin{aligned} d(A, B) &= \inf\{|\mu| : \mu \notin W(A + iB)\} \\ &\ge \inf\{r(E + iF) : (A + E) + i(B + F) \text{ is [illegible] }\}. \end{aligned} \tag{2.3}$$

[illegible] (2.3) [illegible] $r(\cdot)$ [illegible] $\|\cdot\|$ [illegible]. Let $T = A+iB$. Since $W(T)$ is compact, there is a boundary point μ with minimum modulus. We may replace (T, μ) by $(e^{it}T, e^{it}\mu)$ for a suitable $t \in [0, 2\pi)$ so

that there is a left support line of $W(T)$ passing through μ. Then for any $\varepsilon > 0$, we can let $E + iF = (\varepsilon - \mu)I$ so that $0 \notin W(T + (E + iF))$ and hence $T + (E + iF)$ is diagonalizable by congruence. Since $\|E+iF\| = r(E+iF) \leq |\mu|+\varepsilon$ and ε is arbitrary, we get the desired inequality. □

Let (A, B) be a Hermitian pair such that $0 \in W(A + iB)$. Since $W(A + iB)$ is closed, $\inf\{|\mu| : \mu \notin W(A + iB)\}$ is not attained by any element not in $W(A + iB)$. Also,

$$\inf\{r(E + iF) : (A + E) + i(B + F) \text{ is diagonalizable by congruence}\}$$

is not always attainable. For example, if

$$A = \begin{pmatrix} 0 & 10 \\ 10 & 0 \end{pmatrix} \qquad \text{and} \qquad B = \begin{pmatrix} 11 & 0 \\ 0 & -1 \end{pmatrix},$$

then $W(A + iB)$ is an elliptical disk with minor axis joining the numbers $11i$ and $-i$, and major axis joining the numbers $10 + 5i$ and $-10 + 5i$. Clearly, $d(A, B) = 1$, and $-i$ is the boundary point of $W(A + iB)$ nearest to the origin. Suppose $E + iF$ satisfies $r(E + iF) \leq 1$. We claim that $T = (A + E) + i(B + F)$ is not diagonalizable by congruence. Suppose it is not true and that $S \in M_2$ is invertible such that S^*TS is in diagonal form. Note that $0 \in W(T)$. It follows that $W(S^*TS)$ is a line segment containing 0. Thus, there exists a complex unit ξ such that ξS^*TS is Hermitian. So, ξT is Hermitian and $\xi W(T)$ is a real line segment containing 0. Let $x, y, z \in \mathbb{C}^n$ be unit vectors such that $x^*(A + iB)x = 11i$, $y^*(A + iB)y = 10 + 5i$, and $z^*(A + iB)z = -10 + 5i$. Let $x^*Tx = \mu_1$, $y^*Ty = \mu_2$, and $z^*Tz = \mu_3$. Then $|11i - \mu_1| \leq 1$, $|10 + 5i - \mu_2| \leq 1$, and $|-10 + 5i - \mu_3| \leq 1$. Thus, $W(T)$ cannot be a line segment. Hence, T is not diagonalizable.

Next, we turn to our main result.

THEOREM 2.3. [illegible] (A, B) [illegible]

$$c(A, B) = \min\{r(E+iF) : (A+E)+i(B+F) \text{ [illegible]}\} \tag{2.4}$$

[illegible]

$$c(A, B) = \inf\{\|E+iF\| : (A+E)+i(B+F) \text{ [illegible]}\}. \tag{2.5}$$

We need two lemmas to prove Theorem 2.3. The first one is a standard result characterizing diagonalizability of a pair by congruence when one of the matrices is invertible. The second presents a perhaps surprising difference between the numerical radius and the spectral norm. This difference is the reason that the result in Theorem 2.3 contains a "min" for the numerical radius but only an "inf" for the spectral norm.

LEMMA 2.4 (see [2, Table 4.5.15, part 1 (b)]). [illegible] $A, B \in H_n$ [illegible] A [illegible] $A + iB$ [illegible] $A^{-1}B$ [illegible]

LEMMA 2.5 (see [3, Theorem 1.3.6 (b)]). [illegible] $t \in (0, 1/2]$ [illegible]

$$X = \begin{pmatrix} 0 & it \\ it & 1 \end{pmatrix}.$$

[illegible] $r(X) = 1 < \|X\|$

[illegible] 2.3. Suppose that

$$\min\{|z| : z \in W(A + iB)\}$$

occurs at $z = re^{i\theta}$; then replacing $A+iB$ by $e^{-i\theta}(A+iB)$ if necessary, we may assume that $z = i\gamma$. After a unitary similarity, we may assume with loss of generality that $B = B_1 \oplus [\gamma]$ with $B_1 \in M_{n-1}$. This implies that $a_{nn} = 0$, so write $A = \left(\begin{smallmatrix} A_{11} & A_{12} \\ A_{12}^* & 0 \end{smallmatrix}\right)$ with $A_{11} \in M_{n-1}$. Let

$$E = \operatorname{diag}(d_1, \ldots, d_{n-2}) \oplus \begin{pmatrix} 0 & t \\ t & 0 \end{pmatrix} \quad \text{and} \quad F = 0_{n-2} \oplus \operatorname{diag}(0, -\gamma).$$

Using a Schur complement argument, for example, we can show that for any $t \neq 0$ we can choose $d_1, \ldots, d_{n-2}$ with $\gamma > d_j > 0$ such that $\tilde{A} = A + E$ is invertible. We claim that $\tilde{A} + i\tilde{B}$ is not diagonalizable by congruence.

First, note that $\tilde{B} = B + F = B_1 \oplus 0$ has rank $n-1$ and hence so has $\tilde{A}^{-1}\tilde{B}$.

Write

$$\tilde{A}^{-1} = \begin{pmatrix} X & Y \\ Y^* & Z \end{pmatrix}, \text{ where } X \in M_{n-1},\ Z \in M_1.$$

Note that $0 = a_{nn} = \tilde{a}_{nn} = \det(X)/\det(\tilde{A}^{-1})$. Thus, X is singular. Hence XB_1 has at least one zero eigenvalue. Therefore, the rank $n-1$ matrix

$$\tilde{A}^{-1}\tilde{B} = \begin{pmatrix} XB_1 & 0 \\ Y^*\tilde{B}_1 & 0 \end{pmatrix}$$

has at most $n-2$ nonzero eigenvalues. Thus, $\tilde{A}^{-1}\tilde{B}$ is not diagonalizable, and our claim is proved.

Now, by Lemma 2.5, taking $t \in (0, \gamma/2)$ ensures $r(E+iF) = \gamma$, establishing (2.4).

Taking $t = \epsilon > 0$ ensures $\|E + iF\| \leq \gamma + \epsilon$ and establishes (2.5). □

A slightly more careful argument shows that if, in the proof above, $A_{12} \neq 0$, then we can take $t = 0$ in constructing $(E + iF)$ such that $(A + iB) + (E + iF)$ is not diagonalizable by congruence. The resulting $(E + iF)$ will have $\|E + iF\| = \gamma$. Thus generically, the infimum in (2.5) is attained.

Here is an instance where the infimum in (2.5) is not attained. Take the 2×2 matrices $A = 0$ and $B = I$. Clearly $c(A, B) = 1$. Let E and F be Hermitian and such that

$$(A + iB) + (E + iF) \text{ is not diagonalizable by congruence.} \tag{2.6}$$

Since both A and B are invariant under unitary similarity, we may assume without loss of generality that F is diagonal. Note that $\max\{\|E\|, \|F\|\} \leq \|E + iF\|$, so if $\|E + iF\| \leq 1$ and if the pair $(A + E, B + F)$ is not definite, then F must be of the form

$$\begin{pmatrix} -1 & 0 \\ 0 & t \end{pmatrix} \text{ or } \begin{pmatrix} t & 0 \\ 0 & -1 \end{pmatrix}.$$

In either case $B + F$ is diagonal, so the condition (2.6) requires that $A + E = E$ has nonzero off-diagonal. However, for such E and F it is the case that $\|E + iF\| > 1$.

REFERENCES

[1] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd. ed., The Johns Hopkins University Press, Baltimore, MD, 1996.
[2] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1985.
[3] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.
[4] V. Istratescu, *Introduction to Linear Operator Theory*, Marcel Dekker, New York, 1981.
[5] G. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 2, pp. 306–325

THE GEOMETRY OF THE REACHABILITY SET FOR LINEAR DISCRETE-TIME SYSTEMS WITH POSITIVE CONTROLS*

LUCA BENVENUTI† AND LORENZO FARINA†

Abstract. In this paper we study the geometrical properties of the set of reachable states of a single input discrete-time linear time invariant (LTI) system with positive controls. This set is a cone and it can be expressed as the direct sum of a linear subspace and a proper cone. In order to give a complete geometrical characterization of the reachable set, we provide a formula to evaluate the dimension of the largest reachable subspace and necessary and sufficient conditions for polyhedrality of the proper cone in terms of eigenvalues location.

Key words. dynamical systems in control, discrete-time systems, positive matrices and their generalizations

AMS subject classifications. 37N35, 93C55, 15A48

DOI. 10.1137/040612531

1. Introduction. In this paper we study the geometrical properties of the set of reachable states x_k of a single input discrete–time LTI system of the form:

$$x_{k+1} = F\,x_k + g\,u_k \qquad k = 0, 1, \dots \tag{1.1}$$

with $F \in \mathbb{R}^{n\times n}$, $g \in \mathbb{R}^n$ when the input function u_k is nonnegative for all times k. This situation is frequently encountered, for example, in medical, ecological, chemical and economical applications where the controls have a unidirectional influence [2]. Moreover, this may also occur in electro-mechanical applications (see the examples discussed in [15]).

It is worth noting that nonnegativity of the input implies that the reachable set is a convex cone. In fact, the set of states reachable in k steps can be written as

$$\begin{aligned}\mathcal{R}_k(F,g) &= \left\{x : x = \textstyle\sum_{i=0}^{k-1} F^{k-i-1} g\,u(i),\ u(i) \geq 0\right\} \\ &= \operatorname{cone}\left(g, Fg, \dots F^{k-1}g\right).\end{aligned}$$

In what follows, we will consider the geometrical properties of the reachability cone $\mathcal{R}(F,g)$ of a reachable pair (F,g) defined as

$$\mathcal{R}(F,g) = \operatorname{cl}\left\{\sum_{k=1}^{\infty}\mathcal{R}_k(F,g)\right\} = \operatorname{cl}\left\{\operatorname{cone}\left(g, Fg, F^2g\dots\right)\right\}, \tag{1.2}$$

where the sum of two cones, as proved in [12, Theorem 3.8], coincides with the set of all finite nonnegative combinations of vectors belonging to the two cones.

*Received by the editors July 29, 2004; accepted for publication (in revised form) by P. Van Dooren October 18, 2005; published electronically April 21, 2006.
http://www.siam.org/journals/simax/28-2/61253.html
†Dipartimento di Informatica e Sistemistica "A. Ruberti," Università degli Studi di Roma "La Sapienza," Via Eudossiana 18, 00184 Roma, Italy (luca.benvenuti@uniroma1.it, lorenzo.farina@uniroma1.it).

The reachability set $\mathcal{R}(F,g) \subseteq \mathbb{R}^n$ is a convex cone contained in the subspace spanned by the vectors $g, Fg, \ldots, F^{n-1}g$. Without loss of generality, we will assume in the sequel only reachable pairs (F,g), that is, $\mathbb{R}^n$ is the smallest subspace containing $\mathcal{R}(F,g)$. Therefore, since the reachability set $\mathcal{R}(F,g)$ is a convex cone, it can be written as

$$\mathcal{R}(F,g) = \mathcal{S}(F,g) \oplus \mathcal{K}(F,g),$$

where $\mathcal{S}(F,g)$ is the [illegible] of $\mathcal{R}(F,g)$ (the largest subspace contained in $\mathcal{R}(F,g)$) and

$$\mathcal{K}(F,g) = \mathcal{R}(F,g) \bigcap \mathcal{S}(F,g)^{\perp}$$

is a proper cone contained in the subspace $\mathcal{S}(F,g)^{\perp}$ complementary to $\mathcal{S}(F,g)$ in $\mathbb{R}^n$ (see [12, p. 65]).

The problem of characterizing the geometrical properties of the reachable set $\mathcal{R}(F,g)$ of linear system has been studied by Evans and Murthy, and Son in [9, 16] for discrete-time systems and by Brammer, Saperstone and Yorke, and Ohta et al. in [6, 15, 11] for continuous-time systems. Evans and Murthy, and Brammer derived conditions for complete controllability, i.e., $\mathcal{R}(F,g) = \mathcal{S}(F,g) = \mathbb{R}^n$ for discrete and continuous-time, respectively. Ohta et al. provided a simple formula to evaluate the dimension of the largest reachable subspace, i.e., the dimension of $\mathcal{S}(F,g)$ for single-input continuous-time systems. Saperstone and Yorke, and Son consider complete controllability in the case of bounded inputs for continuous- and discrete-time systems, respectively. Results in the related area of controllability of positive systems can be found in [17, 7, 14] and in the references cited therein.

In this paper we deal with single-input discrete-time systems and provide a complete geometrical characterization of the reachable set $\mathcal{R}(F,g)$, i.e., of both $\mathcal{S}(F,g)$ and $\mathcal{K}(F,g)$. More precisely, we give the dimension of the largest reachable subspace $\mathcal{S}(F,g)$ (analogous to those found by Ohta et al. in [11] for the continuous-time case) and provide necessary and sufficient conditions for polyhedrality of $\mathcal{K}(F,g)$ in terms of eigenvalues location. Some preliminary results have appeared in [8]. Polyhedrality of $\mathcal{K}(F,g)$ is relevant in the positive realization problem and its applications (optical filters and charge routing networks design, hidden Markov modeling, ...) as shown in [4]. Moreover, polyhedrality of $\mathcal{K}(F,g)$ is related to reachability with nonnegative inputs of every state from the origin in a finite number of steps, as discussed at the end of this paper.

2. Definitions. A set $\mathcal{K} \subset \mathbb{R}^m$ is said to be a [illegible] provided that $\alpha\mathcal{K} \subseteq \mathcal{K}$ for all $\alpha \geq 0$. If a cone $\mathcal{K} \subseteq \mathbb{R}^m$ contains an open ball of $\mathbb{R}^m$, then it is said to be [illegible] and if $\mathcal{K} \cap \{-\mathcal{K}\} = \{0\}$, it is said to be [illegible]. A cone which is closed, convex, solid and pointed is said to be a [illegible]. A cone $\mathcal{K}$ is said to be [illegible] if it is expressible as the intersection of a finite family of closed half-spaces. The notation cone$(v_1, \ldots, v_M)$ indicates the convex cone consisting of all nonnegative linear combinations of vectors $v_1, \ldots, v_M$, with M possibly infinite.

Given a square matrix F, $p_F(\lambda)$ is its characteristic polynomial, σ_F denotes the set of its eigenvalues and $\deg \lambda_i$, with $\lambda_i \in \sigma_F$, is the size of the largest block containing λ_i in the Jordan canonical form of F. If the matrix F has at least one nonnegative real eigenvalue, then ω_F equals the maximal nonnegative real eigenvalue of F; otherwise $\omega_F = 0$. Using the above definitions, the set σ_F can be partitioned in

the following disjoint subsets:

$$\begin{aligned}
\sigma_F^{(1)} &= \{\lambda_i \in \sigma_F : |\lambda_i| > \omega_F\} \\
\sigma_F^{(2)} &= \{\lambda_i \in \sigma_F : |\lambda_i| = \omega_F \text{ and } \deg \lambda_i > \deg \omega_F\} \\
\sigma_F^{(3)} &= \{\lambda_i \in \sigma_F : |\lambda_i| = \omega_F \text{ and } \deg \lambda_i \leq \deg \omega_F\} \\
\sigma_F^{(4)} &= \{\lambda_i \in \sigma_F : |\lambda_i| < \omega_F\}
\end{aligned}$$

so that $\sigma_F := \sigma_F^{(0)} = \sigma_F^{(1)} \cup \sigma_F^{(2)} \cup \sigma_F^{(3)} \cup \sigma_F^{(4)}$. Moreover, given a set of eigenvalues $\sigma_F^{(k)}$, we define

$$\rho(\sigma_F^{(k)}) = \max_{\lambda_i \in \sigma_F^{(k)}} \{|\lambda_i|\}$$

and every eigenvalue $\lambda_i \in \sigma_F^{(k)}$ such that $|\lambda_i| = \rho(\sigma_F^{(k)})$ will be called a [illegible] of $\sigma_F^{(k)}$.

If F is nonderogatory,[1] then without loss of generality (w.l.o.g.) we can assume the matrix to be in following pseudo-Jordan form

$$F = \left(\begin{array}{cc|ccc} J\left(\sigma_F^{(1)}\right) & 0 & 0 & 0 & 0 \\ 0 & J'\left(\sigma_F^{(2)}\right) & * & 0 & 0 \\ \hline 0 & 0 & J''\left(\sigma_F^{(2)}\right) & 0 & 0 \\ 0 & 0 & 0 & J\left(\sigma_F^{(3)}\right) & 0 \\ 0 & 0 & 0 & 0 & J\left(\sigma_F^{(4)}\right) \end{array}\right) = \left(\begin{array}{cc} A' & * \\ 0 & A \end{array}\right), \quad g = \left(\begin{array}{c} b' \\ b \end{array}\right), \tag{2.1}$$

where

$$\begin{aligned}
J\left(\sigma_F^{(k)}\right) &= \operatorname{diag}_{\lambda_i \in \sigma_F^{(k)}} \left(J_{\deg \lambda_i}(\lambda_i)\right) \qquad k = 1, 3, 4 \\
J'\left(\sigma_F^{(2)}\right) &= \operatorname{diag}_{\lambda_i \in \sigma_F^{(2)}} \left(J_{\deg \lambda_i - \deg \omega_F}(\lambda_i)\right) \\
J''\left(\sigma_F^{(2)}\right) &= \operatorname{diag}_{\lambda_i \in \sigma_F^{(2)}} \left(J_{\deg \omega_F}(\lambda_i)\right)
\end{aligned}$$

and $J_k(\lambda)$ is a $k \times k$ upper triangular matrix of the form

$$J_k(\lambda) = \begin{pmatrix} \lambda & 1 & & & 0 \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ 0 & & & & \lambda \end{pmatrix}.$$

[1] A matrix is nonderogatory if its characteristic polynomial equals its minimal polynomial.

The dimension of the matrix A' is

$$\mu = \sum_{\lambda_i \in \sigma_F^{(1)}} \deg \lambda_i + \sum_{\lambda_i \in \sigma_F^{(2)}} (\deg \lambda_i - \deg \omega_F)$$

and that of A is

$$\chi = n - \mu = \sum_{\lambda_i \in \sigma_F^{(2)}} \deg \omega_F + \sum_{\lambda_i \in \sigma_F^{(3)} \cup \sigma_F^{(4)}} \deg \lambda_i,$$

where summation over the empty set is considered to be zero.

3. Main results. As stated in the introduction, we begin this section by presenting a result which provides the dimension of the largest reachable subspace $\mathcal{S}(F, g)$ (analogous to those found by Ohta et al. in [11] for the continuous-time case) and how the cone $\mathcal{K}(F, g)$ can be generated.

THEOREM 3.1. (F, g) $\mathcal{S}(F, g)$

$$\mu = \sum_{\lambda_i \in \sigma_F^{(1)}} \deg \lambda_i + \sum_{\lambda_i \in \sigma_F^{(2)}} (\deg \lambda_i - \deg \omega_F).$$

$$\mathcal{K}(F, g) = \mathrm{cl}\left\{\mathrm{cone}\left(b, Ab, A^2 b, \dots\right)\right\}$$

The proof is the discrete-time counterpart of that contained in [11] for continuous-time systems and can be found in [3]. □

It is worth stating the following corollaries which directly follow from the above theorem and characterize the two special cases of $\mathcal{R}(F, g) = \mathcal{S}(F, g) = \mathbb{R}^n$ and $\mathcal{R}(F, g) = \mathcal{K}(F, g)$.

COROLLARY 3.2 (see [9]). (F, g) $\mathcal{R}(F, g) = \mathbb{R}^n$ $\mu = n$ F $F = A'$ $g = b'$

COROLLARY 3.3. (F, g) $\mathcal{R}(F, g)$ $\mu = 0$

$$\sigma_F^{(1)} \bigcup \sigma_F^{(2)} = \varnothing,$$

$\rho(\sigma_F) \in \sigma_F$ $\deg \rho(\sigma_F) \geq \deg \lambda_i$ λ_i $|\lambda_i| = \rho(\sigma_F)$ $F = A$ $g = b$

Secondly, we present hereafter a Lemma which provides conditions for polyhedrality of $\mathcal{K}(F, g)$. Define

$$\mathcal{K}_i(A, b) := \mathrm{cone}\left(b, Ab, A^2 b, \dots, A^{i-1} b\right) = \mathcal{R}_i(F, g) \bigcap \mathcal{S}(F, g)^{\perp}$$

and

$$\hat{\mathcal{K}}(A, b) := \sum_{i=1}^{\infty} \mathcal{K}_i(A, b) = \mathrm{cone}\left(b, Ab, A^2 b, \dots\right)$$

so that we have

$$\mathcal{K}(F, g) = \mathrm{cl}\, \hat{\mathcal{K}}(A, b) =: \mathcal{K}(A, b). \tag{3.1}$$

Moreover, by definition,

$$\mathcal{K}_1(A,b) \subseteq \mathcal{K}_2(A,b) \subseteq \mathcal{K}_3(A,b) \subseteq \dots$$

and if $\mathcal{K}_N(A,b) = \mathcal{K}_{N+1}(A,b)$, then $\mathcal{K}(F,g) = \mathcal{K}_N(A,b) = \mathcal{K}_i(A,b)\ \forall i \geq N$.

First note that if $\mathcal{K}(F,g) \neq \{0\}$ and $\omega_F = 0$ then the matrix A is nilpotent and $A^\chi = 0$. Hence

$$\mathcal{K}(F,g) = \hat{\mathcal{K}}(A,b) = \mathcal{K}_\chi(A,b) = \operatorname{cone}\left(b, Ab, \dots, A^{\chi-1}b\right)$$

is polyhedral. Consequently, w.l.o.g. in the following we will assume $\omega_F > 0$.

LEMMA 3.4. [illegible] (F,g) [illegible] $\omega_F > 0$ [illegible] $\mathcal{K}(F,g)$ [illegible] $\mathcal{S}(F,g)^\perp$ [illegible] r [illegible]

$$\lim_{k\to\infty} \frac{A^{rk+h}b}{\|A^{rk+h}b\|} = v_\infty^{(h)} \neq 0 \qquad h = 0, \dots, r-1$$

[illegible] $v_\infty^{(i)} \neq v_\infty^{(j)}$ [illegible] $i \neq j$ [illegible] N [illegible]

$$\mathcal{K}_{N+1}(A^r, A^h b) + \operatorname{cone}\left(v_\infty^{(h)}\right) = \mathcal{K}_N(A^r, A^h b) + \operatorname{cone}\left(v_\infty^{(h)}\right) \tag{3.2}$$

[illegible] $h = 0, \dots, r-1$

[illegible]. (Necessity) From polyhedrality and A-invariance of $\mathcal{K}(F,g)$, as proved in [1], it follows that the dominant eigenvalues of σ_A are among the k-roots of $\rho(\sigma_A)^k = \omega_F^k$ for some positive integer k. Moreover, $\rho(\sigma_A) = \omega_F \in \sigma_A$ and $\deg \omega_F \geq \deg \lambda_i$ for each λ_i such that $|\lambda_i| = \omega_F$ by definition of A, so that from Lemma 4 in [18], it follows that the limits

$$\lim_{k\to\infty} \frac{A^{rk+h}b}{\|A^{rk+h}b\|} = v_\infty^{(h)} \qquad h = 0, \dots, r-1 \tag{3.3}$$

exist with $v_\infty^{(i)} \neq v_\infty^{(j)}$ for $i \neq j$, they are nonzero and r is the minimal value for which the dominant eigenvalues with maximal degree are among the rth roots of ω_F^r. Note that

$$\operatorname{cone}\left(Av_\infty^{(h)}\right) = \operatorname{cone}\left(Av_\infty^{(h+1)\bmod r}\right).$$

Moreover, the vectors $v_\infty^{(i)}$'s are nonnegatively linearly independent. In fact, otherwise we would have

$$v_\infty^{(i)} = \sum_{\substack{k=0 \\ k\neq i}}^{r-1} \alpha_k v_\infty^{(k)} \qquad \alpha_k \geq 0$$

for, at least, one value of i. Consequently, multiplying both sides by A^h with $h = 0, \dots, r-1$, we would obtain

$$v_\infty^{(i+h)\bmod r} = \sum_{\substack{k=0 \\ k\neq (i+h)\bmod r}}^{r-1} \alpha_k v_\infty^{(k)} \qquad h = 0, \dots, r-1.$$

From the above equations, it directly follows that

$$v_\infty^{(0)} = v_\infty^{(1)} = \ldots = v_\infty^{(r-1)},$$

thus contradicting the fact that $v_\infty^{(i)} \neq v_\infty^{(j)}$ for $i \neq j$.

Consider now the cone $\hat{\mathcal{K}}(A^r, A^h b)+\text{cone}(v_\infty^{(h)})$. By definition, its extremal vectors are among the vectors $A^i b$ and the vector $v_\infty^{(h)}$ with i appropriate; let $v_{m,h} = A^i b = A^{ri_m+h} b$ be an extremal vector of the cone. We will prove that $v_{m,h}$ is also an extremal vector of $\mathcal{K}(A, b)$. Suppose that $v_{m,h}$ is not an extremal vector of $\mathcal{K}(A, b)$, that is,

$$v_{m,h} = \sum_{\substack{i=0 \\ i\neq h}}^{r-1} \sum_{j=1}^{n_i} \alpha_{j,i} v_{j,i} + \sum_{\substack{j=1 \\ j\neq m}}^{n_h} \alpha_{j,h} v_{j,h} + \sum_{k=0}^{r-1} \alpha_k v_\infty^{(k)} \qquad \alpha_{j,i}, \alpha_{j,h}, \alpha_k \geq 0, \tag{3.4}$$

where $v_{j,i}$ are the extremal vectors of $\mathcal{K}(A, b)$ which are also extremal vectors of $\hat{\mathcal{K}}(A^r, A^i b) + \text{cone}(v_\infty^{(i)})$. By applying $A^{rk}/\|A^{r(k+i_m)+h}\|$ with $k \to \infty$ to both sides of the above equation, one would obtain

$$\begin{aligned} \lim_{k\to\infty} \frac{A^{r(k+i_m)+h} b}{\|A^{r(k+i_m)+h} b\|} &= v_\infty^{(h)} \\ &= \sum_{\substack{i=0 \\ i\neq h}}^{r-1} \left(\sum_{j=1}^{n_i} \alpha'_{j,i} \right) v_\infty^{(i)} + \sum_{\substack{j=1 \\ j\neq m}}^{n_h} \alpha'_{j,h} v_\infty^{(h)} + \sum_{k=0}^{r-1} \alpha'_k v_\infty^{(k)} \end{aligned}$$

with $\alpha'_{j,i}, \alpha'_{j,h}, \alpha'_k \geq 0$ proportional to $\alpha_{j,i}, \alpha_{j,h}$ and α_k, respectively, up to a positive constant.

Since $v_\infty^{(h)} \notin \text{cone}(v_\infty^{(0)}, \ldots, v_\infty^{(h-1)}, v_\infty^{(h+1)}, \ldots, v_\infty^{(r-1)})$ then one would have $\sum_{j=1}^{n_i} \alpha'_{j,i} = 0$ for every $i \neq h$ and consequently $\alpha'_{j,i} = \alpha_{j,i} = 0$ for every $i \neq h$ and $j = 1, \ldots, n_i$. Moreover, $\alpha_k = 0$ for $k \neq h$. Then from (3.4), one obtains

$$v_{m,h} = \sum_{\substack{j=1 \\ j\neq m}}^{n_h} \alpha_{j,h} v_{j,h} + \alpha_h v_\infty^{(h)} \quad \alpha_{j,h}, \alpha_k \geq 0,$$

that is, $v_{m,h}$ would be a nonnegative linear combination of all the other extremal vectors of $\hat{\mathcal{K}}(A^r, A^h b) + \text{cone}(v_\infty^{(h)})$ which contradicts the fact that $v_{m,h}$ is an extremal vector.

Since $\mathcal{K}(F, g)$ is a polyhedral cone by hypothesis, then it has a finite number of extremal vectors. Since each extremal vector $A^i b$ of the cones $\hat{\mathcal{K}}(A^r, A^h b) + \text{cone}(v_\infty^{(h)})$ is also an extremal vector of $\mathcal{K}(A, b)$, then the number of extremal vectors of $\hat{\mathcal{K}}(A^r, A^h b) + \text{cone}(v_\infty^{(h)})$ is also finite. Consequently, there exists a finite value N_h for which

$$\mathcal{K}_{N_h+1}(A^r, A^h b) + \text{cone}\left(v_\infty^{(h)}\right) = \mathcal{K}_{N_h}(A^r, A^h b) + \text{cone}\left(v_\infty^{(h)}\right)$$

holds for $h = 0, \ldots, r-1$ so that (3.2) follows by taking $N = \max_h N_h$.

(Sufficiency) If there exists a positive integer r such that the following limits

$$\lim_{k\to\infty} \frac{A^{rk+h} b}{\|A^{rk+h} b\|} = v_\infty^{(h)} \neq 0 \qquad h = 0, \ldots, r-1$$

exist with $v_\infty^{(i)} \neq v_\infty^{(j)}$ for $i \neq j$, and there exists a value N such that (3.2) holds for every $h = 0, \dots, r-1$, then from the definition of $\mathcal{K}(F, g)$ it follows that

$$\begin{aligned}\mathcal{K}(F,g) &= \mathrm{cl}\left\{\sum_{i=1}^{\infty} \mathcal{K}_i(A,b)\right\} \\ &= \sum_{i=1}^{\infty}\sum_{h=0}^{r-1} \mathcal{K}_i(A^r, A^h b) + \sum_{h=0}^{r-1} \mathrm{cone}\left(\lim_{i\to\infty} \frac{A^{ri+h}b}{\|A^{ri+h}b\|}\right) \\ &= \sum_{i=1}^{\infty}\sum_{h=0}^{r-1} \left(\mathcal{K}_i(A^r, A^h b) + \mathrm{cone}\left(v_\infty^{(h)}\right)\right).\end{aligned}$$

Moreover, from (3.2) it is immediate to check that

$$\sum_{i=1}^{\infty} \left(\mathcal{K}_i(A^r, A^h b) + \mathrm{cone}\left(v_\infty^{(h)}\right)\right) = \mathcal{K}_N(A^r, A^h b) + \mathrm{cone}\left(v_\infty^{(h)}\right)$$

so that

$$\mathcal{K}(F,g) = \sum_{h=0}^{r-1} \left(\mathcal{K}_N(A^r, A^h b) + \mathrm{cone}\left(v_\infty^{(h)}\right)\right).$$

Hence, the cone $\mathcal{K}(F, g)$ has a finite number of extremal vectors, so that it is polyhedral. □

1. In order to illustrate the previous theorem, consider the matrices

$$F = \mathrm{diag}\,(-2, 1, -1, -0.8), \qquad g = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

In this case we have

$$A = \mathrm{diag}\,(1, -1, -0.8), \qquad b = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

and, for $r = 2$ the following limits exist

$$\lim_{k\to\infty} \frac{A^{2k}b}{\|A^{2k}b\|} = v_\infty^{(0)} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \quad \lim_{k\to\infty} \frac{A^{2k+1}b}{\|A^{2k+1}b\|} = v_\infty^{(1)} = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$$

and for $N = 1$ equality (3.2) holds for $h = 0, 1$ as Figure 3.1 clearly shows. Moreover, the figure also makes clear that, as shown in the proof of Lemma 3.4, any extremal vector of the form $A^i b$ of $\hat{\mathcal{K}}\left(A^2, b\right) + v_\infty^{(0)}$ and of $\hat{\mathcal{K}}\left(A^2, Ab\right) + v_\infty^{(1)}$ is also an extremal vector of $\mathcal{K}(A, b)$.

1. The conditions of the previous Lemma may not hold if either the limits $v_\infty^{(h)}$ do not exist or condition (3.2) doesn't hold. As an example of the first possibility, consider a pair (F, g) such that

$$A = \begin{pmatrix} \cos\varphi & \sin\varphi & 0 \\ -\sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

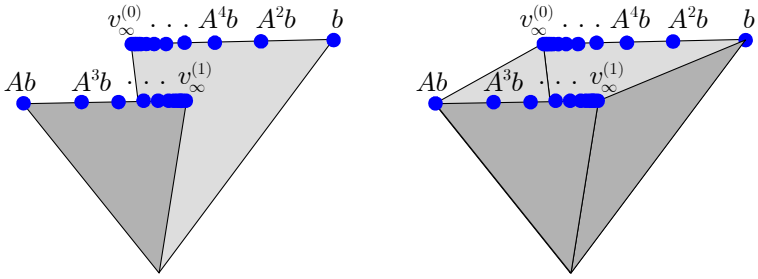


FIG. 3.1. *The cones* $\hat{\mathcal{K}}\left(A^2, b\right) + v_\infty^{(0)}$ *and* $\hat{\mathcal{K}}\left(A^2, Ab\right) + v_\infty^{(1)}$ *(left) and the cone* $\mathcal{K}(A, b)$ *(right).*

If φ/π is an irrational number, then the cone $\mathcal{K}(F, g)$ is not polyhedral and the $v_\infty^{(h)}$'s do not exist.

To show this, suppose there exists an invariant polyhedral proper cone and consider any of its extremal vectors v. Since the third component of v remains unchanged under A and the first two components are rotated by an angle φ in the (x_1, x_2) plane, then it is easily seen that, as k goes to infinity, the cone

$$\text{cone}\left(v, Av, A^2v, \ldots, A^kv\right)$$

is an ice cream cone, thus contradicting the polyhedrality hypothesis. Analogously, by taking $v = b$, we see that the limits $v_\infty^{(h)}$ do not exist.

As an example of the second possibility, consider a pair (F, g) such that

$$A = \text{diag}\left(\lambda_1, \lambda_2, \lambda_3\right), \qquad b = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

with $\lambda_1 > \lambda_2 > \lambda_3 > 0$. In this case $r = 1$ and

$$v_\infty^{(0)} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

If there exists a finite value N such that (3.2) holds, then A^Nb is a nonnegative linear combination of the vectors b, Ab, $\ldots$, $A^{N-1}b$ and $v_\infty^{(0)}$. Consequently, there exist nonnegative numbers α_0, α_1, $\ldots$, α_{N-1} such that

$$\lambda_i^N - \alpha_{N-1}\lambda_i^{N-1} - \ldots - \alpha_1\lambda_i - \alpha_0 = 0$$

holds for $i = 1, 2$. This is a contradiction since the above polynomial has only one positive real root from the Descartes rule of signs.

. . . . 2. The sum of two cones, as defined in the introduction, coincides with the set of all finite nonnegative combinations of vectors belonging to the two cones. Consequently, in condition (3.2), it makes no sense to "subtract" the term cone$(v_\infty^{(h)})$ from each side of the equation. In fact, (3.2) may hold even if

$$\mathcal{K}_{N+1}(A^r, A^hb) = \mathcal{K}_N(A^r, A^hb)$$

does not. To see this it suffices to consider a pair (F, g) such that

$$A = \mathrm{diag}\,(1, 0.8, -0.8)\,, \qquad b = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

In this case $r = 1$ and

$$v_\infty^{(0)} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

Consequently, straightforward calculations show that

$$A^2 b = 0.64\; b + 0.36\; v_\infty^{(0)},$$

that is, (3.2) holds with $N = 2$, while $A^2 b$ cannot be written as a nonnegative linear combination of b and Ab, i.e., $\mathcal{K}_2(A, b) \neq \mathcal{K}_3(A, b)$. Moreover, in this case, $\mathcal{K}_i(A, b) \neq \mathcal{K}_{i+1}(A, b)$ for any $i > 0$.

In what follows we provide the main result of the paper, that is, a spectral characterization of polyhedrality of the cone $\mathcal{K}(F, g)$.

THEOREM 3.5. [illegible] (F, g) [illegible] $\omega_F > 0$ [illegible] $\mathcal{K}(F, g)$ [illegible] $\mathcal{S}(F, g)^\perp$ [illegible]

1. $\deg \omega_F \leq 2$.
2. [illegible] $\sigma_F^{(2)} \bigcup \sigma_F^{(3)}$ [illegible] rth [illegible] ω_F^r [illegible] r.
3. [illegible] r [illegible] $\sigma_F^{(4)}$ [illegible] $2\pi/r$

[illegible]

1. $\deg \omega_F = 1$.
2. [illegible] $\sigma_F^{(4)}$ [illegible]
3. [illegible] $\sigma_F^{(2)} \bigcup \sigma_F^{(3)}$ [illegible] rth [illegible] ω_F^r [illegible] r [illegible] $\sigma_F^{(4)}$ [illegible] sth [illegible] $\rho(\sigma_F^{(4)})^s$ [illegible] s.
4. [illegible] r [illegible] s [illegible] $\sigma_F^{(4)}$ [illegible] $2\pi/\widetilde{r}$ [illegible] $\widetilde{r}$ [illegible] r [illegible] s

[illegible]. The proof is divided into three cases:

[illegible] $\deg \omega_F = 1$ and [illegible] hold;

[illegible] $\deg \omega_F = 2$ and [illegible] hold;

[illegible] $\deg \omega_F = 1$ and [illegible] hold.

[illegible] In order to prove polyhedrality of $\mathcal{K}(F, g)$, we will show that for each of the three cases considered, conditions of Lemma 3.4 hold. In particular, note that from [illegible] or [illegible] and from Lemma 4 in [18], it follows that the limits in Lemma 3.4 exist with $v_\infty^{(i)} \neq v_\infty^{(j)}$ for $i \neq j$ and they are nonzero. Hence, this remains true for all the three cases above considered. As a consequence, we will show that condition (3.2) holds for the three cases (see (3.7), (3.10) and (3.13)).

(Case 1) From condition , the dominant eigenvalues of σ_{A^r} are equal to ω_F^r with $\deg \omega_F^r = 1$. Hence, without loss of generality, we can assume

$$A^r = \left(\begin{array}{c|c} A_{11} & 0 \\ \hline 0 & A_{22} \end{array}\right) \qquad A^h b = \left(\begin{array}{c} b_1 \\ \hline b_2 \end{array}\right)$$

with

$$A_{11} = \begin{pmatrix} J''\left(\sigma_F^{(2)}\right) & 0 \\ 0 & J''\left(\sigma_F^{(3)}\right) \end{pmatrix}^r = \omega_F^r \cdot I$$

of appropriate dimension and $A_{22} = J(\sigma_F^{(4)})^r$. Moreover, $\rho(\sigma_{A_{22}}) = \rho(\sigma_F^{(4)})^r < \omega_F^r$ and, from condition , A_{22} has no real positive eigenvalues. Hence, by Lemma 3.11, there exists a monic polynomial $q(\lambda)$ such that

$$g(\lambda) = (\lambda - \omega_F^r) \cdot p_{A_{22}}(\lambda) \cdot q(\lambda) = \lambda^m - \alpha_{m-1}\lambda^{m-1} - \ldots - \alpha_1\lambda - \alpha_0$$

with m finite and $\alpha_k \geq 0$ for $k = 0, 1, \ldots, m-1$ and $\alpha_k > 0$ for some k.

Since $g(\omega_F^r) = 0$, then

$$g(A_{11}) = A_{11}^m - \alpha_{m-1}A_{11}^{m-1} - \ldots - \alpha_1 A_{11} - \alpha_0 I = 0. \tag{3.5}$$

Moreover, since $p_{A_{22}}(A_{22}) = 0$, then

$$g(A_{22}) = A_{22}^m - \alpha_{m-1}A_{22}^{m-1} - \ldots - \alpha_1 A_{22} - \alpha_0 I = 0. \tag{3.6}$$

From (3.5) and (3.6) it follows that

$$g(A^r) = (A^r)^m - \alpha_{m-1}(A^r)^{m-1} - \ldots - \alpha_1 A^r - \alpha_0 I = 0$$

so that

$$(A^r)^m A^h b = \alpha_{m-1}(A^r)^{m-1}A^h b + \ldots + \alpha_1 A^r A^h b + \alpha_0 A^h b, \tag{3.7}$$

that is, $\mathcal{K}_{N+1}(A^r, A^h b) = \mathcal{K}_N(A^r, A^h b)$ holds with $N = m$. Finally, in view of Lemma 3.4, $\mathcal{K}(F, g)$ is a polyhedral cone.

(Case 2) From condition and and in view of the definition of r, the matrix A^r can be written, without loss of generality up to a similarity transformation, as

$$A^r = \left(\begin{array}{c|c|c} A_{11} & 0 & 0 \\ \hline 0 & A_{22} & 0 \\ \hline 0 & 0 & A_{33} \end{array}\right) \qquad A^h b = \left(\begin{array}{c} b_1 \\ \hline b_2 \\ \hline b_3 \end{array}\right)$$

with

$$A_{11} = \left(\begin{array}{cc|c|c} \omega_F^r & 1 & & \\ 0 & \omega_F^r & & 0 \\ \hline & & \ddots & \\ \hline & 0 & & \begin{array}{cc} \omega_F^r & 1 \\ 0 & \omega_F^r \end{array} \end{array}\right) \qquad b_1 = \left(\begin{array}{c} b_1^{1,1} \\ b_1^{1,2} \\ \hline \vdots \\ \hline b_1^{d_1,1} \\ b_1^{d_1,2} \end{array}\right)$$

with A_{11} and $A_{22} = \omega_F^r \cdot I$ of appropriate dimensions and $A_{33} = J(\sigma_F^{(4)})^r$. Moreover, $\rho(\sigma_{A_{33}}) = \rho(\sigma_F^{(4)})^r < \omega_F^r$ and, from condition , A_{33} has no real positive eigenvalues. Hence, by Lemma 3.11, there exists a monic polynomial $q(\lambda)$ such that

$$g(\lambda) = (\lambda - \omega_F^r) \cdot p_{A_{33}}(\lambda) \cdot q(\lambda) = \lambda^m - \alpha_{m-1}\lambda^{m-1} - \ldots - \alpha_1\lambda - \alpha_0$$

with m finite and $\alpha_k \geq 0$ for $k = 0, 1, \ldots, m-1$ and $\alpha_k > 0$ for some k.

Since $g(\omega_F^r) = 0$, then

$$g(A_{22}) = A_{22}^m - \alpha_{m-1}A_{22}^{m-1} - \ldots - \alpha_1 A_{22} - \alpha_0 I = 0. \tag{3.8}$$

Moreover, since $p_{A_{33}}(A_{33}) = 0$, then

$$g(A_{33}) = A_{33}^m - \alpha_{m-1}A_{33}^{m-1} - \ldots - \alpha_1 A_{33} - \alpha_0 I = 0. \tag{3.9}$$

From (3.8) and (3.9) it follows that

$$\left(\begin{array}{c|c} A_{22} & 0 \\ \hline 0 & A_{33} \end{array}\right)^m = \alpha_{m-1}\left(\begin{array}{c|c} A_{22} & 0 \\ \hline 0 & A_{33} \end{array}\right)^{m-1} + \ldots + \alpha_1\left(\begin{array}{c|c} A_{22} & 0 \\ \hline 0 & A_{33} \end{array}\right) + \alpha_0 I.$$

Moreover, note that

$$\lim_{k\to\infty} \frac{A^{rk}A^h b}{\|A^{rk}A^h b\|} = \left(\begin{array}{c} v_1/\|v_1\| \\ \hline 0 \\ \hline 0 \end{array}\right) = v_\infty^{(h)} \quad \text{with } v_1 = \left(\begin{array}{c} b_1^{1,2} \\ \hline 0 \\ \hline \vdots \\ \hline b_1^{d_1,2} \\ 0 \end{array}\right).$$

Hence, the following holds

$$(A^r)^m A^h b = \alpha_m v_\infty^{(h)} + \alpha_{m-1}(A^r)^{m-1}A^h b + \ldots + \alpha_1 A^r A^h b + \alpha_0 A^h b \tag{3.10}$$

for any $h = 0, \ldots, r-1$ with

$$\alpha_m = m\frac{\|v_1\|}{\omega_F^r}\left(\omega_F^{mr} - \alpha_{m-1}\frac{m-1}{m}\omega_F^{(m-1)r} - \ldots - \alpha_1\frac{1}{m}\omega_F^r\right)$$

as one can easily check by substitution. Since

$$\begin{aligned}\alpha_m\frac{\omega_F^r}{m\,\|v_1\|} &= \omega_F^{mr} - \alpha_{m-1}\frac{m-1}{m}\omega_F^{(m-1)r} - \ldots - \alpha_1\frac{1}{m}\omega_F^r \\ &> \omega_F^{mr} - \alpha_{m-1}\omega_F^{(m-1)r} - \ldots - \alpha_1\omega_F^r - \alpha_0 = g\left(\omega_F^r\right) = 0,\end{aligned}$$

then $\alpha_m > 0$. Hence,

$$\mathcal{K}_{N+1}(A^r, A^h b) + \operatorname{cone}\left(v_\infty^{(h)}\right) = \mathcal{K}_N(A^r, A^h b) + \operatorname{cone}\left(v_\infty^{(h)}\right)$$

with $N = m$ holds for any $h = 0, \ldots, r-1$. Finally, in view of Lemma 3.4, $\mathcal{K}(A, b)$ is a polyhedral cone.

(Case 3) From conditions , , and in view of the definition of $\widetilde{r}$, the dominant eigenvalues of $\sigma_{A^{\widetilde{r}}}$ are equal to $\omega_F^{\widetilde{r}}$ with $\deg \omega_F^{\widetilde{r}} = 1$, and the subdominant eigenvalues are equal to $\rho(\sigma_F^{(4)})^{\widetilde{r}}$ with $\deg \rho(\sigma_F^{(4)})^{\widetilde{r}} = 1$. Hence, without loss of generality, we can write

$$A^{\widetilde{r}} = \left(\begin{array}{c|c|c} A_{11} & 0 & 0 \\ \hline 0 & A_{22} & 0 \\ \hline 0 & 0 & A_{33} \end{array}\right) \qquad A^h b = \left(\begin{array}{c} b_1 \\ \hline b_2 \\ \hline b_3 \end{array}\right)$$

with

$$A_{11} = \begin{pmatrix} J''\left(\sigma_F^{(2)}\right) & 0 \\ 0 & J''\left(\sigma_F^{(3)}\right) \end{pmatrix}^{\tilde{r}} = \omega_F^{\tilde{r}} \cdot I.$$

$A_{22} = \rho(\sigma_F^{(4)})^{\widetilde{r}} \cdot I$ of appropriate dimensions and

$$A_{33} = \operatorname*{diag}_{\substack{\lambda_i \in \sigma_F^{(4)} \\ |\lambda_i| \neq \rho(\sigma_F^{(4)})}} \left(J_{\deg \lambda_i}(\lambda_i)\right)^{\widetilde{r}}.$$

Consequently, $\rho(\sigma_{A_{33}}) < \rho(\sigma_F^{(4)})^{\widetilde{r}}$ and, from condition , A_{33} has no real positive eigenvalues. Hence, by Lemma 3.11, there exists a monic polynomial $q(\lambda)$ such that

$$g(\lambda) = (\lambda - \rho(\sigma_A^{(4)})^{\widetilde{r}}) \cdot p_{A_{33}}(\lambda) \cdot q(\lambda) = \lambda^m - \alpha_{m-1}\lambda^{m-1} - \ldots - \alpha_1\lambda - \alpha_0$$

with m finite and $\alpha_k \geq 0$ for $k = 0, 1, \ldots, m-1$.

Since $g(\rho(\sigma_A^{(4)})^{\widetilde{r}}) = 0$, then

$$g(A_{22}) = A_{22}^m - \alpha_{m-1}A_{22}^{m-1} - \ldots - \alpha_1 A_{22} - \alpha_0 I = 0. \tag{3.11}$$

Moreover, since $p_{A_{33}}(A_{33}) = 0$, then

$$g(A_{33}) = A_{33}^m - \alpha_{m-1}A_{33}^{m-1} - \ldots - \alpha_1 A_{33} - \alpha_0 I = 0. \tag{3.12}$$

From (3.11) and (3.12) it follows that

$$\left(\begin{array}{c|c} A_{22} & 0 \\ \hline 0 & A_{33} \end{array}\right)^m = \alpha_{m-1}\left(\begin{array}{c|c} A_{22} & 0 \\ \hline 0 & A_{33} \end{array}\right)^{m-1} + \ldots + \alpha_1\left(\begin{array}{c|c} A_{22} & 0 \\ \hline 0 & A_{33} \end{array}\right) + \alpha_0 I.$$

Moreover, note that

$$\lim_{k\to\infty} \frac{A^{\widetilde{r}k}A^h b}{\|A^{\widetilde{r}k}A^h b\|} = \left(\begin{array}{c} b_1/\|b_1\| \\ \hline 0 \\ \hline 0 \end{array}\right) = v_\infty^{(h)}$$

with $h = 0, \ldots, r-1$. Hence, the following holds

$$(A^{\widetilde{r}})^m A^h b = \alpha_m v_\infty^{(h)} + \alpha_{m-1}(A^{\widetilde{r}})^{m-1}A^h b + \ldots + \alpha_1(A^{\widetilde{r}})A^h b + \alpha_0 A^h b \tag{3.13}$$

for any $h = 0, \dots, r-1$ and with

$$\alpha_m = \|b_1\| \cdot \left(\omega_F^{m\widetilde{r}} - \alpha_{m-1}\omega_F^{(m-1)\widetilde{r}} - \dots - \alpha_1 \omega_F^{\widetilde{r}} - \alpha_0 \right) = \|b_1\| \cdot g(\omega_F^{\widetilde{r}}).$$

Since the polynomial $g(\lambda)$ has only one real root in $\rho(\sigma_A^{(4)})^{\widetilde{r}}$ and $\lim_{\lambda\to\infty} g(\lambda) = +\infty$, then $g(\omega_F^{\widetilde{r}}) > 0$, that is, $\alpha_m > 0$. Hence,

$$\mathcal{K}_{N+1}(A^{\tilde{r}}, A^h b) + \operatorname{cone}\left(v_\infty^{(h)}\right) = \mathcal{K}_N(A^{\tilde{r}}, A^h b) + \operatorname{cone}\left(v_\infty^{(h)}\right)$$

with $N = m$ holds for any $h = 0, \dots, r-1$. Finally, in view of Lemma 3.4, $\mathcal{K}(A, b)$ is a polyhedral cone.

We begin by proving that if $\mathcal{K}(A, b)$ is a polyhedral proper cone and $\deg \omega_F > 1$, then necessarily $\deg \omega_F = 2$. Assume then $\deg \omega_F = m \geq 2$ and let

$$A = \left(\begin{array}{c|ccc|c} \omega_F & 1 & 0 & \cdots & 0 & 0 \\ \hline 0 & & A_{22} & & & 0 \\ \hline 0 & & 0 & & & A_{33} \end{array}\right) \qquad A^h b = \left(\begin{array}{c} b_1 \\ \hline b_2 \\ \hline b_3 \end{array}\right)$$

with

$$A_{22} = \left(\begin{array}{c|c} J_{m-1}(\omega_F) & 0 \\ \hline 0 & A_{22}^* \end{array}\right), \quad A_{22}^* = \operatorname*{diag}_{\substack{\lambda_i \in \sigma_F^{(4)} \\ \lambda_i \neq 0}} \left(J_{\deg \lambda_i}(\lambda_i)\right)$$

and $\sigma_{A_{33}}$ contains the remaining eigenvalues of A.

From Lemma 3.4 it follows that there exists a value N such that (3.2) holds for any $h = 0, 1, \dots, r-1$. Since the vector $v_\infty^{(h)}$ is of the form

$$v_\infty^{(h)} = \left(\begin{array}{c} * \\ \hline 0 \\ \hline * \end{array}\right),$$

i.e., it has the entries corresponding to A_{22} equal to zero, then there exist nonnegative numbers $\alpha_0, \dots, \alpha_{N-1}$ such that the following holds:

$$A_{22}^{Nr+h} b_2 = \alpha_{N-1} A_{22}^{(N-1)r+h} b_2 + \dots + \alpha_1 A_{22}^{r+h} b_2 + \alpha_0 A_{22}^h b_2.$$

Let

$$P = \left[A_{22}^h b_2 \; A_{22}^{h+1} b_2 \; \dots A_{22}^{Nr-1+h} b_2\right] = A_{22}^h \left[b_2 \; A_{22} b_2 \; \dots A_{22}^{Nr-1} b_2\right];$$

then the matrix A_+ solution of the equation $A_{22}P = PA_+$ has a characteristic polynomial equal to

$$p_{A_+}(\lambda) = \lambda^{Nr} - \alpha_{N-1}\lambda^{(N-1)r} - \dots - \alpha_1 \lambda^r - \alpha_0.$$

Consequently, from the Descartes rule of signs, the polynomial $p_{A_+}(\lambda)$ has only one positive real root. Since P is full row-rank,[2] then from Lemma 3.10, $p_{A_{22}}(\lambda)$ divides $p_{A_+}(\lambda)$, so that also $p_{A_{22}}(\lambda)$ has only one positive real root, that is, ω_F, and consequently $\deg \omega_F = 2$.

[2]It is immediate to verify that condition (3.2) holds for any $M \geq N$ so that by choosing N in P large enough, full-row rankness of P follows from reachability of the pair (A_{22}, b_2) and from the fact that A_{22} has no zero eigenvalues.

We are now able to split the proof into two parts: $\deg \omega_F = 1$ or $\deg \omega_F = 2$. We begin with $\deg \omega_F = 2$, i.e., with the necessity of Case 2. From polyhedrality and A-invariance of $\mathcal{K}(F, g)$, as proved in [1], it follows that condition . holds. We will prove necessity of condition . by contradiction. Assume then that there exists an eigenvalue $\tilde{\lambda} \in \sigma_F^{(4)}$ such that $\tilde{\lambda} \neq 0$ and having an argument $\varphi = 2\pi m/r$ for some positive integer m.

Without loss of generality up to a similarity transformation, we can write

$$A = \left(\begin{array}{c|c|c|c|c} A_{11} & I & 0 & 0 & 0 \\ \hline 0 & A_{11} & 0 & 0 & 0 \\ \hline 0 & 0 & A_{22} & 0 & 0 \\ \hline 0 & 0 & 0 & A_{33} & 0 \\ \hline 0 & 0 & 0 & 0 & A_{44} \end{array}\right) \qquad b = \left(\begin{array}{c} b_1^1 \\ \hline b_1^2 \\ \hline b_2 \\ \hline b_3 \\ \hline b_4 \end{array}\right)$$

with

$$A_{11} = \operatorname*{diag}_{\substack{\lambda_i \in \sigma_F^{(2)} \\ \gamma_i \in \sigma_F^{(3)} \\ \deg \gamma_i = 2}} (\lambda_i), \qquad A_{22} = \operatorname*{diag}_{\substack{\lambda_i \in \sigma_F^{(3)} \\ \deg \lambda_i = 1}} (\lambda_i), \qquad A_{33} = \operatorname*{diag}_{\substack{\lambda_i \in \sigma_F^{(4)} \\ \lambda_i \neq 0}} (J_{\deg \lambda_i}(\lambda_i))$$

and $\sigma_{A_{44}}$ containing all the zero eigenvalues, if any.

From Lemma 3.4 it follows that there exists a finite value N such that (3.2) holds for every $h = 0, \ldots, r-1$. Since the vector $v_\infty^{(h)}$ has the form

$$v_\infty^{(h)} = \left(\begin{array}{c} * \\ \hline 0 \\ \hline 0 \\ \hline 0 \\ \hline 0 \end{array}\right),$$

then there exist nonnegative numbers $\alpha_0, \ldots, \alpha_{N-1}$ such that the following holds:

$$\hat{A}^{Nr+h}\hat{b} = \alpha_{N-1}\hat{A}^{(N-1)r+h}\hat{b} + \ldots + \alpha_1 \hat{A}^{r+h}\hat{b} + \alpha_0 \hat{A}^h \hat{b}$$

with

$$\hat{A} = \left(\begin{array}{c|c|c} A_{11} & 0 & 0 \\ \hline 0 & A_{22} & 0 \\ \hline 0 & 0 & A_{33} \end{array}\right) \qquad \hat{b} = \left(\begin{array}{c} b_1^2 \\ \hline b_2 \\ \hline b_3 \end{array}\right).$$

Let

$$P = \left[\hat{A}^h\hat{b}\ \hat{A}^{h+1}\hat{b}\ \ldots \hat{A}^{Nr-1+h}\hat{b}\right] = \hat{A}^h\left[\hat{b}\ \hat{A}\hat{b}\ \ldots \hat{A}^{Nr-1}\hat{b}\right],$$

then the matrix A_+ solution of the equation $\hat{A}P = PA_+$ has a characteristic polynomial equal to

$$p_{A_+}(\lambda) = \lambda^{Nr} - \alpha_{N-1}\lambda^{(N-1)r} - \ldots - \alpha_1\lambda^r - \alpha_0.$$

Consequently, from the Descartes rule of signs, the polynomial $p_{A_+}(\lambda)$ has only one positive real root. Since P is full row-rank (see footnote 2), then from Lemma 3.10,

$p_{\hat{A}}(\lambda)$ divides $p_{A_+}(\lambda)$ so that, from the Frobenius theorem (see [5, Theorem 2.20[3]]), the whole spectrum of A_+ goes into itself under any rotation of the complex plane by $2\pi m/r^+$, with r^+ multiple of r. Since there is a $\widetilde{\lambda}$ with an argument $\varphi = 2\pi m/r$ for some positive integer m, then the polynomial $p_{A_+}(\lambda)$ necessarily has a real positive root $|\widetilde{\lambda}|$ other than ω_F, which is a contradiction since $p_{A_+}(\lambda)$ has only one positive real root.

This concludes the proof of necessity of case 2.

We continue assuming $\deg \omega_F = 1$. In view of Lemma 3.4, we consider two possibilities:

(a) $v_\infty^{(h)} \in \mathcal{K}_N(A^r, A^h b)$ for every $h = 0, \dots, r-1$ so that (3.2) reduces to

$$\mathcal{K}_{N+1}(A^r, A^h b) = \mathcal{K}_N(A^r, A^h b)$$

or

(b) $v_\infty^{(h)} \notin \mathcal{K}_N(A^r, A^h b)$ for some h's, so that

$$\mathcal{K}_{N+1}(A^r, A^h b) + \operatorname{cone}\left(v_\infty^{(h)}\right) = \mathcal{K}_N(A^r, A^h b) + \operatorname{cone}\left(v_\infty^{(h)}\right).$$

In what follows we will prove that in possibility (a) necessity of case 1 holds while, necessity of case 3 does in possibility (b).

We begin with possibility (a). From polyhedrality and A-invariance of $\mathcal{K}(F, g)$, as proved in [1], it follows that condition holds. We will prove necessity of condition by contradiction. Assume then that there exists an eigenvalue $\tilde{\lambda} \in \sigma_F^{(4)}$ such that $\tilde{\lambda} \neq 0$ and having an argument $\varphi = 2\pi m/r$ for some positive integer m.

Without loss of generality up to a similarity transformation, we can write

$$A = \left(\begin{array}{c|c|c} A_{11} & 0 & 0 \\ \hline 0 & A_{22} & 0 \\ \hline 0 & 0 & A_{33} \end{array}\right) \qquad b = \left(\begin{array}{c} b_1 \\ \hline b_2 \\ \hline b_3 \end{array}\right) \tag{3.14}$$

with

$$A_{11} = \operatorname*{diag}_{\lambda_i \in \sigma_F^{(2)} \cup \sigma_F^{(3)}} (\lambda_i), \quad A_{22} = \operatorname*{diag}_{\substack{\lambda_i \in \sigma_F^{(4)} \\ \lambda_i \neq 0}} \left(J_{\deg \lambda_i}(\lambda_i)\right)$$

and $\sigma_{A_{33}}$ containing all the zero eigenvalues, if any.

[3]Since $\rho(\sigma_{A_+}) \geq \omega_F > 0$, then at least one α_k must be positive, so that there exists $k \geq 0$ such that $\alpha_0 = ... = \alpha_{k-1} = 0$ and $\alpha_k > 0$. Consequently, A_+ can be written as

$$A_+ = \left(\begin{array}{c|c} A_{11}^+ & 0 \\ \hline A_{21}^+ & A_{22}^+ \end{array}\right)$$

with

$$A_{11}^+ = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ & \ddots & & \vdots \\ 0 & \dots & 1 & 0 \end{pmatrix}, \qquad A_{22}^+ = \begin{pmatrix} 0 & 0 & \dots & 0 & \alpha_k \\ 1 & 0 & \dots & 0 & \alpha_{k+1} \\ 0 & 1 & & 0 & \alpha_{k+2} \\ \vdots & & \ddots & & \vdots \\ 0 & \dots & 0 & 1 & \alpha_{N-1} \end{pmatrix}.$$

Since $\alpha_k > 0$, then A_{22}^+ is an irreducible nonnegative matrix (see Theorem 1.3 (d) in [5, p. 27]). Hence, the Frobenius theorem applies to the submatrix A_{22}^+ and, w.r.t. the rotational symmetry, to the whole matrix A_+ since the spectrum of A_{11}^+ contains only zero eigenvalues.

From the assumption considered in possibility (a), there exist nonnegative numbers $\alpha_0, \ldots, \alpha_{N-1}$ such that the following holds:

$$\hat{A}^{Nr+h}\hat{b} = \alpha_{N-1}\hat{A}^{(N-1)r+h}\hat{b} + \ldots + \alpha_1\hat{A}^{r+h}\hat{b} + \alpha_0\hat{A}^h\hat{b}$$

with

$$\hat{A} = \left(\begin{array}{c|c} A_{11} & 0 \\ \hline 0 & A_{22} \end{array}\right) \qquad \hat{b} = \left(\begin{array}{c} b_1 \\ \hline b_2 \end{array}\right).$$

Let

$$P = \left[\hat{A}^h\hat{b}\ \hat{A}^{h+1}\hat{b}\ \ldots \hat{A}^{Nr-1+h}\hat{b}\right] = \hat{A}^h\left[\hat{b}\ \hat{A}\hat{b}\ \ldots \hat{A}^{Nr-1}\hat{b}\right],$$

then the matrix A_+ solution of the equation $\hat{A}P = PA_+$ has a characteristic polynomial equal to

$$p_{A_+}(\lambda) = \lambda^{Nr} - \alpha_{N-1}\lambda^{(N-1)r} - \ldots - \alpha_1\lambda^r - \alpha_0.$$

Consequently, from the Descartes rule of signs, the polynomial $p_{A_+}(\lambda)$ has only one positive real root. Since P is full row-rank (see footnote 2), then from Lemma 3.10, $p_{\hat{A}}(\lambda)$ divides $p_{A_+}(\lambda)$ so that, from the Frobenius theorem (see [5, Theorem 2.20] and footnote 3), the whole spectrum of A_+ goes into itself under any rotation of the complex plane by $2\pi m/r^+$, with r^+ multiple of r. Since there is a $\widetilde{\lambda}$ with an argument $\varphi = 2\pi m/r$ for some positive integer m, then the polynomial $p_{A_+}(\lambda)$ necessarily has a real positive root $|\widetilde{\lambda}|$ other than ω_F, which is a contradiction since $p_{A_+}(\lambda)$ has only one positive real root. This concludes the proof of necessity of case 1.

Let's tackle now possibility (b). From polyhedrality and A-invariance of $\mathcal{K}(F, g)$, as proved in [1], it follows that the first statement in condition holds. We will prove necessity of the remaining conditions (namely, , the second statement in , and) under the assumption that condition does not hold. In fact, since in this subcase we have $\deg \omega_F = 1$ and the first statement in condition (which is the same as condition) holds, then if condition would hold, then also case 1 would. Hence—as proved in the sufficiency part of case 1—we would have $\mathcal{K}_{N+1}(A^r, A^h b) = \mathcal{K}_N(A^r, A^h b)$ for any $h = 0, \ldots, r-1$. This is a contradiction.

We assume then that there exists an eigenvalue $\tilde{\lambda} \in \sigma_F^{(4)}$ such that $\tilde{\lambda} \neq 0$ and having an argument $\varphi = 2\pi m/r$ for some positive integer m.

Without loss of generality up to a similarity transformation, we can write the matrices A and b as in (3.14).

In this case, there exists a finite value N such that (3.2) holds for every $h = 0, \ldots, r-1$. Since the vector $v_\infty^{(h)}$ has the form

$$v_\infty^{(h)} = \left(\begin{array}{c} * \\ \hline 0 \\ \hline 0 \end{array}\right),$$

then there exist nonnegative numbers $\alpha_0, \ldots, \alpha_{N-1}$ such that the following holds:

$$A_{22}^{Nr+h}b_2 = \alpha_{N-1}A_{22}^{(N-1)r+h}b_2 + \ldots + \alpha_1 A_{22}^{r+h}b_2 + \alpha_0 A_{22}^h b_2.$$

Let

$$P = \left[A_{22}^h b_2\ A_{22}^{h+1}b_2\ \ldots A_{22}^{Nr-1+h}b_2\right] = A_{22}^h\left[b_2\ A_{22}b_2\ \ldots A_{22}^{Nr-1}b_2\right],$$

then the matrix A_+ solution of the equation $A_{22}P = PA_+$ has a characteristic polynomial equal to

$$p_{A_+}(\lambda) = \lambda^{Nr} - \alpha_{N-1}\lambda^{(N-1)r} - \ldots - \alpha_1\lambda^r - \alpha_0.$$

Consequently, from the Descartes rule of signs, the polynomial $p_{A_+}(\lambda)$ has only one positive real root, that is, $\rho\left(\sigma_{A_+}\right)$. Moreover, the polynomial

$$\hat{p}_{A_+}(\lambda) := \lambda^N - \alpha_{N-1}\lambda^{N-1} - \ldots - \alpha_1\lambda - \alpha_0$$

has only one positive real root in $\rho(\sigma_{A_+})^r$. Since P is full row-rank (see footnote 2), then from Lemma 3.10, $p_{A_{22}}(\lambda)$ divides $p_{A_+}(\lambda)$. Since by assumption $p_{A_{22}}(\tilde{\lambda}) = 0$, then $\hat{p}_{A_+}(\tilde{\lambda}^r) = 0$ where $\tilde{\lambda}^r$ is positive real, so that $\rho(\sigma_{A_+}) = |\tilde{\lambda}|$. Finally, since $\rho(\sigma_{A_+}) \geq \rho(\sigma_{A_{22}}) = \rho(\sigma_F^{(4)})$, then $|\tilde{\lambda}| = \rho(\sigma_F^{(4)})$. Hence, from the Frobenius theorem (see [5, Theorem 2.20] and footnote 3), condition and the second statement of condition hold. Necessity of condition will be proved by contradiction. Assume then that there exists an eigenvalue $\hat{\lambda} \in \sigma_{A_{22}} \subseteq \sigma_{A_+}$ such that $\hat{\lambda} \neq 0$, $|\hat{\lambda}| < \rho(\sigma_F^{(4)})$ and having an argument $\varphi = 2\pi m/\tilde{r}$ for some positive integer m. The whole spectrum of A_+ goes into itself under any rotation of the complex plane by $2\pi m/r^+$, with r^+ multiple of s. Moreover, since $p_{A_{22}}(\hat{\lambda}) = 0$, then $\hat{p}_{A_+}(\hat{\lambda}^r) = 0$ so that all the rth roots of $\hat{\lambda}^r$ are roots of $p_{A_+}(\lambda)$. Consequently, since $\tilde{r}$ is the least common multiple between r and s, then the polynomial $p_{A_+}(\lambda)$ would necessarily have a real positive root $|\hat{\lambda}|$ other then $\rho(\sigma_F^{(4)})$ which is a contradiction since from the Descartes rule of signs, the polynomial $p_{A_+}(\lambda)$ has only one positive real root.

This concludes the proof of necessity of case 3. □

, 2. In order to illustrate the above theorem, consider the following pair

$$F = \operatorname{diag}(1, \lambda_2, \lambda_3), \qquad g = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

with λ_2, λ_3 real and such that $|\lambda_3| < |\lambda_2| < 1$. Hence, $A = F$, $b = g$, $\omega_F = 1$ and $\deg \omega_F = 1$. Furthermore, condition holds and condition holds with $r = 1$. Lastly, conditions and hold and condition holds with $r = 1$ and $s \leq 2$.

When $\lambda_2 = -0.9$ and $\lambda_3 = -0.6$, then also condition holds. Moreover, as expected, condition fails since the dominant eigenvalue of $\sigma_F^{(4)}$ is $\lambda_2 = -0.9$ so that $s = 2$, $\tilde{r} = 2$ and $\lambda_3 = -0.6$ has a phase equal to π. Hence, the cone $\mathcal{K}(A, b)$ is polyhedral as shown on the left-hand side of Figure 3.2.

When $\lambda_2 = 0.9$ and $\lambda_3 = -0.8$, then condition fails since the eigenvalue $\lambda_2 = 0.9$ has a phase equal to 2π. By contrast, condition holds with $s = 1$ so that $\tilde{r} = 1$ and consequently condition b4 holds since $\lambda_3 = -0.8$ has a phase which is not an integer multiple of 2π. Hence, the cone $\mathcal{K}(A, b)$ is polyhedral as shown in the middle picture of Figure 3.2.

Finally, when $\lambda_2 = -0.9$ and $\lambda_3 = 0.8$, then condition fails since the eigenvalue $\lambda_3 = 0.8$ has a phase equal to 2π. Moreover, also condition fails since the dominant eigenvalue of $\sigma_F^{(4)}$ is $\lambda_2 = -0.9$ so that $s = 2$, $\tilde{r} = 2$ and $\lambda_3 = 0.8$ has a phase which is an integer multiple of π. Hence, the cone $\mathcal{K}(A, b)$ is not polyhedral as shown in the right-hand side of Figure 3.2.

Note that, when $\chi = 2$ the cone $\mathcal{K}(F, g)$ is always polyhedral since obviously any cone in $\mathbb{R}^2$ is polyhedral. In fact, in this case, the conditions of the theorem are always met as one can easily check.

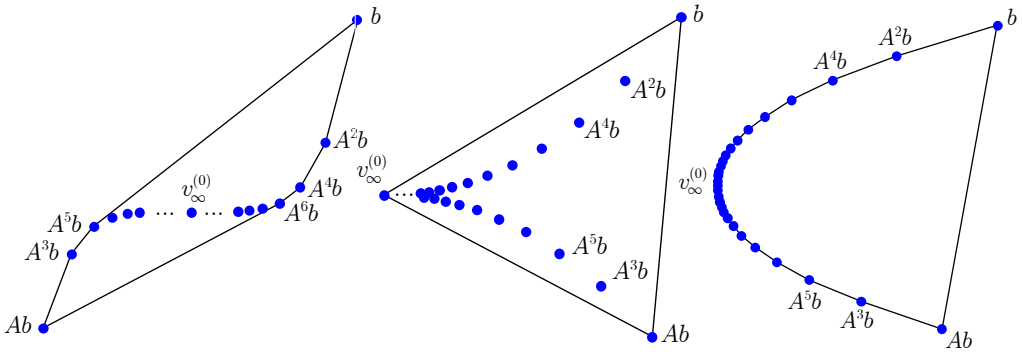


FIG. 3.2. *Planar section of the cone $\mathcal{K}(A,b)$ with the plane $x_1 = 0$ for the three cases considered in Example* 2.

Moreover, from the proof of the previous theorem, immediately follows the next corollaries which provide a geometrical and the corresponding spectral characterizations of systems for which the cone $\mathcal{K}(F,g)$ is reachable in a finite number of steps. This property is clearly equivalent to requiring polyhedrality of $\hat{\mathcal{K}}(A,b)$, or that the condition $\mathcal{K}(F,g) = \hat{\mathcal{K}}(A,b)$ holds.

COROLLARY 3.6. (F,g) $\omega_F > 0$ $\hat{\mathcal{K}}(A,b)$ $\mathcal{S}(F,g)^\perp$ N

$$\mathcal{K}_{N+1}(A,b) = \mathcal{K}_N(A,b).$$

COROLLARY 3.7. (F,g) $\omega_F > 0$ $\hat{\mathcal{K}}(A,b)$ $\mathcal{S}(F,g)^\perp$

1. $\deg \omega_F = 1$.
2. $\sigma_F^{(2)} \bigcup \sigma_F^{(3)}$ rth ω_F^r r.
3. r $\sigma_F^{(4)}$ $2\pi/r$

Moreover, the following theorem characterizes the case in which cone $\mathcal{K}(F,g)$ is reachable in at most n steps, that is, $\mathcal{K}(F,g)$ is simplicial.

THEOREM 3.8. (F,g) $\chi > 0$ $\mathcal{K}(F,g)$ $\mathcal{S}(F,g)^\perp$ χ

$$p(\lambda) := \prod_{\lambda_i \in \sigma_A^{(3)} \cup \sigma_A^{(4)}} (\lambda - \lambda_i)$$

Note that characterizing states reachable in an infinite number of steps is not trivial. In fact, as shown in the middle picture of Figure 3.2, there may well be states reachable in an infinite number of steps even if $\mathcal{K}(F,g)$ is polyhedral. It would be interesting to fully characterize the set of states reachable in an infinite number of steps which is, to the the best of our knowledge, an open question.

3. As a concluding remark we note that, in the multiple-input case, the situation is far more complicated and polyhedrality of $\mathcal{K}(F,G)$, in general, does not

depend only on the spectrum of F, as in the single-input case considered in this paper. In fact, in this case, the reachable set $\mathcal{R}(F,G)$ is

$$\mathcal{R}(F,G) = \sum_{i=1}^{m} \mathrm{cl}\left\{\mathrm{cone}\left(g_i, Fg_i, F^2 g_i \dots\right)\right\} = \mathcal{S}(F,G) \oplus \sum_{i=1}^{m} \mathcal{K}(F,g_i),$$

where m is the number of inputs and g_i is the ith column of G. It is immediate to realize that $\mathcal{K}(F,G)$ may be polyhedral even if the cones $\mathcal{K}(F,g_i)$ are not such.

Appendix.

LEMMA 3.9. [illegible] (A,b) [illegible] $p_A(\lambda)$ [illegible] A [illegible]. From reachability follows that A is similar to the companion matrix of its characteristic polynomial $p_A(\lambda)$ so that, from Theorem 3.3.15 in [10, p. 147], the theorem is proved. □

LEMMA 3.10. [illegible]

$$AP = PB \qquad A \in R^{n\times n}, \qquad P \in R^{n\times p}, \qquad B \in R^{p\times p}, \qquad p \geq n$$

[illegible] $\mathrm{rank}(P) = n$ [illegible] $p_A(\lambda)$ [illegible] $p_B(\lambda)$ [illegible]. First we note that there is no loss of generality in assuming that the columns of P can be freely rearranged. In fact,

$$A(PT) = (PT)T^{-1}BT = (PT)C,$$

where $C = TBT^{-1}$ is similar to B, i.e., $p_B(\lambda) = p_C(\lambda)$ being T a permutation matrix and, as such, invertible. Then

$$AQ = QC \qquad Q = PT = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}$$

with $Q_1 \in R^{n\times n}$ full rank (invertible).

Moreover, let S be the matrix such that $J = S^{-1}CS$ is the real Jordan canonical form of C. Then we can write

$$AQS = QSS^{-1}CS \rightarrow A(QS) = (QS)J$$

with

$$QS = \begin{pmatrix} Q_1 S & Q_2 S \end{pmatrix},$$

where, in particular, $Q_1 S$ is full rank (invertible) being such both Q_1 and S. Consequently, we can write

$$A\begin{pmatrix} Q_1 S & Q_2 S \end{pmatrix} = \begin{pmatrix} Q_1 S & Q_2 S \end{pmatrix} \begin{pmatrix} J_1 & * \\ 0 & J_2 \end{pmatrix}$$

and, in particular,

$$AQ_1 S = Q_1 S J_1 \rightarrow A = (Q_1 S) J_1 (Q_1 S)^{-1}$$

so that

$$p_A(\lambda) = p_{J_1}(\lambda).$$

The theorem is proved by noting that $p_B(\lambda) = p_C(\lambda) = p_J(\lambda) = p_{J_1}(\lambda) p_{J_2}(\lambda)$. □

LEMMA 3.11 (see [13]). [illegible] $p(\lambda)$ [illegible] $\rho > \max\{|\lambda_i| : p(\lambda_i) = 0\}$ [illegible] $q_\rho(\lambda)$ [illegible]

$$(\lambda - \rho) \cdot p(\lambda) \cdot q_\rho(\lambda) = \lambda^N - \alpha_{N-1}\lambda^{N-1} - \ldots - \alpha_1\lambda - \alpha_0$$

[illegible] N [illegible] $\alpha_k \geq 0$ [illegible] $k = 0, 1, \ldots, N-1$

REFERENCES

[1] G. P. BARKER AND R. E. L. TURNER, *Some observations on the spectra of cone preserving maps*, Linear Algebra Appl., 6 (1973), pp. 149–153.

[2] L. BENVENUTI, A. DE SANTIS, AND L. FARINA, eds., *Positive Systems*, in Proceedings of the First Multidisciplinary International Symposium on Positive Systems: Theory and Applications, Rome, 2003.

[3] L. BENVENUTI AND L. FARINA, *The geometry of the reachability cone for linear discrete–time systems*, Technical report 1, Dipartimento di Informatica e Sistemistica "A. Ruberti," Università degli Studi di Roma "La Sapienza," Rome, 2004.

[4] L. BENVENUTI AND L. FARINA, *A tutorial on the positive realization problem*, IEEE Trans. Automat. Control, 49 (2004), pp. 651–664.

[5] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[6] R. F. BRAMMER, *Controllability in linear autonomous systems with positive controllers*, SIAM J. Control Optim., 10 (1972), pp. 339–353.

[7] R. BRU, S. ROMERO, AND E. SANCHEZ, *Canonical forms of reachability and controllability of positive discrete-time control systems*, Linear Algebra Appl., 310 (2000), pp. 49–310.

[8] L. FARINA AND L. BENVENUTI, *Polyhedral reachable set with positive control*, Math. Control Signals Systems, 10 (1997), pp. 364–380.

[9] M. E. EVANS AND D. N. P. MURTHY, *Controllability of discrete–time systems with positive controls*, IEEE Trans. Automat. Control, 22 (1977), pp. 942–945.

[10] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[11] Y. OHTA, H. MAEDA, AND S. KODAMA, *Reachability, observability and realizability of continuous-time positive systems*, SIAM J. Control Optim., 22 (1984), pp. 171–180.

[12] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, New Jersey, 1970.

[13] M. ROITMAN AND Z. RUBINSTEIN, *On linear recursion with nonnegative coefficients*, Linear Algebra Appl., 167 (1992), pp. 151–155.

[14] V. G. RUMCHEV AND D. J. G. JAMES, *Controllability of positive linear discrete-time systems*, Int. J. Control, 50 (1989), pp. 845–857.

[15] S. H. SAPERSTONE AND J. A. YORKE, *Controllability of linear oscillatory systems using positive controls*, SIAM J. Control Optim., 9 (1971), pp. 253–262.

[16] N. K. SON, *Controllability of linear discrete–time systems with constrained controls in Banach spaces*, Control Cybernet., 10 (1981), pp. 5–16.

[17] M. E. VALCHER, *Controllability and reachability criteria for discrete–time positive systems*, Int. J. Control, 65 (1996), pp. 511–536.

[18] C. WENDE AND L. DAMING, *Nonnegative realizations of systems over nonnegative quasi-fields*, Acta Math. Sinica, 5 (1989), pp. 252–261.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 2, pp. 326–347

STRUCTURED CONDITION NUMBERS FOR INVARIANT SUBSPACES*

RALPH BYERS† AND DANIEL KRESSNER‡

Abstract. Invariant subspaces of structured matrices are sometimes better conditioned with respect to structured perturbations than with respect to general perturbations. Sometimes they are not. This paper proposes an appropriate condition number $c_{\mathbb{S}}$, for invariant subspaces subject to structured perturbations. Several examples compare $c_{\mathbb{S}}$ with the unstructured condition number. The examples include block cyclic, Hamiltonian, and orthogonal matrices. This approach extends naturally to structured generalized eigenvalue problems such as palindromic matrix pencils.

Key words. structured eigenvalue problem, invariant subspace, perturbation theory, condition number, deflating subspace, block cyclic, Hamiltonian, orthogonal, palindromic

AMS subject classifications. 65F15, 65F35

DOI. 10.1137/050637601

1. Introduction.

An invariant subspace $\mathcal{X} \subseteq \mathbb{C}^n$ of a matrix $A \in \mathbb{C}^{n\times n}$ is a linear subspace that stays invariant under the action of A, i.e., $Ax \in \mathcal{X}$ for all $x \in \mathcal{X}$. The computation of such an invariant subspace to solve a real-world problem is virtually always affected by some error, e.g., due to the limitations of finite-precision arithmetic. Instead of $\mathcal{X}$, it is usually the case that only a (hopefully nearby) invariant subspace $\hat{\mathcal{X}}$ of a slightly perturbed matrix $A+E$ is computed, where E represents measurement, modeling, discretization, or roundoff errors. It is therefore important to analyze the influence of perturbations in the entries of A on the accuracy of the invariant subspace $\mathcal{X}$. Stewart [33, 35] developed such a perturbation analysis, yielding a measure on the worst-case sensitivity of $\mathcal{X}$. This measure, the condition number $c(\mathcal{X})$, is most appropriate if the only information available on E is that its norm is below a certain perturbation threshold ϵ. Often, however, more information is available, i.e., it is known that the perturbation E preserves some structure of A. For example, if A is a real matrix, then it is reasonable to assume that E is also a real matrix. Also, for many classes of structured eigenvalue problems, such as Hamiltonian eigenvalue problems, it is more natural to study and analyze perturbations that respect the structure.

In this paper, we analyze the influence of structured perturbations $E : A + E \in \mathbb{S}$, where $\mathbb{S}$ is a linear matrix subspace or a smooth submanifold of $\mathbb{C}^{n\times n}$ or $\mathbb{R}^{n\times n}$. This will lead to the notion of a structured condition number $c_{\mathbb{S}}(\mathcal{X})$ for an invariant subspace $\mathcal{X}$. It occasionally happens that $c_{\mathbb{S}}(\mathcal{X}) \ll c(\mathcal{X})$, in which case the standard condition number $c(\mathcal{X})$ becomes an inappropriate measure on the actual worst-case sensitivity

*Received by the editors August 5, 2005; accepted for publication (in revised form) by N. J. Higham December 1, 2005; published electronically April 21, 2006.
http://www.siam.org/journals/simax/28-2/63760.html

†Department of Mathematics, University of Kansas, 405 Snow Hall, Lawrence, KS 66045 (byers@math.ku.edu). This author was partially supported by National Science Foundation awards 0098150 and 0112375 and by the Deutsche Forschungsgemeinschaft Research Center Mathematics for Key Technologies.

‡Department of Computing Science, Umeå University, Umeå 90187, Sweden (kressner@cs.umu.se) This author has been supported by a DFG Emmy Noether fellowship and in part by the Swedish Foundation for Strategic Research under the Frame Programme Grant A3 02:128.

of $\mathcal{X}$. An extreme example is provided by

$$A = \left[\begin{array}{cc|cc} 0 & -1-\alpha & 2 & 0 \\ 1+\alpha & 0 & 0 & 2 \\ \hline 0 & 0 & 0 & 1-\alpha \\ 0 & 0 & -1+\alpha & 0 \end{array}\right], \quad \mathcal{X} = \operatorname{span}\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \right\}, \tag{1.1}$$

where $\alpha \geq 0$ is considered to be tiny. While $c(\mathcal{X}) = \frac{1}{2\alpha}$, we will see that the structured condition number is given by $c_{\mathbb{S}}(\mathcal{X}) = 1/2$ if the set $\mathbb{S}$ of perturbed matrices is restricted to matrices of the form $A + E = \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ \hat{A}_{21} & \hat{A}_{22} \end{bmatrix}$ with $\hat{A}_{ij} = \begin{bmatrix} \beta_{ij} & \gamma_{ij} \\ -\gamma_{ij} & \beta_{ij} \end{bmatrix}$ for some $\beta_{ij}, \gamma_{ij} \in \mathbb{R}$.

Structured condition numbers for eigenvectors have been studied in [14, 17] and for invariant subspaces in [24, 26, 39], mostly for special cases. The (structured) perturbation analysis of quadratic matrix equations is a closely related area, which is comprehensively treated in [23, 40]. In this paper, we aim to provide a more general framework for studying structured condition numbers for invariant subspaces, which applies to all structures that form smooth manifolds.

The rest of this paper is organized as follows. In section 2, we briefly summarize known first-order perturbation results for invariant subspace along with associated notions, such as Sylvester operators and canonical angles. Two conceptually different approaches to the structured perturbation analysis of invariant subspaces for linear structures are described in section 3. One approach is based on a Kronecker product formulation and pattern matrices, much in the spirit of [9, 14, 22, 31, 41]. Although such an approach yields a computable formula for the structured condition number $c_{\mathbb{S}}(\mathcal{X})$, it gives little or no first hand information on the relationship between $c_{\mathbb{S}}(\mathcal{X})$ and $c(\mathcal{X})$. The other approach, possibly offering more insight into this relationship, is based on the observation that for several relevant structures, the Sylvester operator associated with an invariant subspace admits an orthogonal decomposition into two operators, one of them is confined to the structure. This property also allows one to develop global perturbation results and to deal with invariant subspaces that are stable under structured perturbations but unstable under unstructured perturbations. Both approaches extend to structures that form smooth manifolds, as shown in section 3.4. Illustrating the results, section 4 explains how structured condition numbers for product, Hamiltonian, and orthogonal eigenvalue problems can be derived in a considerably simple manner. The results extend to deflating subspaces of generalized eigenvalue problems; see section 5 and apply to structured matrix pencils including palindromic matrix pencils.

2. Preliminaries. Given a k-dimensional invariant subspace $\mathcal{X}$ of a matrix $A \in \mathbb{C}^{n\times n}$, we need some basis for $\mathcal{X}$ to begin with. Let the columns of the matrix $X \in \mathbb{C}^{n\times k}$ form such a basis. It is convenient to assume that this basis is orthonormal, which implies that $X^H X$ equals the $k \times k$ identity matrix I_k. If the columns of $X_\perp \in \mathbb{C}^{n\times k}$ form an orthonormal basis for $\mathcal{X}^\perp$, then the orthogonal complement of $\mathcal{X}$, then A has a [illegible] :

$$[X, X_\perp]^H A [X, X_\perp] = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \tag{2.1}$$

where $A_{11} \in \mathbb{C}^{k\times k}$ and $A_{22} \in \mathbb{C}^{(n-k)\times(n-k)}$.

An entity closely associated with $\mathcal{X}$ is the so-called [illegible]

$$\mathbf{T} : R \mapsto A_{22}R - RA_{11}. \tag{2.2}$$

This operator is invertible if and only if A_{11} and A_{22} have no eigenvalue in common, i.e., $\lambda(A_{11}) \cap \lambda(A_{22}) = \emptyset$; see [36, Thm. V.1.3]. The [illegible] of A_{11} and A_{22}, $\operatorname{sep}(A_{11}, A_{22})$, is defined as the smallest singular value of $\mathbf{T}$:

$$\operatorname{sep}(A_{11}, A_{22}) := \min_{R \neq 0} \frac{\|\mathbf{T}(R)\|_F}{\|R\|_F} = \min_{R \neq 0} \frac{\|A_{22}R - RA_{11}\|_F}{\|R\|_F}. \tag{2.3}$$

If $\mathbf{T}$ is invertible, this definition implies $\operatorname{sep}(A_{11}, A_{22}) = 1/\|\mathbf{T}^{-1}\|$, where $\|\cdot\|$ is the norm on the space of linear operators $\mathbb{R}^{k\times(n-k)} \to \mathbb{R}^{k\times(n-k)}$ induced by the Frobenius norm. Note that neither the invertibility of $\mathbf{T}$ nor the value of $\operatorname{sep}(A_{11}, A_{22})$ depend on the choice of orthonormal bases for $\mathcal{X}$ and $\mathcal{X}^\perp$. This justifies the following definition.

DEFINITION 2.1. [illegible] simple[illegible]

We are now prepared to state a first-order perturbation expansion for simple invariant subspaces, which can be proved by the implicit function theorem [37, 39, 25].

THEOREM 2.2. [illegible] A [illegible] (2.1) [illegible] $\mathcal{X}$ [illegible] X [illegible] $A+E \in \mathcal{B}(A)$ [illegible] A [illegible] $\mathcal{B}(A) \subset \mathbb{C}^{n\times n}$ [illegible] [illegible] A [illegible] $f : \mathcal{B}_A \to \mathbb{C}^{n\times k}$ [illegible] $X = f(A)$ [illegible] $\hat{X} = f(A+E)$ [illegible] $A+E$ [illegible] $X^H(\hat{X} - X) = 0$ [illegible]

$$\hat{X} = X - X_\perp \mathbf{T}^{-1}(X_\perp^H E X) + \mathcal{O}(\|E\|_F^2), \tag{2.4}$$

[illegible] $\mathbf{T} : R \mapsto A_{22}R - RA_{11}$

2.1. Canonical angles, a perturbation bound, and $c(\mathcal{X})$. In order to obtain perturbation bounds and condition numbers for invariant subspaces we require the notions of angles and distances between two subspaces.

DEFINITION 2.3. [illegible] X [illegible] Y [illegible] k [illegible] $\mathcal{X}$ [illegible] $\mathcal{Y}$ [illegible] $\sigma_1 \le \sigma_2 \le \cdots \le \sigma_k$ [illegible] $X^H Y$ [illegible] canonical angles [illegible] $\mathcal{X}$ [illegible] $\mathcal{Y}$ [illegible] $\theta_i(\mathcal{X}, \mathcal{Y}) := \arccos \sigma_i$ [illegible] $i = 1, \ldots, k$ [illegible] [illegible] $\Theta(\mathcal{X}, \mathcal{Y}) := \operatorname{diag}(\theta_1(\mathcal{X}, \mathcal{Y}), \ldots, \theta_k(\mathcal{X}, \mathcal{Y}))$

Canonical angles can be used to measure the distance between two subspaces. In particular, it can be shown that any unitarily invariant norm $\|\cdot\|_\gamma$ on $\mathbb{C}^{k\times k}$ defines a unitarily invariant metric d_γ on the space of k-dimensional subspaces via $d_\gamma(\mathcal{X}, \mathcal{Y}) = \|\sin[\Theta(\mathcal{X}, \mathcal{Y})]\|_\gamma$; see [36, p. 93].

In the case that one of the subspaces is spanned by a nonorthonormal basis, as in Theorem 2.2, the following lemma provides a useful tool for computing canonical angles.

LEMMA 2.4 (see [36]). [illegible] $\mathcal{X}$ [illegible] $[I_k, 0]^H$ [illegible] $\mathcal{Y}$ [illegible] $[I_k, R^H]^H$ [illegible] $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_k$ [illegible] [illegible] R [illegible] $\theta_i(\mathcal{X}, \mathcal{Y}) = \arctan \sigma_i$ [illegible] $i = 1, \ldots, k$

This yields the following perturbation bound for invariant subspaces.

COROLLARY 2.5. [illegible] 2.2

$$\|\Theta(\mathcal{X}, \hat{\mathcal{X}})\|_F \le \frac{\|E\|_F}{\operatorname{sep}(A_{11}, A_{22})} + \mathcal{O}(\|E\|_F^2), \tag{2.5}$$

[illegible] $\hat{\mathcal{X}} = \operatorname{range}(\hat{X})$

[illegible]. Without loss of generality, we may assume $X = [I, 0]^T$. Since $X^T(\hat{X} - X) = 0$ the matrix $\hat{X}$ must have the form $[I, R^H]^H$ for some $R \in \mathbb{C}^{(n-k)\times k}$. Together with the perturbation expansion (2.4) this implies

$$\|R\|_F = \|\hat{X} - X\|_F \leq \|E\|_F / \operatorname{sep}(A_{11}, A_{22}).$$

Inequality (2.5) is proved by applying Lemma 2.4 combined with the expansion $\arctan z = z + \mathcal{O}(z^3)$. ∎

The derived bound (2.5) is approximately tight. To see this, let V be a matrix such that $\|V\|_F = 1$ and $\|\mathbf{T}^{-1}(V)\|_F = 1/\operatorname{sep}(A_{11}, A_{22})$. Plugging $E = \epsilon X_\perp V X^H$ with $\epsilon > 0$ into the perturbation expansion (2.4) yields

$$\|\Theta(\mathcal{X}, \hat{\mathcal{X}})\|_F = \|\hat{X} - X\|_F + \mathcal{O}(\|\hat{X} - X\|_F^3) = \epsilon / \operatorname{sep}(A_{11}, A_{22}) + \mathcal{O}(\epsilon^2).$$

Hence, we obtain the following [illegible] $\mathcal{X}$:

$$c(\mathcal{X}) := \lim_{\epsilon \to 0} \sup \left\{ \|\Theta(\mathcal{X}, \hat{\mathcal{X}})\|_F / \epsilon : \ E \in \mathbb{C}^{n\times n}, \|E\|_F \leq \epsilon \right\} \tag{2.6}$$
$$= 1/\operatorname{sep}(A_{11}, A_{22}) = \|\mathbf{T}^{-1}\|;$$

see also [33, 36]. The condition number $c(\mathcal{X})$ extends to invariant subspaces $\mathcal{X}$ which are not simple by the convention $c(\mathcal{X}) = \infty$. Unlike eigenvalues, invariant subspaces with infinite condition numbers are generally discontinuous with respect to changes in the matrix entries, i.e., they are unstable under unstructured perturbations [36].

2.2. On the computation of sep. To obtain a computable formula for the quantity $\operatorname{sep}(A_{11}, A_{22})$, a convenient (but computationally expensive) approach is to express the Sylvester operator $\mathbf{T}$, see (2.2), in terms of Kronecker products:

$$\operatorname{vec}(\mathbf{T}(R)) = K_{\mathbf{T}} \cdot \operatorname{vec}(R), \tag{2.7}$$

where the $k(n-k) \times k(n-k)$ matrix $K_{\mathbf{T}}$ is given by

$$K_{\mathbf{T}} = I_k \otimes A_{22} - A_{11}^T \otimes I_{n-k}. \tag{2.8}$$

Here, "$\otimes$" denotes the Kronecker product of two matrices and the vec operator stacks the columns of a matrix in their natural order into one long vector [12]. Note that A_{11}^T denotes the complex transpose of A_{11}. Combining (2.3) with (2.7) yields the formula

$$\operatorname{sep}(A_{11}, A_{22}) = \sigma_{\min}(K_{\mathbf{T}}) = \sigma_{\min}(I_k \otimes A_{22} - A_{11}^T \otimes I_{n-k}), \tag{2.9}$$

where $\sigma_{\min}$ denotes the smallest singular value of a matrix.

Computing the separation based on a singular value decomposition of $K_{\mathbf{T}}$ is costly in terms of memory and computational time. A cheaper estimate of sep can be obtained by applying a norm estimator [15] to $K_{\mathbf{T}}^{-1}$. This amounts to the solution of a few linear equations $K_{\mathbf{T}} x = c$ and $K_{\mathbf{T}}^H x = d$ for particular chosen right-hand sides c and d or, equivalently, the solution of a few Sylvester equations $A_{22}X - XA_{11} = C$ and $A_{22}^H X - X A_{11}^H = D$. This approach becomes particularly attractive when A_{11} and A_{22} are already in Schur form; see [1, 5, 18, 19].

3. The structured condition number $c_{\mathbb{S}}(\mathcal{X})$. The condition number $c(\mathcal{X})$ for a simple invariant subspace $\mathcal{X}$ of A provides a first-order bound on the sensitivity of $\mathcal{X}$. This bound is strict in the sense that for any sufficiently small $\epsilon > 0$ there exists a perturbation E with $\|E\|_F = \epsilon$ such that $\|\Theta(\mathcal{X}, \hat{\mathcal{X}})\|_F \approx c(\mathcal{X})\epsilon$. If, however, it is

known that the set of admissible perturbations is restricted to a subset $\mathbb{S} \subseteq \mathbb{C}^{n\times n}$, then $c(\mathcal{X})$ may severely overestimate the actual worst-case sensitivity of $\mathcal{X}$. To avoid this effect, we introduce an appropriate notion of structured condition numbers in the sense of Rice [32] as follows.

DEFINITION 3.1. [illegible] $\mathbb{S} \subseteq \mathbb{C}^{n\times n}$ [illegible] $\mathcal{X}$ [illegible] $A \in \mathbb{S}$ [illegible] structured condition number for $\mathcal{X}$ [illegible] $\mathbb{S}$ [illegible]

$$c_{\mathbb{S}}(\mathcal{X}) := \lim_{\epsilon\to 0} \sup_{\substack{A+E\in\mathbb{S}\\ \|E\|_F\le\epsilon}} \inf\{\|\Theta(\mathcal{X},\hat{\mathcal{X}})\|_F/\epsilon : \hat{\mathcal{X}} \text{ is an invariant subspace of } A+E\}.$$

Note that the structured condition number $c_{\mathbb{S}}(\mathcal{X})$ may be finite even when $\mathcal{X}$ is not simple. This reflects the fact that (as in (1.1) with "$\alpha = 0$") an invariant subspace may be unstable with respect to unstructured perturbation ($c(\mathcal{X}) = \infty$) but stable with respect to structured perturbations ($c_{\mathbb{S}}(\mathcal{X}) < \infty$). If $\mathbb{S} = \mathbb{C}^{n\times n}$, then $c_{\mathbb{S}}(\mathcal{X}) = c(\mathcal{X})$.

If $\mathcal{X}$ is simple, then Definition 3.1 simplifies to

$$c_{\mathbb{S}}(\mathcal{X}) = \lim_{\epsilon\to 0} \sup\big\{\|\Theta(\mathcal{X},\hat{\mathcal{X}})\|_F/\epsilon :\ A+E\in\mathbb{S}, \|E\|_F\le\epsilon\big\}, \tag{3.1}$$

where $\hat{\mathcal{X}}$ is defined in the sense of Theorem 2.2.

As the supremum in (3.1) is taken over a set which is potentially smaller than for the unstructured condition number in (2.6), it is clear that $c_{\mathbb{S}}(\mathcal{X}) \le c(\mathcal{X})$. Much of the following discussion will be concerned with the question of how far can $c_{\mathbb{S}}(\mathcal{X})$ be below $c(\mathcal{X})$. As a first step, we provide a useful connection between the structured condition number and $\mathbf{T}^{-1}$.

LEMMA 3.2. [illegible] $\mathcal{X}$ [illegible] A [illegible] (2.1) [illegible] $\mathcal{X}$ [illegible] $\mathbb{S} \subseteq \mathbb{C}^{n\times n}$ [illegible]

$$c_{\mathbb{S}}(\mathcal{X}) = \lim_{\epsilon\to 0} \sup\big\{\|\mathbf{T}^{-1}(X_\perp^H E X)\|_F/\epsilon :\ A+E\in\mathbb{S}, \|E\|_F\le\epsilon\big\}, \tag{3.2}$$

[illegible] $\mathbf{T}$ [illegible] $\mathbf{T} : R \mapsto A_{22}R - RA_{11}$ [illegible]. This statement can be concluded from Theorem 2.2 along the line of arguments that led to the expression (2.6) for the standard condition number. ∎

3.1. A Kronecker product approach. In the following, we consider perturbations that are linearly structured, i.e., E is known to belong to some linear matrix subspace $\mathbb{L}$. In this case, Lemma 3.2 implies

$$c_{A+\mathbb{L}}(\mathcal{X}) = \sup\big\{\|\mathbf{T}^{-1}(X_\perp^H E X)\|_F :\ E\in\mathbb{L}, \|E\|_F = 1\big\}, \tag{3.3}$$

provided that $\mathcal{X}$ is simple.

The Kronecker product representation of $\mathbf{T}$ described in section 2.2 can be used to turn (3.3) into a computable formula for $c_{A+\mathbb{L}}(\mathcal{X})$. Very similar approaches have been used to obtain expressions for structured condition numbers in the context of eigenvalues [14, 22, 31, 41] and matrix functions [9]. Given an m-dimensional linear matrix subspace $\mathbb{L} \subseteq \mathbb{K}^{n\times n}$ with $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, one can always find an $n^2 \times m$ pattern matrix $M_{\mathbb{L}}$ such that for every $E \in \mathbb{L}$ there exists a uniquely defined parameter vector $p \in \mathbb{K}^m$ with

$$\operatorname{vec}(E) = M_{\mathbb{L}} p, \quad \|E\|_F = \|p\|_2.$$

This implies

$$\operatorname{vec}(\mathbf{T}^{-1}(X_\perp^H E X)) = K_{\mathbf{T}}^{-1}(X^T \otimes X_\perp^H)\operatorname{vec}(E) = K_{\mathbf{T}}^{-1}(X^T \otimes X_\perp^H)M_{\mathbb{L}}p, \tag{3.4}$$

where $K_{\mathbf{T}}$ is defined as in (2.8). Consequently, we have the formula

$$c_{A+\mathbb{L}}(\mathcal{X}) = \sup_{\|p\|_2=1} \|K_{\mathbf{T}}^{-1}(X^T \otimes X_\perp^H)M_{\mathbb{L}}p\|_2 = \|K_{\mathbf{T}}^{-1}(X^T \otimes X_\perp^H)M_{\mathbb{L}}\|_2, \tag{3.5}$$

provided that either $\mathbb{K} = \mathbb{C}$ or all of $\mathbb{K}$, A, and $\mathcal{X}$ are real.

If $\mathbb{K} = \mathbb{R}$ but A or $\mathcal{X}$ is complex, then problems occur because the supremum in (3.5) is taken with respect to real vectors p but $K_{\mathbf{T}}^{-1}(X^T \otimes X_\perp^H)M$ could be a complex matrix. Nevertheless, one has the following bounds to address such cases; see also [6].

LEMMA 3.3. [illegible] $\mathbb{L} \subseteq \mathbb{R}^{n\times n}$ [illegible] $M_{\mathbb{L}}$ [illegible] $\mathcal{X}$ [illegible] $A \in \mathbb{C}^{n\times n}$ [illegible]

$$\|K_{\mathbf{T}}^{-1}(X^T \otimes X_\perp^H)M_{\mathbb{L}}\|_2/\sqrt{2} \le c_{A+\mathbb{L}}(\mathcal{X}) \le \|K_{\mathbf{T}}^{-1}(X^T \otimes X_\perp^H)M_{\mathbb{L}}\|_2.$$

[illegible]. Let $B = K_{\mathbf{T}}^{-1}(X^T \otimes X_\perp^H)M_{\mathbb{L}}$ and decompose $B = B^{(R)} + \imath B^{(I)}$ with real matrices $B^{(R)}$ and $B^{(I)}$. Then

$$\frac{1}{\sqrt{2}}\left\|\begin{bmatrix} B_R & -B_I \\ B_I & B_R \end{bmatrix}\right\|_2 \le \left\|\begin{bmatrix} B_R \\ B_I \end{bmatrix}\right\|_2 \le \left\|\begin{bmatrix} B_R & -B_I \\ B_I & B_R \end{bmatrix}\right\|_2 = \|B\|_2.$$

Using $\left\|\begin{bmatrix} B_R \\ B_I \end{bmatrix}\right\|_2 = c_{A+\mathbb{L}}(\mathcal{X})$, this concludes the proof. ∎

3.2. An orthogonal decomposition approach. Although (3.5) provides an explicit expression for $c_{A+\mathbb{L}}(\mathcal{X})$, it tells little about the relationship to the unstructured condition number $c(\mathcal{X})$. In this section, we provide an alternative approach by decomposing the associated Sylvester operator $\mathbf{T} : R \mapsto A_{22}R - RA_{11}$ with respect to the structure.

For this purpose, assume the invariant subspace $\mathcal{X}$ to be simple, and let the columns of X and $X_\perp$ form orthonormal bases of $\mathcal{X}$ and $\mathcal{X}^\perp$, respectively. We set

$$\mathcal{N} := \{X_\perp^H E X : \ E \in \mathbb{L}\},$$

which can be considered as the structure induced by $\mathbb{L}$ in the $(2,1)$ block in a block Schur decomposition (2.1). Moreover, let $\mathcal{M}$ denote the preimage of $\mathcal{N}$ under $\mathbf{T}$. As we assume $\mathcal{X}$ to be simple, we can simply write $\mathcal{M} := \mathbf{T}^{-1}(\mathcal{N})$. Lemma 3.2 shows that the structured condition number of $\mathcal{X}$ is given by

$$c_{A+\mathbb{L}}(\mathcal{X}) = \|\mathbf{T}_s^{-1}\|,$$

where $\mathbf{T}_s$ is the restriction of $\mathbf{T}$ to $\mathcal{M} \to \mathcal{N}$, i.e., $\mathbf{T}_s := \mathbf{T}\big|_{\mathcal{M}\to\mathcal{N}}$. The operator $\mathbf{T}_s$ can be considered as the part of $\mathbf{T}$ that acts on the linear spaces induced by the structure.

In all examples considered in this paper, we additionally have the property that the operator $\mathbf{T}^\star : Q \mapsto A_{22}^H Q - Q A_{11}^H$ satisfies $\mathbf{T}^\star : \mathcal{N} \to \mathcal{M}$. Note that $\mathbf{T}^\star$ is the Sylvester operator dual to $\mathbf{T}$:

$$\langle \mathbf{T}(R), Q\rangle = \langle R, \mathbf{T}^\star(Q)\rangle$$

with the matrix inner product $\langle X, Y\rangle = \operatorname{trace}(Y^H X)$. This implies $\mathbf{T} : \mathcal{M}^\perp \to \mathcal{N}^\perp$, where $^\perp$ denotes the orthogonal complement w.r.t. the matrix inner product. Hence, $\mathbf{T}$ decomposes orthogonally into $\mathbf{T}_s$ and $\mathbf{T}_u := \mathbf{T}\big|_{\mathcal{M}^\perp \to \mathcal{N}^\perp}$, and we have

$$c(\mathcal{X}) = \max\{\|\mathbf{T}_s^{-1}\|, \|\mathbf{T}_u^{-1}\|\}. \tag{3.6}$$

Hence, comparing $c(\mathcal{X})$ with $c_{A+\mathbb{L}}(\mathcal{X})$ amounts to comparing $\|\mathbf{T}_u^{-1}\|$ with $\|\mathbf{T}_s^{-1}\|$.

REMARK 3.4. The conditions $\mathbf{T} : \mathcal{M} \to \mathcal{N}$ and $\mathbf{T}^\star : \mathcal{N} \to \mathcal{M}$ imply $\mathbf{T}^{-1}|_{\mathcal{N}\to\mathcal{M}} = \mathbf{T}_s^{-1}$ and $\mathbf{T}^{-\star}|_{\mathcal{M}\to\mathcal{N}} = \mathbf{T}_s^{-\star}$. Hence $\|\mathbf{T}_s^{-1}\| = \sqrt{\|(\mathbf{T}^{-\star} \circ \mathbf{T}^{-1})|_{\mathcal{N}\to\mathcal{N}}\|}$, and the power method can be applied to $\mathbf{T}^{-\star} \circ \mathbf{T}^{-1}$ in order to estimate $\|\mathbf{T}_s^{-1}\|$.

EXAMPLE 3.5. Consider the embedding of a complex matrix $B + \imath C$, with $B, C \in \mathbb{R}^{n\times n}$, into a real $2n \times 2n$ matrix of the form $A = \left[\begin{smallmatrix} B & C \\ -C & B \end{smallmatrix}\right]$. Let the columns of $Y + \imath Z$ and $Y_\perp + \imath Z_\perp$, where $Y, Z \in \mathbb{R}^{n\times k}$ and $Y_\perp, Z_\perp \in \mathbb{R}^{n\times(n-k)}$, form orthonormal bases for an invariant subspace of $B+\imath C$ and its orthogonal complement, respectively. Then the columns of $X = \left[\begin{smallmatrix} Y & Z \\ -Z & Y \end{smallmatrix}\right]$ and $X_\perp = \left[\begin{smallmatrix} Y_\perp & Z_\perp \\ -Z_\perp & Y_\perp \end{smallmatrix}\right]$ form orthonormal bases for an invariant subspace $\mathcal{X}$ of A and $\mathcal{X}^\perp$, respectively. This corresponds to the block Schur decomposition

$$[X, X_\perp]^T A [X, X_\perp] =: \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline 0 & A_{22} \end{array}\right] = \left[\begin{array}{cc|cc} B_{11} & C_{11} & B_{12} & C_{12} \\ -C_{11} & B_{11} & -C_{12} & B_{12} \\ \hline 0 & 0 & B_{22} & C_{22} \\ 0 & 0 & -C_{22} & B_{22} \end{array}\right],$$

and the associated Sylvester operator is given by $\mathbf{T} : R \mapsto A_{22}R - RA_{11}$.

If we consider perturbations having the same structure as A, then $\mathbb{L} = \left\{\left[\begin{smallmatrix} F & G \\ -G & F \end{smallmatrix}\right]\right\}$ and

$$\mathcal{N} := X_\perp^T \mathbb{L} X = \left\{\left[\begin{array}{cc} F_{21} & G_{21} \\ -G_{21} & F_{21} \end{array}\right]\right\}, \quad \mathcal{N}^\perp = \left\{\left[\begin{array}{cc} F_{21} & G_{21} \\ G_{21} & -F_{21} \end{array}\right]\right\}.$$

Moreover, we have $\mathbf{T} : \mathcal{N} \to \mathcal{N}$ and $\mathbf{T}^\star : \mathcal{N} \to \mathcal{N}$. The restricted operator $\mathbf{T}_s := \mathbf{T}\big|_{\mathcal{N}\to\mathcal{N}}$ becomes singular only if $B_{11} + \imath C_{11}$ and $B_{22} + \imath C_{22}$ have eigenvalues in common, while $\mathbf{T}_u := \mathbf{T}\big|_{\mathcal{N}^\perp \to \mathcal{N}^\perp}$ becomes singular if $B_{11} + \imath C_{11}$ and $B_{22} - \imath C_{22}$ have eigenvalues in common. Thus, there are situations in which the unstructured condition number $c(\mathcal{X}) = \max\{\|\mathbf{T}_s^{-1}\|, \|\mathbf{T}_u^{-1}\|\}$ can be significantly larger than the structured condition number $c_{\mathbb{S}}(\mathcal{X}) = c_{A+\mathbb{L}}(\mathcal{X}) = \|\mathbf{T}_s^{-1}\|$, e.g., if $\imath\gamma$ is nearly an eigenvalue of $B_{11} + \imath C_{11}$ while $-\imath\gamma$ is nearly an eigenvalue of $B_{22} + \imath C_{22}$ for some $\gamma \in \mathbb{R}$.

The introductionary example (1.1) is a special case of Example 3.5, where the unstructured condition number tends to infinity as the parameter α tends to zero. The results above imply that the structured condition number is given by

$$c_{\mathbb{S}}(\mathcal{X}) = \inf_{|\beta|^2+|\gamma|^2=1} \left\{ \left\| \begin{bmatrix} 0 & 1-\alpha \\ -1+\alpha & 0 \end{bmatrix} \begin{bmatrix} \beta & \gamma \\ -\gamma & \beta \end{bmatrix} - \begin{bmatrix} \beta & \gamma \\ -\gamma & \beta \end{bmatrix} \begin{bmatrix} 0 & -1-\alpha \\ 1+\alpha & 0 \end{bmatrix} \right\|_F \right\}^{-1} = \frac{1}{2}.$$

There is evidence to believe that $c_{\mathbb{S}}(\mathcal{X}) = 1/2$ holds even if $\alpha = 0$. However, all our arguments so far rest on the perturbation expansion in Theorem 2.2, which requires the invariant subspace to be simple; a condition that is not satisfied if $\alpha = 0$. This restriction will be removed in the following section by adapting the global perturbation analysis for invariant subspaces proposed by Stewart [35] and refined by Demmel [10]; see also [8].

3.3. Global perturbation bounds. In addition to the block Schur decomposition (2.1) we now consider the [illegible] block Schur decomposition

$$[X, X_\perp]^H (A+E)[X, X_\perp] = \begin{bmatrix} A_{11}+E_{11} & A_{12}+E_{12} \\ E_{21} & A_{22}+E_{22} \end{bmatrix} =: \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ E_{21} & \hat{A}_{22} \end{bmatrix}. \tag{3.7}$$

In order to obtain a formula for $\hat{X}$, a basis for the perturbed invariant subspace $\hat{\mathcal{X}}$ close to $\mathcal{X} = \operatorname{span}(X)$, we look for an invertible matrix of the form $W = \left[\begin{smallmatrix} I & 0 \\ -R & I \end{smallmatrix}\right]$ so that

$$W^{-1} \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ E_{21} & \hat{A}_{22} \end{bmatrix} W = \begin{bmatrix} \hat{A}_{11} - \hat{A}_{12}R & \hat{A}_{12} \\ E_{21} + R\hat{A}_{11} - \hat{A}_{22}R - R\hat{A}_{12}R & \hat{A}_{22} + R\hat{A}_{12} \end{bmatrix}$$

is in block upper triangular form. This implies that R is a solution of the algebraic Riccati equation

$$\hat{A}_{22}R - R\hat{A}_{11} + R\hat{A}_{12}R = E_{21}. \tag{3.8}$$

To solve this quadratic matrix equation and for deriving the structured condition number with respect to a linear matrix space $\mathbb{L}$ we need to require the following two conditions on $\mathbb{L}$.

A1: Let $\mathcal{N} = \{X_\perp^H F X : F \in \mathbb{L}\}$ and $\hat{\mathbf{T}} : R \mapsto \hat{A}_{22}R - R\hat{A}_{11}$. Then there exists a linear matrix space $\mathcal{M}$, having the same dimension as $\mathcal{N}$, such that $\hat{\mathbf{T}} : \mathcal{M} \to \mathcal{N}$ and $R\hat{A}_{12}R \in \mathcal{N}$ for all $R \in \mathcal{M}$.

A2: The restricted operator $\hat{\mathbf{T}}_s := \hat{\mathbf{T}}\big|_{\mathcal{M}\to\mathcal{N}}$ is invertible.

THEOREM 3.6. [illegible] 1 [illegible] 2 [illegible] $(4\|\hat{\mathbf{T}}_s^{-1}\|^2 \, \|\hat{A}_{12}\|_F \, \|E_{21}\|_F) < 1$ [illegible] $R \in \mathcal{M}$ [illegible] (3.8) [illegible]

$$\begin{aligned} \|R\|_F &\le \frac{2\|\hat{\mathbf{T}}_s^{-1}\| \, \|E_{21}\|_F}{1 + \sqrt{1 - 4\|\hat{\mathbf{T}}_s^{-1}\|^2 \, \|\hat{A}_{12}\|_F \, \|E_{21}\|_F}} \\ &< 2\|\hat{\mathbf{T}}_s^{-1}\| \, \|E_{21}\|_F. \end{aligned} \tag{3.9}$$

[illegible]. The result can be proved by constructing an iteration

$$R_0 \leftarrow 0, \quad R_{i+1} \leftarrow \hat{\mathbf{T}}_s^{-1}(E_{21} - R_i\hat{A}_{12}R_i),$$

which is well defined because $R_i \in \mathcal{M}$ implies $R_i\hat{A}_{12}R_i \in \mathcal{N}$. This approach is very similar to the technique used by Stewart; see [33, 35] or [36, Thm. V.2.11]. In fact, it can be shown in precisely the same way as in [36, Thm. V.2.11] that all iterates R_i satisfy a bound of the form (3.9) and converge to a solution of (3.8). □

Having obtained a solution R of (3.8), a basis for an invariant subspace $\hat{\mathcal{X}}$ of $A+E$ is given by $\hat{X} = X - X_\perp R$. Together with Lemma 2.4, this leads to the following global version of Corollary 2.5.

COROLLARY 3.7. [illegible] 3.6 [illegible] $\hat{\mathcal{X}}$ [illegible] $A+E$ [illegible]

$$\|\tan\Theta(\mathcal{X}, \hat{\mathcal{X}})\|_F \le \frac{2\|\hat{\mathbf{T}}_s^{-1}\| \, \|E_{21}\|_F}{1 + \sqrt{1 - 4\|\hat{\mathbf{T}}_s^{-1}\|^2 \, \|\hat{A}_{12}\|_F \, \|E_{21}\|_F}}. \tag{3.10}$$

The quantity $\|\hat{\mathbf{T}}_s^{-1}\|$ in the bound (3.10) can be related to $\|\mathbf{T}_s^{-1}\|$, the norm of the inverse of the unperturbed Sylvester operator, using the following lemma.

LEMMA 3.8. [illegible] $\mathbf{T} : R \mapsto A_{22}R - RA_{11}$ [illegible] (2.1) [illegible] $\mathbf{T} : \mathcal{M} \to \mathcal{N}$ [illegible] $\mathbf{T}_s := \mathbf{T}|_{\mathcal{M}\to\mathcal{N}}$ [illegible] $1/\|\mathbf{T}_s^{-1}\| > \|E_{11}\|_F + \|E_{22}\|_F$ [illegible] $\hat{\mathbf{T}}_s$ [illegible]

$$\|\hat{\mathbf{T}}_s^{-1}\| \leq \frac{\|\mathbf{T}_s^{-1}\|}{1 - \|\mathbf{T}_s^{-1}\|(\|E_{11}\|_F + \|E_{22}\|_F)}. \tag{3.11}$$

[illegible] Under the given assumptions we have

$$\|I - \mathbf{T}_s^{-1} \circ \hat{\mathbf{T}}_s\| = \sup_{\substack{R\in\mathcal{M} \\ \|R\|_F=1}} \|\mathbf{T}_s^{-1}(E_{22}R - RE_{11})\|_F \leq \|\mathbf{T}_s^{-1}\|(\|E_{11}\|_F + \|E_{22}\|_F) < 1.$$

Thus, the Neumann series

$$\sum_{i=0}^{\infty} (I - \mathbf{T}_s^{-1} \circ \hat{\mathbf{T}}_s)^i \circ \mathbf{T}_s^{-1}$$

converges to $\hat{\mathbf{T}}_s^{-1}$, which proves (3.11). □

Combining Corollary 3.7 with the expansion $\arctan z = z + \mathcal{O}(z^3)$ and Lemma 3.8 yields

$$\|\Theta(\mathcal{X}, \hat{\mathcal{X}})\|_F \leq \|\mathbf{T}_s^{-1}\| \, \|E\|_F + \mathcal{O}(\|E\|_F^2). \tag{3.12}$$

This implies that $c_{A+\mathbb{L}}(\mathcal{X})$, the structured condition number for $\mathcal{X}$, is bounded from above by $\|\mathbf{T}_s^{-1}\|$, even if the operator $\mathbf{T}$ itself is not invertible. To show that the structured condition number and $\|\mathbf{T}_s^{-1}\|$ are actually equal, we require the extra assumption that $\mathbf{T}^\star : \mathcal{N} \to \mathcal{M}$.

THEOREM 3.9. [illegible] $\mathcal{M}$ [illegible] $\hat{\mathbf{T}}$ [illegible] $E \in \mathbb{L}$ [illegible] $\mathbf{T} : R \mapsto A_{22}R - RA_{11}$ [illegible] $\mathbf{T}^\star : \mathcal{N} \to \mathcal{M}$ [illegible] $\mathbf{T}_s := \mathbf{T}|_{\mathcal{M}\to\mathcal{N}}$ [illegible] $c_{A+\mathbb{L}}(\mathcal{X}) = \|\mathbf{T}_s^{-1}\|$

[illegible]. By Lemma 3.8, it follows that $\hat{\mathbf{T}}_s$ is invertible for all sufficiently small perturbations E. Thus, the discussion provided above proves $c_{A+\mathbb{L}}(\mathcal{X}) \leq \|\mathbf{T}_s^{-1}\|$. It remains to construct perturbations $E \in \mathbb{L}$ so that

$$\lim_{\|E\|_F\to 0} \|\Theta(\mathcal{X}, \hat{\mathcal{X}})\|_F / \|E\|_F \geq \|\mathbf{T}_s^{-1}\|,$$

where $\hat{\mathcal{X}}$ denotes an invariant subspace of $A + E$ nearest to $\mathcal{X}$. For this purpose, we choose $E_{21} \in \mathcal{N}$ such that $\|E_{21}\|_F = 1$, $\|\mathbf{T}^{-1}(E_{21})\|_F = \|\mathbf{T}_s^{-1}\|$, and consider the perturbation $E = \epsilon X_\perp E_{21} X^H$. Because of (3.12) we may assume that the nearest invariant subspace $\hat{\mathcal{X}}$ of $A + E$ satisfies $\|\Theta(\mathcal{X}, \hat{\mathcal{X}})\|_2 < \pi/2$ for sufficiently small $\epsilon > 0$. In other words, none of the vectors in $\hat{\mathcal{X}}$ is orthogonal to $\mathcal{X}$. This implies the existence of a matrix R such that the columns of $\hat{X} = X - X_\perp R$ form a basis for $\hat{\mathcal{X}}$. Equivalently, R satisfies the matrix equation

$$\mathbf{T}(R) + RA_{12}R = \epsilon E_{21}.$$

If we decompose $R = R_s + R_u$, where $R_s \in \mathcal{M}$ and $R_u \in \mathcal{M}^\perp$, then $\mathbf{T}(R_s) \in \mathcal{N}$ while $\mathbf{T}^\star : \mathcal{N} \to \mathcal{M}$ implies $\mathbf{T}(R_u) \in \mathcal{N}^\perp$. Similarly, $RA_{12}R = Q_s + Q_u$ with $Q_s \in \mathcal{N}$ and

$Q_u \in \mathcal{N}^\perp$. Consequently, $\mathbf{T}(R_s) + Q_s = \epsilon E_{21}$ and since $\|Q_s\|_F = \mathcal{O}(\epsilon^2)$ it follows that

$$\lim_{\epsilon\to 0} \|R\|_F/\epsilon \geq \lim_{\epsilon\to 0} \|R_s\|_F/\epsilon = \|\mathbf{T}^{-1}(E_{21})\|_F = \|\mathbf{T}_s^{-1}\|.$$

Combining this inequality with $\|\Theta(\mathcal{X}, \hat{\mathcal{X}})\|_F = \|R\|_F + \mathcal{O}(\epsilon^2)$ yields the desired result. ∎

Let us briefly summarize the discussion on structured condition numbers. If $\mathcal{X}$ is simple, then $c_{A+\mathbb{L}}(\mathcal{X})$ is given by $\|\mathbf{T}_s^{-1}\|$. This equality also holds for the case that $\mathcal{X}$ is not simple but stable under structured perturbations, provided that the assumptions of Theorem 3.9 are satisfied. It is easy to see that all these extra assumptions are satisfied by the introductory example (1.1), showing that $c_{A+\mathbb{L}}(\mathcal{X}) = 1/2$ also holds for $\alpha = 0$.

3.4. Extension to nonlinear structures. So far we have mainly considered structures $\mathbb{S}$ that form (affine) linear matrix spaces. Nevertheless, the results from the previous subsections can be used to address a smooth manifold $\mathbb{S}$ by observing that the structured condition number with respect to $\mathbb{S}$ equals the one with respect to the tangent space of $\mathbb{S}$ at A. This is a consequence of the following theorem, which is much in the spirit of the corresponding result in [22, Thm. 2.1] for structured eigenvalue condition numbers.

THEOREM 3.10. $\mathbb{S}$ $\mathcal{X}$ $A \in \mathbb{S}$ (2.1) $\mathcal{X}$ $\mathbb{S}$

$$c_{\mathbb{S}}(\mathcal{X}) = \sup\left\{\|\mathbf{T}^{-1}(X_\perp^H E X)\|_F : \ E \in T_A\mathbb{S}, \|E\|_F = 1\right\}, \tag{3.13}$$

$\mathbf{T}$ $\mathbf{T} : R \mapsto A_{22}R - RA_{11}$ $T_A\mathbb{S}$ $\mathbb{S}$ A

. Let $E \in T_A\mathbb{S}$ with $\|E\|_F = 1$. Then there is a sufficiently smooth curve $G_E : (-\epsilon, \epsilon) \to \mathbb{K}^{n\times n}$ ($\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$) satisfying $G_E(0) = 0$, $G_E'(0) = E$ and $A + G_E(t) \in \mathbb{S}$ for all t. We have $G_E(t) = Et + \mathcal{O}(|t|^2)$ and, by Lemma 3.2,

$$\begin{aligned} c_{A+G_E(\cdot)}(\mathcal{X}) &= \lim_{\epsilon\to 0} \sup\left\{\|\mathbf{T}^{-1}(X_\perp^H G_E(t) X)\|_F/\epsilon : \ |t| \leq \epsilon\right\} \\ &= \lim_{\epsilon\to 0} \sup\left\{\|\mathbf{T}^{-1}(X_\perp^H E t X)\|_F/\epsilon : \ |t| \leq \epsilon\right\} \\ &= \|\mathbf{T}^{-1}(X_\perp^H E X)\|_F. \end{aligned}$$

The curves $A + G_E(\cdot)$ form a covering of an open neighborhood of $A \in \mathbb{S}$, implying

$$c_{\mathbb{S}}(\mathcal{X}) = \sup\left\{c_{A+G_E(\cdot)}(\mathcal{X}) : \ E \in T_A\mathbb{S}, \|E\|_F = 1\right\},$$

which proves (3.13). ∎

Theorem 3.10 admits the derivation of an explicit expression for $c_{\mathbb{S}}(\mathcal{X})$, e.g., by applying the Kronecker product approach from section 3.1 to $T_A\mathbb{S}$. This requires the computation of a pattern matrix for $T_A\mathbb{S}$; an issue which has been discussed for automorphism groups in [22].

4. Examples. In this section, we illustrate the applicability of the theory developed in the preceding section for product, Hamiltonian, and orthogonal eigenvalue problems.

4.1. Block cyclic matrices. Let us consider a matrix product

$$\Pi = A^{(p)} A^{(p-1)} \cdots A^{(1)},$$

where $A^{(1)}, \ldots, A^{(p)} \in \mathbb{C}^{n \times n}$. Computing invariant subspaces of matrix products has applications in several areas, such as model reduction, periodic discrete-time systems, and bifurcation analysis; see [42] for a recent survey. In many of these applications it is reasonable to consider factorwise perturbations, i.e., the perturbed product $\Pi = (A^{(p)} + E^{(p)})(A^{(p-1)} + E^{(p-1)}) \cdots (A^{(1)} + E^{(1)})$. What seems to be a multilinearly structured eigenvalue problem can be turned into a linearly structured eigenvalue problem associated with the block cyclic matrix

$$A = \begin{bmatrix} 0 & & & A^{(p)} \\ A^{(1)} & \ddots & & \\ & \ddots & \ddots & \\ & & A^{(p-1)} & 0 \end{bmatrix}.$$

To see this, let the columns of the block diagonal matrix $X = X^{(1)} \oplus X^{(2)} \oplus \cdots \oplus X^{(p)}$ with $X^{(1)}, \ldots, X^{(p)} \in \mathbb{C}^{n \times k}$ form a basis for an invariant subspace $\mathcal{X}$ of A. By direct computation, it can be seen that the columns of $X^{(1)}$ form a basis for an invariant subspace of Π. Vice versa, the periodic Schur decomposition [4, 13] shows that any basis $X^{(1)}$ for an invariant subspace of Π can be extended to a basis $X^{(1)} \oplus X^{(2)} \oplus \cdots \oplus X^{(p)}$ for an invariant subspace $\mathcal{X}$ of A.

To perform a structured perturbation analysis for an invariant subspace $\mathcal{X}$ admitting an orthonormal basis $X = X^{(1)} \oplus X^{(2)} \oplus \cdots \oplus X^{(p)}$, we first note that there is an orthonormal basis $X_\perp$ of $\mathcal{X}^\perp$ having the form $X_\perp = X_\perp^{(1)} \oplus X_\perp^{(2)} \oplus \cdots \oplus X_\perp^{(p)}$. This leads to the block Schur decomposition

$$[X, X_\perp]^T A [X, X_\perp] = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix},$$

where $A_{11} \in \operatorname{cyc}(k, k, p)$, $A_{12} \in \operatorname{cyc}(k, n-k, p)$, $A_{22} \in \operatorname{cyc}(n-k, n-k, p)$, and $\operatorname{cyc}(n_1, n_2, p)$ denotes the set of $p \times p$ block cyclic matrices with $n_1 \times n_2$ blocks. The corresponding Sylvester operator is given by $\mathbf{T} : R \mapsto A_{22} R - R A_{11}$.

Factorwise perturbations in Π correspond to block cyclic perturbations in A, i.e., $\mathbb{S} = \operatorname{cyc}(n, n, p)$. The set $\mathcal{N} = X_\perp^T \mathbb{S} X$ coincides with $\operatorname{cyc}(n-k, k, p)$ and we have $\mathbf{T} : \mathcal{M} \to \mathcal{N}$, where $\mathcal{M}$ equals $\operatorname{diag}(n-k, k, p)$, the set of $p \times p$ block diagonal matrices with $(n-k) \times k$ blocks. Moreover, it can be directly verified that $\mathbf{T}^\star : \mathcal{N} \to \mathcal{M}$. Letting $\mathbf{T}_s = \mathbf{T}\big|_{\mathcal{M} \to \mathcal{N}}$ and $\mathbf{T}_u = \mathbf{T}\big|_{\mathcal{M}^\perp \to \mathcal{N}^\perp}$, we thus have $c_{\mathbb{S}}(\mathcal{X}) = \|\mathbf{T}_s^{-1}\|$ and $c(\mathcal{X}) = \max\{\|\mathbf{T}_s^{-1}\|, \|\mathbf{T}_u^{-1}\|\}$. Note that $\mathcal{M}^\perp$, $\mathcal{N}^\perp$ coincide with the set of all $p \times p$ block matrices with $(n-k) \times k$ blocks that are zero in their block diagonal or block cyclic part, respectively.

Although $\mathbf{T}_s$ is invertible if and only if $\mathbf{T}$ is invertible [25], the following example reveals that there may be significant difference between $\|\mathbf{T}_s^{-1}\|$ and $\|\mathbf{T}^{-1}\|$ (and consequently between the structured and unstructured condition numbers for $\mathcal{X}$).

EXAMPLE 4.1 (see [25]). Let $p = 2$, $A_{11} = \left[\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right]$, and $A_{22} = \left[\begin{smallmatrix} 0 & C \\ D & 0 \end{smallmatrix}\right]$, where

$$C = \begin{bmatrix} 10^5 & 10^5 \\ 0 & 10^{-5} \end{bmatrix}, \quad D = \begin{bmatrix} 10^{-5} & 0 \\ 0 & 10^5 \end{bmatrix}.$$

Then the structured condition number is given by

$$c_{\mathbb{S}}(\mathcal{X}) = \left\| \begin{bmatrix} C & -I_2 \\ 0 & D \end{bmatrix}^{-1} \right\|_2 = \sqrt{2} \times 10^5,$$

while the unstructured condition number is much higher,

$$c(\mathcal{X}) = \max\left\{ c_{\mathbb{S}}(\mathcal{X}), \left\| \begin{bmatrix} D & -I_2 \\ 0 & C \end{bmatrix}^{-1} \right\|_2 \right\} = 10^{10}.$$

Other and more detailed approaches to the perturbation analysis for invariant subspaces of (generalized) matrix products, yielding similar results, can be found in [3, 27].

4.2. Hamiltonian matrices. A Hamiltonian matrix is a $2n \times 2n$ matrix A of the form

$$A = \begin{bmatrix} -B & G \\ Q & B^T \end{bmatrix}, \quad G = G^T, \quad Q = Q^T,$$

where $B, G, Q \in \mathbb{R}^{n\times n}$. Hamiltonian matrices arise from, e.g., linear-quadratic optimal control problems and certain quadratic eigenvalue problems; see [2, 29] and the references therein. A particular property of A is that its eigenvalues are symmetric with respect to the imaginary axis. Hence, if A has no purely imaginary eigenvalues, there are n eigenvalues having negative real part. The invariant subspace $\mathcal{X}$ belonging to these n eigenvalues is called the stable invariant subspace. For all $x \in \mathcal{X}$ we have $Jx \perp \mathcal{X}$ with $J = \left[\begin{smallmatrix} 0 & I_n \\ -I_n & 0 \end{smallmatrix}\right]$, a property which makes $\mathcal{X}$ an isotropic vector space [30]. If the columns of $X \in \mathbb{R}^{2n\times n}$ form an orthonormal basis for $\mathcal{X}$, the isotropy of $\mathcal{X}$ implies that $[X, JX]$ is an orthogonal matrix and we have the structured block Schur decomposition

$$[X, X_\perp]^T A [X, X_\perp] = \begin{bmatrix} -\tilde{B} & \tilde{G} \\ 0 & \tilde{B}^T \end{bmatrix}, \quad \tilde{G} = \tilde{G}^T.$$

The corresponding Sylvester operator is given by $\mathbf{T} : R \mapsto \tilde{B}^T R + R\tilde{B}$.

If we restrict the set $\mathbb{S}$ of admissible perturbations to be Hamiltonian, then $\mathcal{N} = X_\perp^T \mathbb{S} X$ equals $\mathrm{symm}(n)$, the set of $n \times n$ symmetric matrices, while $\mathcal{N}^\perp = \mathrm{skew}(n)$, the set of $n\times n$ skew-symmetric matrices. It can be directly seen that $\mathbf{T} : \mathcal{N} \to \mathcal{N}$ and, moreover, $\mathbf{T}^\star = \mathbf{T}$. Thus, by letting $\mathbf{T}_s = \mathbf{T}\big|_{\mathcal{N}\to\mathcal{N}}$ and $\mathbf{T}_u = \mathbf{T}\big|_{\mathcal{N}^\perp\to\mathcal{N}^\perp}$, we have $c_{\mathbb{S}}(\mathcal{X}) = \|\mathbf{T}_s^{-1}\|$ and $c(\mathcal{X}) = \max\{\|\mathbf{T}_s^{-1}\|, \|\mathbf{T}_u^{-1}\|\}$. It is known that the expression $\|\tilde{B}^T R + R\tilde{B}\|_F / \|R\|_F$, $R \neq 0$, is always minimized by a symmetric matrix R [7, Thm. 8], which implies $\|\mathbf{T}_u^{-1}\| \leq \|\mathbf{T}_s^{-1}\|$. Hence, the structured and unstructured condition numbers for the stable invariant subspace of a Hamiltonian matrix are always the same.

A more general perturbation analysis for (block) Hamiltonian Schur forms, based on the technique of splitting operators and Lyapunov majorants, can be found in [24].

4.3. Orthogonal matrices. As an orthogonal matrix $A \in \mathbb{R}^{n\times n}$ is normal, the block Schur decomposition associated with a simple invariant subspace $\mathcal{X}$ is block diagonal:

$$[X, X_\perp]^T A [X, X_\perp] = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}.$$

Here, we will assume for convenience that X and $X_\perp$ are real. Both diagonal blocks, $A_{11} \in \mathbb{R}^{k\times k}$ and $A_{22} \in \mathbb{R}^{(n-k)\times(n-k)}$, are again orthogonal matrices.

The set of orthogonal matrices $\mathbb{S} = \{A : A^T A = I\}$ forms a smooth real manifold and the tangent space of $\mathbb{S}$ at A is given by $T_A\mathbb{S} = \{AW : W \in \mathrm{skew}(n)\}$. According to Theorem 3.10, this implies that the structured condition number is given by

$$\begin{aligned} c_{\mathbb{S}}(\mathcal{X}) &= \sup\left\{\|\mathbf{T}^{-1}(X_\perp^T AWX)\|_F : W \in \mathrm{skew}(n), \|AW\|_F = 1\right\} \\ &= \sup\left\{\|\mathbf{T}^{-1}(A_{22}X_\perp^T WX)\|_F : W \in \mathrm{skew}(n), \|W\|_F = 1\right\} \\ &= \sup\left\{\|\mathbf{T}^{-1}(A_{22}W_{21})\|_F : W_{21} \in \mathbb{R}^{(n-k)\times k}, \|W_{21}\|_F = 1\right\} \\ &= \sup\left\{\|\mathbf{T}^{-1}(\tilde{W}_{21})\|_F : \tilde{W}_{21} \in \mathbb{R}^{(n-k)\times k}, \|\tilde{W}_{21}\|_F = 1\right\} = c(\mathcal{X}), \end{aligned}$$

where $\mathbf{T} : R \mapsto A_{22}R - RA_{11}$. Here we used the fact that the "off-diagonal" block $W_{21} = X_\perp^T WX$ of a skew-symmetric matrix W has no particular structure. Hence, there is no difference between structured and unstructured condition numbers for invariant subspaces of orthogonal matrices.

5. Extension to matrix pencils. In this section, we extend the results of section 3 to deflating subspaces of matrix pencils. The exposition is briefer than for the standard eigenvalue problem as many of the results can be derived by similar techniques.

Throughout this section it is assumed that our matrix pencil $A - \lambda B$ of interest, with $n \times n$ matrices A and B, is regular, i.e., $\det(A - \lambda B) \not\equiv 0$. The roots $\lambda \in \mathbb{C}$ (if any) of $\det(A - \lambda B) = 0$ are the finite eigenvalues of the pencil. In addition, if B is not invertible, then the pencil has infinite eigenvalues. A k-dimensional subspace $\mathcal{X}$ is called a [illegible] of $A - \lambda B$ if $A\mathcal{X}$ and $B\mathcal{X}$ are both contained in a subspace $\mathcal{Y}$ of dimension k. The regularity of $A - \lambda B$ implies that such a subspace $\mathcal{Y}$ is uniquely defined; we call $\mathcal{Y}$ a [illegible] and $(\mathcal{X}, \mathcal{Y})$ a [illegible]; see [36] for a more detailed introduction.

Let $(\mathcal{X}, \mathcal{Y})$ be such a pair of deflating subspaces and let the columns of $X, X_\perp, Y, Y_\perp$ form orthonormal bases for $\mathcal{X}, \mathcal{X}^\perp, \mathcal{Y}, \mathcal{Y}^\perp$, respectively. Then $A - \lambda B$ admits the following [illegible]:

$$[Y, Y_\perp]^H (A - \lambda B)\,[X, X_\perp] = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} - \lambda \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix}. \tag{5.1}$$

The eigenvalues of $A-\lambda B$ are the union of the eigenvalues of the $k\times k$ pencil $A_{11}-\lambda B_{11}$ and the $(n-k)\times(n-k)$ pencil $A_{22} - \lambda B_{22}$.

An entity closely associated with (5.1) is the [illegible]

$$\mathbf{T} : (R_r, R_l) \mapsto (A_{22}R_r - R_l A_{11}, B_{22}R_r - R_l B_{11}), \tag{5.2}$$

where R_r and R_l are $(n-k)\times k$ matrices. It can be shown [34] that $\mathbf{T}$ is invertible if and only if the matrix pencils $A_{11} - \lambda B_{11}$ and $A_{22} - \lambda B_{22}$ have no eigenvalues in common. Clearly, this property is independent of the choice of orthonormal bases for $\mathcal{X}$ and $\mathcal{Y}$, justifying the following definition.

DEFINITION 5.1. [illegible] simple [illegible] $A_{11} - \lambda B_{11}$ [illegible] $A_{22} - \lambda B_{22}$ [illegible] (5.1) [illegible]

Provided that $\mathbf{T}$ is invertible, the [illegible] $A_{11} - \lambda B_{11}$ and $A_{22} - \lambda B_{22}$ can be defined via the norm of the inverse of $\mathbf{T}$:

$$\begin{aligned} \mathrm{dif}[(A_{11}, B_{11}), (A_{22}, B_{22})] &:= 1/\sup\left\{\|\mathbf{T}^{-1}(E_{21}, F_{21})\|_F : \|(E_{21}, F_{21})\|_F = 1\right\} \\ &= 1/\|\mathbf{T}^{-1}\|, \end{aligned}$$

where we let $\|(E_{21}, F_{21})\|_F = \sqrt{\|E_{21}\|_F^2 + \|F_{21}\|_F^2}$. Not surprisingly, it turns out that $\mathbf{T}^{-1}$ governs the sensitivity of $(\mathcal{X}, \mathcal{Y})$ with respect to perturbations in A and B.

THEOREM 5.2 (see [38, 25]). $A - \lambda B$ (5.1)
$(\mathcal{X}, \mathcal{Y}) = (\mathrm{span}(X), \mathrm{span}(Y))$ $(A + E, B + F) \in \mathcal{B}(A, B)$
(A, B) $\mathcal{B}(A, B) \subset \mathbb{C}^{n\times n} \times \mathbb{C}^{n\times n}$
(A, B) $f : \mathcal{B}(A, B) \to \mathbb{C}^{n\times k} \times \mathbb{C}^{n\times k}$ $(X, Y) = f(A, B)$, $(\hat{X}, \hat{Y}) = f(A + E, B + F)$
$(A + E) - \lambda(B + F)$
$X^H(\hat{X} - X) = Y^H(\hat{Y} - Y) = 0$

$$(\hat{X}, \hat{Y}) = (X, Y) - (X_\perp R_r, Y_\perp R_l) + \mathcal{O}(\|[E, F]\|^2), \tag{5.3}$$

$(R_r, R_l) = \mathbf{T}^{-1}(Y_\perp^H EX, Y_\perp^H FX)$ $\mathbf{T}$ (5.2)

By using similar techniques as in section 3, it can be concluded from (5.3) that the condition number for $(\mathcal{X}, \mathcal{Y})$, defined as

$$c(\mathcal{X}, \mathcal{Y}) := \lim_{\epsilon\to 0} \sup \big\{ \|(\Theta(\mathcal{X}, \hat{\mathcal{X}}), \Theta(\mathcal{Y}, \hat{\mathcal{Y}}))\|_F/\epsilon : \ E, F \in \mathbb{C}^{n\times n}, \|(E, F)\|_F \le \epsilon \big\},$$

happens to coincide with $\|\mathbf{T}^{-1}\| = 1/\operatorname{dif}[(A_{11}, B_{11}), (A_{22}, B_{22})]$; a result which goes back to Stewart [34, 35]. If $\operatorname{dif}[(A_{11}, B_{11}), (A_{22}, B_{22})] = 0$, then T is not invertible and, by convention, $c(\mathcal{X}, \mathcal{Y}) = \infty$. Algorithms that estimate dif efficiently by solving only a few generalized Sylvester equations can be found in [20, 21].

It may happen that $\mathcal{X}$ and $\mathcal{Y}$ are not equally sensitive to perturbations. In this case, $c(\mathcal{X}, \mathcal{Y})$ overestimates the sensitivity of one of the deflating subspaces; an aspect emphasized by Sun [38, 39], who has also pointed out that separating the influence of the operator $\mathbf{T}^{-1}$ on $\mathcal{X}$ and $\mathcal{Y}$ resolves this difficulty. However, for the purpose of simplifying the presentation we will only consider joint (structured) condition numbers for $(\mathcal{X}, \mathcal{Y})$.

DEFINITION 5.3. $\mathbb{S} \subseteq \mathbb{C}^{n\times n} \times \mathbb{C}^{n\times n}$ $(\mathcal{X}, \mathcal{Y})$ $A - \lambda B$ $(A, B) \in \mathbb{S}$ structured condition number for $(\mathcal{X}, \mathcal{Y})$ $\mathbb{S}$

$$c_{\mathbb{S}}(\mathcal{X}, \mathcal{Y}) := \lim_{\epsilon\to 0} \sup_{\substack{(A+E,B+F)\in\mathbb{S} \\ \|(E,F)\|_F\le\epsilon}} \inf \left\{ \|(\Theta(\mathcal{X}, \hat{\mathcal{X}}), \Theta(\mathcal{Y}, \hat{\mathcal{Y}}))\|_F/\epsilon : \begin{array}{c} (\hat{\mathcal{X}}, \hat{\mathcal{Y}}) \\ (A + E) - \lambda(B + F) \end{array} \right\}.$$

If $\mathbb{S} = \mathbb{C}^{n\times n} \times \mathbb{C}^{n\times n}$, then $c_{\mathbb{S}}(\mathcal{X}, \mathcal{Y}) = c(\mathcal{X}, \mathcal{Y})$.

A straightforward generalization of Lemma 3.2 relates $c_{\mathbb{S}}(\mathcal{X}, \mathcal{Y})$ to the norm of $\mathbf{T}^{-1}$ restricted to a certain subset.

LEMMA 5.4. $(\mathcal{X}, \mathcal{Y})$ $A - \lambda B$ (2.1)
$\mathcal{X}$ $\mathbb{S} \subseteq \mathbb{C}^{n\times n}$

$$c_{\mathbb{S}}(\mathcal{X}) = \lim_{\epsilon\to 0} \sup \big\{ \|\mathbf{T}^{-1}(Y_\perp^H EX, Y_\perp^H FX)\|_F/\epsilon : \ (A + E, B + F) \in \mathbb{S}, \|(E, F)\|_F \le \epsilon \big\},$$

$\mathbf{T}$ (5.2)

5.1. A Kronecker product approach. Using Kronecker products, the generalized Sylvester operator $\mathbf{T}$ can be represented as

$$\operatorname{vec}(\mathbf{T}(R_r, R_l)) = K_{\mathbf{T}} \begin{bmatrix} \operatorname{vec}(R_r) \\ \operatorname{vec}(R_l) \end{bmatrix},$$

with the $2k(n-k) \times 2k(n-k)$ matrix

$$K_{\mathbf{T}} = \begin{bmatrix} I_k \otimes A_{22} & -A_{11}^T \otimes I_{n-k} \\ I_k \otimes B_{22} & -B_{11}^T \otimes I_{n-k} \end{bmatrix}.$$

This implies $c(\mathcal{X}, \mathcal{Y}) = \|\mathbf{T}^{-1}\| = \|K_{\mathbf{T}}^{-1}\|_2$.

In the following, we will assume that the structure $\mathbb{S}$ under consideration takes the form $\mathbb{S} = (A, B) + \mathbb{L}$. Here, $\mathbb{L}$ denotes a linear matrix pencil subspace, i.e., $(E_1, F_1) \in \mathbb{L}$ and $(E_2, F_2) \in \mathbb{L}$ imply $(\alpha E_1 + \beta E_2, \alpha F_1 + \beta F_2) \in \mathbb{L}$ for all $\alpha, \beta \in \mathbb{K}$, where $\mathbb{K} = \mathbb{R}$ if $\mathbb{L}$ is real or $\mathbb{K} = \mathbb{C}$ if $\mathbb{L}$ is complex. Let m be the dimension of $\mathbb{L}$. Then one can always find a $2n^2 \times m$ pattern matrix $M_{\mathbb{L}}$ such that for every $(E, F) \in \mathbb{L}$ there exists a uniquely defined parameter vector $p \in \mathbb{K}^m$ with

$$\begin{bmatrix} \operatorname{vec}(E) \\ \operatorname{vec}(F) \end{bmatrix} = M_{\mathbb{L}} p, \quad \|(E, F)\|_F = \|p\|_2.$$

This yields for $(R_r, R_l) = \mathbf{T}^{-1}(Y_\perp^H E X, Y_\perp^H F X)$ with $(E, F) \in \mathbb{L}$,

$$\begin{bmatrix} \operatorname{vec}(R_r) \\ \operatorname{vec}(R_l) \end{bmatrix} = K_{\mathbf{T}}^{-1} \begin{bmatrix} X \otimes Y_\perp^H & 0 \\ 0 & X \otimes Y_\perp^H \end{bmatrix} M_{\mathbb{L}} p.$$

Hence, Lemma 5.4 implies

$$\begin{aligned} c_{(A,B)+\mathbb{L}}(\mathcal{X}, \mathcal{Y}) &= \sup_{\substack{p \in \mathbb{K}^m \\ \|p\|_2 = 1}} \left\| K_{\mathbf{T}}^{-1} \begin{bmatrix} X \otimes Y_\perp^H & 0 \\ 0 & X \otimes Y_\perp^H \end{bmatrix} M_{\mathbb{L}} p \right\|_2 \\ &= \left\| K_{\mathbf{T}}^{-1} \begin{bmatrix} X \otimes Y_\perp^H & 0 \\ 0 & X \otimes Y_\perp^H \end{bmatrix} M_{\mathbb{L}} \right\|_2. \end{aligned} \tag{5.4}$$

Note that the latter equality only holds provided that either $\mathbb{K} = \mathbb{C}$, or all of $\mathbb{K}$, A, B, $\mathcal{X}$, and $\mathcal{Y}$ are real. Otherwise, inequalities analogous to Lemma 3.3 can be derived.

5.2. An orthogonal decomposition approach. In this section, we extend the orthogonal decomposition approach of section 3.2 to matrix pencils in order to gain more insight into the relationship between the structured and unstructured condition numbers for a pair of deflating subspaces.

For this purpose, assume the pair of deflating subspaces $(\mathcal{X}, \mathcal{Y})$ to be simple, and let the columns of $X, X_\perp, Y, Y_\perp$ form orthonormal bases for $\mathcal{X}, \mathcal{X}^\perp, \mathcal{Y}, \mathcal{Y}^\perp$, respectively. Let

$$\mathcal{N} := \{(Y_\perp^H E X, Y_\perp^H F X) : \ (E, F) \in \mathbb{L}\},$$

and let $\mathcal{M}$ denote the preimage of $\mathcal{N}$ under $\mathbf{T}$, i.e., $\mathcal{M} := \mathbf{T}^{-1}(\mathcal{N})$. Then Lemma 5.4 implies that the structured condition number for $(\mathcal{X}, \mathcal{Y})$ is given by

$$c_{(A,B)+\mathbb{L}}(\mathcal{X}, \mathcal{Y}) = \|\mathbf{T}_s^{-1}\|,$$

where $\mathbf{T}_s$ is the restriction of $\mathbf{T}$ to $\mathcal{M} \to \mathcal{N}$, i.e., $\mathbf{T}_s := \mathbf{T}\big|_{\mathcal{M} \to \mathcal{N}}$.

Let us assume that we additionally have the property that the linear matrix operator

$$\mathbf{T}^\star : (Q_r, Q_l) \mapsto (A_{22}^H Q_r + B_{22}^H Q_l, -Q_r A_{11}^H - Q_l B_{11}^H)$$

satisfies $\mathbf{T}^\star : \mathcal{N} \to \mathcal{M}$. This is equivalent to the condition $\mathbf{T} : \mathcal{M}^\perp \to \mathcal{N}^\perp$, where $^\perp$ denotes the orthogonal complement w.r.t. the inner product $\langle (X_r, X_l), (Y_r, Y_l) \rangle = \mathrm{trace}(Y_r^H X_r + Y_l^H X_l)$. Note that $\mathbf{T}^\star$ can be considered as the linear operator dual to $\mathbf{T}$:

$$\langle \mathbf{T}(R_r, R_l), (Q_r, Q_l) \rangle = \langle (R_r, R_l), \mathbf{T}^\star(Q_r, Q_l) \rangle.$$

The same conclusions as for the matrix case in section 3.2 can be drawn: $\mathbf{T}$ decomposes orthogonally into $\mathbf{T}_s$ and $\mathbf{T}_u := \mathbf{T}\big|_{\mathcal{M}^\perp \to \mathcal{N}^\perp}$, and we have

$$c(\mathcal{X}, \mathcal{Y}) = \max\{\|\mathbf{T}_s^{-1}\|, \|\mathbf{T}_u^{-1}\|\}.$$

5.3. Global perturbation bounds. To derive global perturbation bounds we consider, in addition to (5.1), the generalized block Schur decomposition

$$[Y, Y_\perp]^H \big((A+E) - \lambda(B+F)\big)[X, X_\perp] = \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ E_{21} & \hat{A}_{22} \end{bmatrix} - \lambda \begin{bmatrix} \hat{B}_{11} & \hat{B}_{12} \\ F_{21} & \hat{B}_{22} \end{bmatrix}. \tag{5.5}$$

The following approach follows the work by Stewart [34, 35], which has been refined by Demmel and Kågström in [11]. In order to obtain bases $(\hat{X}, \hat{Y})$ for a nearby pair of perturbed deflating subspaces $(\hat{\mathcal{X}}, \hat{\mathcal{Y}})$ we look for $(n-k) \times k$ matrices R_r and R_l such that the matrix pencil

$$\begin{bmatrix} I_k & 0 \\ R_l & I_{n-k} \end{bmatrix} \left(\begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ E_{21} & \hat{A}_{22} \end{bmatrix} - \lambda \begin{bmatrix} \hat{B}_{11} & \hat{B}_{12} \\ F_{21} & \hat{B}_{22} \end{bmatrix} \right) \begin{bmatrix} I_k & 0 \\ -R_r & I_{n-k} \end{bmatrix}$$

is in block upper triangular form. This is equivalent to the condition that the pair (R_r, R_l) satisfies the following system of quadratic matrix equations:

$$\begin{aligned} \hat{A}_{22} R_r - R_l \hat{A}_{11} + R_l \hat{A}_{12} R_r &= E_{21}, \\ \hat{B}_{22} R_r - R_l \hat{B}_{11} + R_l \hat{B}_{12} R_r &= F_{21}. \end{aligned} \tag{5.6}$$

The following assumptions on the linear structure $\mathbb{L} \subseteq \mathbb{C}^{n \times n} \times \mathbb{C}^{n \times n}$ are related to the solvability of (5.6), along the lines of assumptions A1 and A2 for the matrix case:

A3: Let $\mathcal{N} = \{(Y_\perp^H G X, Y_\perp^H H X) : (G, H) \in \mathbb{L}\}$ and

$$\hat{\mathbf{T}} : (R_r, R_l) \mapsto (\hat{A}_{22} R_r - R_l \hat{A}_{11}, \hat{B}_{22} R_r - R_l \hat{B}_{11}).$$

Then there exists a linear matrix space $\mathcal{M}$, having the same dimension as $\mathcal{N}$, such that $\hat{\mathbf{T}} : \mathcal{M} \to \mathcal{N}$ and $(R_l \hat{A}_{12} R_r, R_l \hat{B}_{12} R_r) \in \mathcal{N}$ for all $(R_r, R_l) \in \mathcal{M}$.

A4: The restricted operator $\hat{\mathbf{T}}_s := \hat{\mathbf{T}}\big|_{\mathcal{M} \to \mathcal{N}}$ is invertible.

THEOREM 5.5. 3 4

$$\kappa := 4 \|\hat{\mathbf{T}}_s^{-1}\|^2 \, \|(\hat{A}_{12}, \hat{B}_{12})\|_F \, \|(E_{21}, F_{21})\|_F < 1,$$

[illegible] $(R_r, R_l) \in \mathcal{M}$ [illegible] (5.6) [illegible]

$$\|(R_r, R_l)\|_F \leq \frac{2\|\hat{\mathbf{T}}_s^{-1}\|\,\|(E_{21}, F_{21})\|_F}{1+\sqrt{1-\kappa}} < 2\|\hat{\mathbf{T}}_s^{-1}\|\,\|(E_{21}, F_{21})\|_F.$$

[illegible]. It follows from A3 that the iteration

$$(R_0, L_0) \leftarrow (0,0), \quad (R_{i+1}, L_{i+1}) \leftarrow \hat{\mathbf{T}}_s^{-1}(E_{21} - L_i\hat{A}_{12}R_i, F_{21} - L_i\hat{B}_{12}R_i)$$

is well defined and $(R_i, L_i) \in \mathcal{M}$ for all i. Its convergence and the bound (5.5) can be proved along the lines of the proof of [36, Thm. V.2.11]. ∎

Any solution (R_r, R_l) of (5.6) yields a pair of deflating subspaces $(\hat{\mathcal{X}}, \hat{\mathcal{Y}})$ of the perturbed pencil $(A+E)-\lambda(B+F)$ with the bases $\hat{X} = X - X_\perp R_r$ and $\hat{Y} = Y - Y_\perp R_l$. Considering the solution constructed in Theorem 5.5, we obtain

$$\|(\tan\Theta(\mathcal{X}, \hat{\mathcal{X}}), \tan\Theta(\mathcal{Y}, \hat{\mathcal{Y}}))\|_F \leq \frac{2\|\hat{\mathbf{T}}_s^{-1}\|\,\|(E_{21}, F_{21})\|_F}{1+\sqrt{1-\kappa}}. \tag{5.7}$$

The proof of Lemma 3.8 can be easily adapted to relate $\|\hat{\mathbf{T}}_s^{-1}\|$ to $\|\mathbf{T}_s^{-1}\|$.

LEMMA 5.6. [illegible] 3 [illegible] $\mathbf{T}$ [illegible] (5.2) [illegible] $\mathbf{T} : \mathcal{M} \to \mathcal{N}$ [illegible] $\mathbf{T}_s := \mathbf{T}|_{\mathcal{M}\to\mathcal{N}}$ [illegible] $1/\|\mathbf{T}_s^{-1}\| > \|(E_{11}, F_{11})\|_F + \|(E_{22}, F_{22})\|_F$ [illegible] $\hat{\mathbf{T}}_s$ [illegible]

$$\|\hat{\mathbf{T}}_s^{-1}\| \leq \frac{\|\mathbf{T}_s^{-1}\|}{1 - \|\mathbf{T}_s^{-1}\|(\|(E_{11}, F_{11})\|_F + \|(E_{22}, F_{22})\|_F)}. \tag{5.8}$$

Combining (5.7) and (5.8) implies $c_{(A,B)+\mathbb{L}}(\mathcal{X}, \mathcal{Y}) \leq \|\mathbf{T}_s^{-1}\|$. Assuming $\mathbf{T}^\star : \mathcal{N} \to \mathcal{M}$, it can be shown that $c_{(A,B)+\mathbb{L}}(\mathcal{X}, \mathcal{Y})$ and $\|\mathbf{T}_s^{-1}\|$ are equal.

THEOREM 5.7. [illegible] 3 [illegible] $\mathcal{M}$ [illegible] $\hat{\mathbf{T}}$ [illegible] $(E, F) \in \mathbb{L}$ [illegible] $\mathbf{T}$ [illegible] (5.2) [illegible] $\mathbf{T}^\star : \mathcal{N} \to \mathcal{M}$ [illegible] $\mathbf{T}_s := \mathbf{T}|_{\mathcal{M}\to\mathcal{N}}$ [illegible] $c_{(A,B)+\mathbb{L}}(\mathcal{X}, \mathcal{Y}) = \|\mathbf{T}_s^{-1}\|$

[illegible]. To adapt the proof of Theorem 3.9 to matrix pencils, we consider perturbations of the form $(E, F) = (\epsilon Y_\perp E_{21} X^H, \epsilon Y_\perp F_{21} X^H)$, where $(E_{21}, F_{21}) \in \mathcal{N}$ is chosen such that $\|(E_{21}, F_{21})\|_F = 1$ and $\|\mathbf{T}^{-1}(E_{21}, F_{21})\|_F = \|\mathbf{T}_s^{-1}\|$. The bound (5.7) implies for sufficiently small $\epsilon > 0$ that the nearest deflating subspace $(\hat{\mathcal{X}}, \hat{\mathcal{Y}})$ of $(A+E) - \lambda(B+F)$ satisfies

$$\max\{\|\Theta(\mathcal{X}, \hat{\mathcal{X}})\|_2, \|\Theta(\mathcal{Y}, \hat{\mathcal{Y}})\|_2\} < \pi/2.$$

This yields the existence of a matrix pair (R, L) such that the columns of $(\hat{X}, \hat{Y}) = (X - X_\perp R, Y - Y_\perp L)$ form bases for $(\hat{\mathcal{X}}, \hat{\mathcal{Y}})$. Equivalently, (R, L) satisfies

$$\mathbf{T}(R, L) + \mathbf{Q}(R, L) = (\epsilon E_{21}, \epsilon F_{21}),$$

where $\mathbf{Q}(R, L) = (L\hat{A}_{12}R, L\hat{B}_{12}R)$. Let us decompose $(R, L) = (R_s, L_s) + (R_u, L_u)$, where $(R_s, L_s) \in \mathcal{M}$ and $(R_u, L_u) \in \mathcal{M}^\perp$. Then $\mathbf{T}(R_s, L_s) \in \mathcal{N}$ and $\mathbf{T}(R_u, L_u) \in \mathcal{N}^\perp$. This implies, as in the proof of Theorem 3.9, that

$$\lim_{\epsilon\to 0} \|(R, L)\|_F/\epsilon \geq \lim_{\epsilon\to 0} \|(R_s, L_s)\|_F/\epsilon = \|\mathbf{T}^{-1}(E_{21}, F_{21})\|_F = \|\mathbf{T}_s^{-1}\|,$$

and consequently $c_{(A,B)+\mathbb{L}}(\mathcal{X}, \mathcal{Y}) \geq \|\mathbf{T}_s^{-1}\|$, which concludes the proof. ∎

5.4. Nonlinear structures. The following theorem shows that the results of sections 5.1 and 5.2 can also be used to address matrix pencil structures that form smooth manifolds.

THEOREM 5.8. $\mathbb{S} \subseteq \mathbb{K}^{n\times n} \times \mathbb{K}^{n\times n}$ $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ $(\mathcal{X}, \mathcal{Y})$ $A-\lambda B$ $(A, B) \in \mathbb{S}$ (5.1) $(\mathcal{X}, \mathcal{Y})$ $\mathbb{S}$

$$c_{\mathbb{S}}(\mathcal{X}, \mathcal{Y}) = \sup\big\{\|\mathbf{T}^{-1}(Y_\perp^H E X, Y_\perp^H F X)\|_F : \ (E, F) \in T_{(A,B)}\mathbb{S}, \|(E, F)\|_F = 1\big\},$$

$\mathbf{T}$ (5.2) $T_{(A,B)}\mathbb{S}$ $\mathbb{S}$ (A, B). The result follows from a rather straightforward extension of the proof of Theorem 3.10. □

5.5. Example: Palindromic matrix pencils. To illustrate the obtained results for structured matrix pencils, let us consider a matrix pencil of the form $A+\lambda A^T$ with $A \in \mathbb{C}^{2n\times 2n}$. A matrix pencil that takes this form is called ; it arises, e.g., from structure-preserving linearizations of palindromic matrix polynomials [16, 28]. The following result provides a structured Schur form.

LEMMA 5.9 (see [16]). $A \in \mathbb{C}^{2n\times 2n}$ $U \in \mathbb{C}^{2n\times 2n}$

$$U^T A U = \begin{bmatrix} 0 & \cdots & 0 & t_{1,2n} \\ & & t_{2,2n-1} & t_{2,2n} \\ 0 & & & \\ t_{2n,1} & t_{2n,2} & \cdots & t_{2n,2n} \end{bmatrix} =: T,$$

T

It should be emphasized that U^T in Lemma 5.9 denotes the complex transpose of U, i.e., $U^T AU$ is similar to A. Nevertheless, $T + \lambda T^T$ is equivalent to $A + \lambda A^T$, implying that the eigenvalues of $A + \lambda A^T$ are given by

$$-t_{1,2n}/t_{2n,1}, \ldots, -t_{n,n+1}/t_{n+1,n}, -t_{n+1,n}/t_{n,n+1}, \ldots, -t_{2n,1}/t_{1,2n}.$$

It follows immediately that the eigenvalues have the following pairing: λ is an eigenvalue of $A + \lambda A^T$ if and only if $1/\lambda$ is an eigenvalue. Zero eigenvalues are included in these pairings as $\lambda = 0$ and $1/\lambda = \infty$.

In the following, we consider the (right) deflating subspace $\mathcal{X}$ belonging to the eigenvalues $-t_{n+1,n}/t_{n,n+1}, \ldots, -t_{2n,1}/t_{1,2n}$. Let the columns of X and $X_\perp$ form orthonormal bases for $\mathcal{X}$ and $\mathcal{X}^\perp$, respectively. Then Lemma 5.9 implies a structured generalized block Schur decomposition of the form

$$[X_\perp, X]^T (A + \lambda A^T)[X, X_\perp] = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} + \lambda \begin{bmatrix} A_{22}^T & A_{12}^T \\ 0 & A_{11}^T \end{bmatrix} \tag{5.9}$$

with $A_{11}, A_{22} \in \mathbb{C}^{n\times n}$. Note that this also shows that $\overline{X_\perp}$, obtained from $X_\perp$ by conjugating its entries, spans a left deflating subspace $\mathcal{Y}$ belonging to the eigenvalues $-t_{n+1,n}/t_{n,n+1}, \ldots, -t_{2n,1}/t_{1,2n}$. We require the following preliminary result for obtaining the structured condition number of $(\mathcal{X}, \mathcal{Y})$ with respect to palindromic perturbations.

LEMMA 5.10. [illegible] $C, D \in \mathbb{C}^{n\times n}$ [illegible]

$$CR + \alpha R^T D^T = F, \tag{5.10}$$

[illegible] $\alpha \in \{1, -1\}$ [illegible] R [illegible] $F \in \mathbb{C}^{n\times n}$ [illegible] $C - \lambda D$

1. [illegible] $\lambda \neq \alpha$ [illegible] $1/\lambda$ [illegible] not [illegible]
2. [illegible] $\lambda = \alpha$ [illegible]

[illegible]. The proof can be found in the appendix. □

The generalized Sylvester operator associated with (5.9) takes the form

$$\mathbf{T} : (R_r, R_l) \mapsto (A_{22}R_r + R_l A_{11}, A_{11}^T R_r + R_l A_{22}^T).$$

Considering the linear space

$$\mathcal{N} := \left\{(X^T EX, -X^T E^T X) : \ E \in \mathbb{C}^{2n\times 2n}\right\} = \left\{(E_{21}, -E_{21}^T) : \ E_{21} \in \mathbb{C}^{n\times n}\right\},$$

we have $\mathbf{T} : \mathcal{N} \to \mathcal{N}$ and $\mathbf{T} : \mathcal{N}^\perp \to \mathcal{N}^\perp$, where $\mathcal{N}^\perp = \left\{(E_{21}, E_{21}^T) : \ E_{21} \in \mathbb{C}^{n\times n}\right\}$. Moreover, $(R_l A_{12} R_r, -R_l A_{12}^T R_r) \in \mathcal{N}$ for all $(R_l, R_r) \in \mathcal{N}$. The restricted Sylvester operators $\mathbf{T}_s = \mathbf{T}\big|_{\mathcal{N}\to\mathcal{N}}$ and $\mathbf{T}_u = \mathbf{T}\big|_{\mathcal{N}^\perp\to\mathcal{N}^\perp}$ can be identified with the matrix operators

$$\mathbf{S}_s : R \mapsto A_{22}R - R^T A_{11}, \quad \mathbf{S}_u : R \mapsto A_{22}R + R^T A_{11},$$

in the sense that

$$\mathbf{T}_s(R, -R^T) = (\mathbf{S}_s(R), -\mathbf{S}_s(R)^T), \quad \mathbf{T}_u(R, R^T) = (\mathbf{S}_u(R), \mathbf{S}_u(R)^T).$$

In particular, $\mathbf{T}_s$ is invertible if and only if $\mathbf{S}_s$ is invertible, which in turn is equivalent to require $A_{22} - \lambda A_{11}^T$ to satisfy the conditions of Lemma 5.10 for $\alpha = -1$. In this case, all assumptions of Theorem 5.7 are satisfied and the structured condition number for the deflating subspace pair $(\mathcal{X}, \mathcal{Y}) = (\mathrm{span}(X), \mathrm{span}(\overline{X_\perp}))$ with respect to $\mathbb{S} = \{(E, -E^T) : \ E \in \mathbb{C}^{2n\times 2n}\}$ is given by

$$c_{\mathbb{S}}(\mathcal{X}, \mathcal{Y}) = \|\mathbf{T}_s^{-1}\| = \sqrt{2}\,\|\mathbf{S}_s^{-1}\| = \frac{\sqrt{2}}{\inf\{\|A_{22}R - R^T A_{11}\|_F : \ R \in \mathbb{C}^{n\times n}, \|R\|_F = 1\}}.$$

On the other hand, the unstructured condition number satisfies

$$c(\mathcal{X}, \mathcal{Y}) = \sqrt{2}\,\max\{\|\mathbf{S}_s^{-1}\|, \|\mathbf{S}_u^{-1}\|\}.$$

This shows that the unstructured condition number can be much larger than the structured condition number, e.g., if $A_{22} - \lambda A_{11}^T$ has a simple eigenvalue close to -1. If one of the eigenvalues of $A_{22} - \lambda A_{11}^T$ happens to be exactly -1, then $(\mathcal{X}, \mathcal{Y})$ is not stable under unstructured perturbations, but Lemma 5.10 implies that it can still be stable under structured perturbations. In these cases, the use of a computational method that yields structured backward errors is likely to be significantly more accurate than other methods.

[illegible] 5.11. For $n = 1$, we obtain

$$\|\mathbf{S}_s^{-1}\| = \frac{1}{|A_{22} - A_{11}|}, \quad \|\mathbf{S}_u^{-1}\| = \frac{1}{|A_{22} + A_{11}|}.$$

Hence, if $A_{22}/A_{11} \approx -1$, then $c(\mathcal{X}, \mathcal{Y}) \gg c_{\mathbb{S}}(\mathcal{X}, \mathcal{Y})$.

6. Conclusions. We have derived directly computable expressions for structured condition numbers of invariant and deflating subspaces for smooth manifolds of structured matrices and matrix pencils. An orthogonal decomposition of the associated Sylvester operators yields global perturbation bounds that remain valid even in cases where the subspace is unstable under unstructured perturbations. It also provides additional insight into the difference between structured and unstructured condition numbers. We have identified structures for which this difference can be significant (block cyclic, palindromic) or negligible (Hamiltonian, orthogonal). Developing efficient structured condition estimators going beyond the simple method mentioned in Remark 3.4 remains an important future task.

The examples suggest some relation between structures that admit the proposed orthogonal decomposition approach and those that admit structured Schur decompositions. However, addressing this question thoroughly requires further investigation.

Appendix.

5.10. We only have to show the case $\alpha = 1$, as $\alpha = -1$ follows from $\alpha = 1$ after replacing D^T by $-D^T$. First, we prove by induction that (5.10) has a solution for any F if the two conditions hold. For $n = 1$, the first condition implies $C \neq -D$ and thus $R = F/(C + D)$. For $n > 1$, using the generalized Schur decomposition of $C - \lambda D$, we may assume without loss of generality that C and D have upper triangular form. Partition the matrices

$$C = \begin{bmatrix} C_{11} & C_{12} \\ 0 & C_{22} \end{bmatrix},\ D = \begin{bmatrix} D_{11} & D_{12} \\ 0 & D_{22} \end{bmatrix},\ F = \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix},\ R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$$

conformally with no void blocks, then (5.10) can be written as

$$F_{11} = C_{11}R_{11} + C_{12}R_{21} + R_{11}^T D_{11} + R_{21}^T D_{12}, \tag{A.1}$$

$$F_{21} = C_{22}R_{21} + R_{12}^T D_{11} + R_{22}^T D_{12}, \tag{A.2}$$

$$F_{12} = C_{11}R_{12} + C_{12}R_{22} + R_{21}^T D_{22}, \tag{A.3}$$

$$F_{22} = C_{22}R_{22} + R_{22}^T D_{22}. \tag{A.4}$$

By the induction assumption, the matrix equation (A.4) is solvable. Thus, R_{22} can be regarded as known, which turns (A.2)–(A.3), after transposing (A.3), into a generalized Sylvester equation associated with the matrix pencils $C_{22}+\lambda D_{11}^T$ and $D_{22}^T+\lambda C_{11}$. Under the given conditions these two pencils have no eigenvalue in common. Hence, (A.2)–(A.3) is solvable and R_{12} as well as R_{21} can be regarded as known. This turns (A.1) into a matrix equation of the form (5.10) of smaller dimension, which is—by the induction assumption—solvable. The uniqueness of the constructed solution R follows from the fact that (5.10) can be regarded as a square linear system of equations in the entries of R.

For the other direction, consider the linear matrix operator $\mathbf{S} : R \mapsto CR + R^T D^T$. We will make use of the fact that the matrix equation (5.10) is uniquely solvable if and only if kernel$(\mathbf{S}) = \{0\}$. Suppose that $\lambda = -1$ is an eigenvalue of $C - \lambda D$ and let x be an associated eigenvector. Then the nonzero matrix $R_0 = xx^T D^T$ satisfies

$$\mathbf{S}(R_0) = Cxx^T D^T + Dxx^T D^T = Cxx^T D^T - Cxx^T D^T = 0,$$

i.e., $R_0 \in$ kernel$(\mathbf{S})$. Now, suppose that $\lambda \neq -1$ and $1/\lambda$ are eigenvalues of $C - \lambda D$ and let x, y be corresponding eigenvectors such that x, y are linearly independent (for

$\lambda = 1$ this is only possible if λ has geometric multiplicity of at least 2). If $\lambda \neq 0$, the nonzero matrix $R_1 = xy^T D^T - yx^T C^T$ satisfies

$$\begin{aligned}\mathbf{S}(R_1) &= Cxy^T D^T - Cyx^T C^T + Dyx^T D^T - Cxy^T D^T \\ &= -\frac{1}{\lambda} Dyx^T C^T + \frac{1}{\lambda} Dyx^T C^T = 0.\end{aligned}$$

Analogously for $\lambda = 0$, the matrix $R_2 = xy^T C^T - yx^T D^T$ is nonzero and satisfies $\mathbf{S}(R_2) = 0$. It remains to be seen that kernel$(\mathbf{S}) \neq \{0\}$ holds if $\lambda = 1$ is an eigenvalue of $C - \lambda D$ with algebraic multiplicity of at least 2 but with geometric multiplicity 1. This is, however, an immediate consequence of the fact that $\mathbf{S}$ cannot be nonsingular at isolated points. Hence, if one of the two conditions of Lemma 5.10 is violated, then (5.10) is not uniquely solvable, which concludes the proof. ∎

Acknowledgments. The authors sincerely thank Ji-guang Sun for helpful remarks and discussions on an earlier version of this paper. Portions of this work were completed while the second author was enjoying the hospitality and support of the Department of Mathematics, University of Kansas.

REFERENCES

[1] R. H. Bartels and G. W. Stewart, *Algorithm* 432: *The solution of the matrix equation* $AX - BX = C$, Comm. ACM, 8 (1972), pp. 820–826.

[2] P. Benner, D. Kressner, and V. Mehrmann, *Skew-Hamiltonian and Hamiltonian eigenvalue problems: Theory, algorithms and applications*, in Proceedings of the Conference on Applied Mathematics and Scientific Computing, Brijuni, Croatia, Z. Drmač, M. Marušić, and Z. Tutek, eds., Springer-Verlag, 2005, 2003, pp. 3–39.

[3] P. Benner, V. Mehrmann, and H. Xu, *Perturbation analysis for the eigenvalue problem of a formal product of matrices*, BIT, 42 (2002), pp. 1–43.

[4] A. Bojanczyk, G. H. Golub, and P. Van Dooren, *The periodic Schur decomposition: Algorithm and applications*, in Proceedings of the SPIE Conference, San Diego, 1992, pp. 31–42.

[5] R. Byers, *A LINPACK-style condition estimator for the equation* $AX - XB^T = C$, IEEE Trans. Automat. Control, 29 (1984), pp. 926–928.

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

[7] R. Byers and S. Nash, *On the singular "vectors" of the Lyapunov operator*, SIAM J. Algebraic Discrete Methods, 8 (1987), pp. 59–66.

[8] F. Chatelin, *Eigenvalues of matrices*, John Wiley & Sons Ltd., Chichester, UK, 1993.

[9] P. I. Davies, *Structured conditioning of matrix functions*, Electron. J. Linear Algebra, 11 (2004), pp. 132–161.

[10] J. W. Demmel, *Computing stable eigendecompositions of matrices*, Linear Algebra Appl., 79 (1986), pp. 163–193.

[11] J. W. Demmel and B. Kågström, *Computing stable eigendecompositions of matrix pencils*, Linear Algebra Appl., 88/89 (1987), pp. 139–186.

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

[13] J. J. Hench and A. J. Laub, *Numerical solution of the discrete-time periodic Riccati equation*, IEEE Trans. Automat. Control, 39 (1994), pp. 1197–1210.

[14] D. J. Higham and N. J. Higham, *Structured backward error and condition of generalized eigenvalue problems*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 493–512.

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

[16] A. Hilliges, C. Mehl, and V. Mehrmann, *On the solution of palindromic eigenvalue problems*, in Proceedings of the Fourth European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS), Jyväskylä, Finland, 2004.

[17] M. E. Hochstenbach and B. Plestenjak, *Backward error, condition numbers, and pseudospectra for the multiparameter eigenvalue problem*, Linear Algebra Appl., 375 (2003), pp. 63–81.

[18] I. Jonsson and B. Kågström, *Recursive blocked algorithm for solving triangular systems,* I, *One-sided and coupled Sylvester-type matrix equations*, ACM Trans. Math. Software, 28 (2002), pp. 392–415.
[19] B. Kågström and P. Poromaa, *Distributed and shared memory block algorithms for the triangular Sylvester equation with* sep^{-1} *estimators*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 90–101.
[20] B. Kågström and P. Poromaa, *Computing eigenspaces with specified eigenvalues of a regular matrix pair* (A, B) *and condition estimation: theory, algorithms and software*, Numer. Algorithms, 12 (1996), pp. 369–407.
[21] B. Kågström and P. Poromaa, *LAPACK-style algorithms and software for solving the generalized Sylvester equation and estimating the separation between regular matrix pairs*, ACM Trans. Math. Software, 22 (1996), pp. 78–103.
[22] M. Karow, D. Kressner, and F. Tisseur, *Structured Eigenvalue Condition Numbers*, Numerical Analysis Report 467, Manchester Centre for Computational Mathematics, Manchester, England, 2005.
[23] M. Konstantinov, D.-W. Gu, V. Mehrmann, and P. Petkov, *Perturbation Theory for Matrix Equations*, Studies in Computational Mathematics 9, North-Holland, Amsterdam, 2003.
[24] M. Konstantinov, V. Mehrmann, and P. Petkov, *Perturbation analysis of Hamiltonian Schur and block-Schur forms*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 387–424.
[25] D. Kressner, *Numerical Methods and Software for General and Structured Eigenvalue Problems*, Ph.D. thesis, TU Berlin, Institut für Mathematik, Berlin, Germany, 2004.
[26] D. Kressner, *Perturbation bounds for isotropic invariant subspaces of skew-Hamiltonian matrices*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 947–961.
[27] W.-W. Lin and J.-G. Sun, *Perturbation analysis for the eigenproblem of periodic matrix pairs*, Linear Algebra Appl., 337 (2001), pp. 157–187.
[28] D. S. Mackey, N. Mackey, C. Mehl, and V. Mehrmann, *Palindromic Polynomial Eigenvalue Problems: Good Vibrations from Good Linearizations*, Numerical Analysis Report 466, Manchester Centre for Computational Mathematics, Manchester, England, 2005.
[29] V. Mehrmann, *The Autonomous Linear Quadratic Control Problem, Theory and Numerical Solution*, Lecture Notes in Control and Information Sciences 163, Springer-Verlag, Berlin, 1991.
[30] V. Mehrmann and D. S. Watkins, *Structure-preserving methods for computing eigenpairs of large sparse skew-Hamiltonian/Hamiltonian pencils*, SIAM J. Sci. Comput., 22 (2000), pp. 1905–1925.
[31] S. Noschese and L. Pasquini, *Eigenvalue condition numbers: Zero-structured versus traditional*, J. Comput. Appl. Math., 185 (2006), pp. 174–189.
[32] J. R. Rice, *A theory of condition*, SIAM J. Numer. Anal., 3 (1966), pp. 287–310.
[33] G. W. Stewart, *Error bounds for approximate invariant subspaces of closed linear operators*, SIAM J. Numer. Anal., 8 (1971), pp. 796–808.
[34] G. W. Stewart, *On the sensitivity of the eigenvalue problem* $Ax = \lambda Bx$, SIAM J. Numer. Anal., 9 (1972), pp. 669–686.
[35] G. W. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.
[36] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.
[37] J.-G. Sun, *Perturbation expansions for invariant subspaces*, Linear Algebra Appl., 153 (1991), pp. 85–97.
[38] J.-G. Sun, *Perturbation analysis of singular subspaces and deflating subspaces*, Numer. Math., 73 (1996), pp. 235–263.
[39] 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.
[40] J.-G. Sun, *Perturbation Analysis of Algebraic Riccati Equations*, Technical report UMINF 02-03, Department of Computing Science, University of Umeå, Umeå, Sweden, 2002.
[41] F. Tisseur, *A chart of backward errors for singly and doubly structured eigenvalue problems*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 877–897.
[42] D. S. Watkins, *Product eigenvalue problems*, SIAM Rev., 47 (2005), pp. 3–40.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 2, pp. 348–359

AN ALGORITHM TO COMPUTE $\mathbf{Sep}_{\boldsymbol{\lambda}}$*

MING GU† AND MICHAEL L. OVERTON‡

This paper is dedicated to James M. Varah on the occasion of his 60th birthday

Abstract. The following problem is addressed: given square matrices A and B, compute the smallest ϵ such that $A + E$ and $B + F$ have a common eigenvalue for some E, F with $\max(\|E\|_2, \|F\|_2) \leq \epsilon$. An algorithm to compute this quantity to any prescribed accuracy is presented, assuming that eigenvalues can be computed exactly.

Key words. eigenvalue perturbation, eigenvalue separation, pencil, pseudospectra

AMS subject classifications. 15A18, 15A22, 15A42, 65F15

DOI. 10.1137/050622584

1. Introduction. The quantity $\mathrm{sep}_\lambda(A, B)$ was introduced by Varah in [Var79] and further investigated by Demmel in [Dem83, Dem86, Dem87]; it measures how much perturbation is required to modify two square matrices A and B so that they have a common eigenvalue. Let $A \in \mathbf{C}^{m\times m}$ and $B \in \mathbf{C}^{n\times n}$. The functions studied by Varah and Demmel are defined slightly differently, namely

$$\mathrm{sep}_\lambda^V(A, B) = \min\{\epsilon \in \mathbf{R} : \exists\ E \in \mathbf{C}^{m\times m}, F \in \mathbf{C}^{n\times n} \text{ with } \|E\| + \|F\| \leq \epsilon \text{ such that } A + E \text{ and } B + F \text{ have a common eigenvalue}\} \tag{1}$$

and

$$\mathrm{sep}_\lambda^D(A, B) = \min\{\epsilon \in \mathbf{R} : \exists\ E \in \mathbf{C}^{m\times m}, F \in \mathbf{C}^{n\times n} \text{ with } \max(\|E\|, \|F\|) \leq \epsilon \text{ such that } A + E \text{ and } B + F \text{ have a common eigenvalue}\}, \tag{2}$$

respectively, where $\|\cdot\|$ denotes the 2-norm. Clearly,

$$\frac{1}{2}\mathrm{sep}_\lambda^V(A, B) \leq \mathrm{sep}_\lambda^D(A, B) \leq \mathrm{sep}_\lambda^V(A, B).$$

The lower bound is tight, with equality holding for normal matrices. A standard argument based on the singular value decomposition shows that

$$\mathrm{sep}_\lambda^V(A, B) = \min_{z\in\mathbf{C}} (\sigma_{\min}(A - zI) + \sigma_{\min}(B - zI)) \tag{3}$$

and

$$\mathrm{sep}_\lambda^D(A, B) = \min_{z\in\mathbf{C}} \max (\sigma_{\min}(A - zI), \sigma_{\min}(B - zI)) , \tag{4}$$

*Received by the editors January 12, 2005; accepted for publication (in revised form) December 8, 2005; published electronically April 21, 2006.
http://www.siam.org/journals/simax/28-2/62258.html

†Department of Mathematics, University of California, Berkeley, Berkeley, CA 94720 (mgu@math.berkeley.edu). This author's research was supported in part by National Science Foundation grant CCF-0515034.

‡Courant Institute of Mathematical Sciences, New York University, New York, NY 10012 (overton@cs.nyu.edu). This author's research was supported in part by National Science Foundation grant DMS-0412049 at NYU and in part by the Dehmel Distinguished Professorship Fund at the University of California, Berkeley.

where $\sigma_{\min}$ denotes least singular value. This also shows that the quantities sep_λ^V and sep_λ^D remain unchanged if the Frobenius norm is substituted for the 2-norm.

Sep_λ may also be defined in terms of pseudospectra. The ϵ-pseudospectrum of A is [ET]

$$\begin{aligned}\Lambda_\epsilon(A) &= \left\{z \in \mathbf{C} : \exists\ E \in \mathbf{C}^{m\times m} \text{ with } \|E\| \le \epsilon \text{ and } \det(A+E-zI) = 0\right\} \\ &= \left\{z \in \mathbf{C} : \sigma_{\min}(A-zI) \le \epsilon\right\},\end{aligned}$$

so $\mathrm{sep}_\lambda^V(A,B)$ is the minimal value of $\epsilon_1+\epsilon_2$ such that $\Lambda_{\epsilon_1}(A)\cap\Lambda_{\epsilon_2}(B)$ is nonempty, while $\mathrm{sep}_\lambda^D(A,B)$ is the minimal value of ϵ such that $\Lambda_\epsilon(A)\cap\Lambda_\epsilon(B)$ is nonempty. Indeed, Trefethen and Embree [TE05, section 1.6] attribute the earliest known definition of pseudospectra to Varah in his Ph.D. thesis [Var67] and the earliest published computer-generated pseudospectral plot to Demmel in [Dem87]. It is well known that the pseudospectrum $\Lambda_\epsilon(A)$ consists of at most n components,[1] and that each component is compact, contains at least one eigenvalue of A and has a piecewise smooth boundary; however, it may not be convex or even simply connected.

Obviously, $\mathrm{sep}_\lambda^V(A,B) = \mathrm{sep}_\lambda^D(A,B) = 0$ if and only if A and B have a common eigenvalue, and it is well known that this holds if and only if the Sylvester equation $AX-XB=0$ has a nontrivial solution $X \in \mathbf{C}^{m\times n}$, or equivalently, that the Kronecker difference $I\otimes A - B^T\otimes I$ is singular [HJ91, section 4.4].[2] Varah's notational choice sep_λ was inspired by its relationship to the quantity sep introduced by Stewart [Ste73] to study angles between subspaces,

$$\mathrm{sep}(A,B) = \min_{X\in\mathbf{C}^{m\times n}} \frac{\|AX-XB\|_F}{\|X\|_F} = \sigma_{\min}(I\otimes A - B^T\otimes I).$$

Varah observed that $\mathrm{sep}(A,B) \le \mathrm{sep}_\lambda^V(A,B)/2$ (so $\mathrm{sep}(A,B) \le \mathrm{sep}_\lambda^D(A,B)$) but that very often, sep and sep_λ differ by several orders of magnitude. This fact is related to the now well known one that pseudospectra and spectra provide very different information for nonnormal matrices, which is the theme of the comprehensive book [TE05]. Thus, even if one is prepared to compute $\mathrm{sep}(A,B)$ via the singular value decomposition of $I\otimes A - B^T\otimes I$, a computation whose complexity is roughly $O(m^3n^3)$ flops, this does not provide a very useful lower bound for $\mathrm{sep}_\lambda(A,B)$.

Upper bounds for sep_λ are immediately obtained by evaluating $\sigma_{\min}(A-zI)$ and $\sigma_{\min}(B-zI)$ for any $z\in\mathbf{C}$, or, more effectively, by applying an optimization method to carry out the minimization in (3) or (4) respectively, perhaps initialized at many systematically generated starting points. However, even though there are only two real variables in each of these minimization problems, solving them is not easy. The main difficulty is that the optimization objectives are nonconvex and may have many local minimizers. No bound is known on the number of possible local minimizers, although it seems a good guess that $m+n$ (or at least its square) might be an upper bound, based on related recent results and conjectures [BLO04]. A second, less crucial, difficulty is that (for reasons to be seen in the next section) the optimization objective in (4) is virtually always nondifferentiable at a local optimizer, and while this may not be the case for the objective in (3), it will be if, as often happens, the local optimizer is an eigenvalue of A or B (i.e., the minimum in (1) is attained with either $E=0$ or $F=0$). This second difficulty may be overcome by using a method for nonsmooth, nonconvex optimization such as that described in [BLO05] instead of a

[1] Throughout, we use *component* to mean *connected component.*

[2] The size of the identity matrix I is context-dependent.

standard method for smooth, nonconvex optimization such as BFGS, but the inability to verify global optimality remains a stumbling block preventing the computation of sep_λ, or even the assessment of the quality of upper bounds, via optimization.

For these reasons, no algorithm to compute sep_λ^V or sep_λ^D or to reliably approximate them has appeared to date. In this paper, we give an algorithm to compute sep_λ^D to any specified accuracy in $O((m+n)m^3n^3)$ flops. Here we are adopting the usual convention for approximate floating point complexity estimates, taking the computation of the eigenvalues of an $m \times m$ matrix or pencil to be an atomic operation requiring $O(m^3)$ flops, and assuming that such eigenvalues are delivered exactly. The main idea is borrowed from an algorithm of Gu [Gu00] for approximating the distance from a matrix pair to the set of "uncontrollable" pairs. Gu's algorithm was later refined to approximate the uncontrollability distance to any prescribed accuracy [BLO04]. As it happens, our algorithm to compute sep_λ^D is substantially less complicated than the algorithm to compute the uncontrollability distance, so readers interested in the latter may find our description of the former to be a good introduction.

The new algorithm to compute sep_λ^D obviously approximates sep_λ^V within a factor of two; we do not see any way to improve this at present. Optimization experiments indicate that very often, e.g., for many randomly generated triangular matrices, $\mathrm{sep}_\lambda^V(A,B)$ equals the trivial upper bound

$$u(A,B) = \min\left(\min_{z\in\Lambda_0(B)} \sigma_{\min}(A - zI), \min_{z\in\Lambda_0(A)} \sigma_{\min}(B - zI)\right).$$

It is tempting to conjecture on the basis of such experiments that $\mathrm{sep}_\lambda^V(A,B)$ can never be much less than $u(A,B)$, and if this were true, it would provide an easy way to approximate $\mathrm{sep}_\lambda^D(A,B)$ as well. However, this is not the case, as can be seen by setting both A and B to Jordan blocks of the same size, with eigenvalues 0 and 1, respectively. Then the objectives in (3) and (4) are both minimized at $z = 0.5$ with $\mathrm{sep}_\lambda^V(A,B) = 2\ \mathrm{sep}_\lambda^D(A,B)$, and $\mathrm{sep}_\lambda^V(A,B)/u(A,B) \to 0$ exponentially as $m \to \infty$.

The importance of the quantity sep_λ is that it measures the distance from a pair (A,B) to the set of pairs $(A+E, B+F)$ for which the corresponding Sylvester equation is singular (i.e., $(A+E)X - X(B+F) = 0$ has a nontrivial solution X). The generic subject of computing the distance from a given matrix or matrix pair to the set of matrices or matrix pairs with certain undesirable properties, such as singularity, instability, or uncontrollability, has been a frequent theme in the literature, one that has been intensively studied and applied by the robust control community in various contexts. We note that Alam and Bora [AB05] have recently proved a result that uses pseudospectra to characterize the so-called Wilkinson distance, i.e., the distance from a matrix to the set of matrices with a multiple eigenvalue, a problem also studied in [Dem83, Dem86]. While computing the Wilkinson distance is superficially similar to the problem of computing sep_λ, it seems to be fundamentally harder. It is perhaps worth mentioning that in applications, lower bounds for such distance functions are more important than upper bounds, as they provide "safety margins." Even though the optimization approach mentioned above often provides good upper bounds on sep_λ, one can never be sure without good lower bounds. Prior to this work, the only nontrivial known lower bound on sep_λ was provided by sep, which, as already noted, is often a poor lower bound despite requiring $O(m^3n^3)$ flops for its computation.

2. The algorithm. For the remainder of the paper we drop the superscript in sep_λ^D and take (2) (equivalently (4)) as the definition of sep_λ. Assume that A and B have no common eigenvalue, so that $\mathrm{sep}_\lambda(A,B) > 0$. The first key observation, based

on the maximum modulus principle, is that the only local minimizers of $\sigma_{\min}(A - zI)$ as a function of z are the eigenvalues of A [BLO03, Theorem 4.2]. Consequently, local minimizers of (4) can be achieved only at a point z where, for some $\epsilon > 0$,

$$\epsilon = \sigma_{\min}(A - zI) = \sigma_{\min}(B - zI). \tag{5}$$

Such points are exactly those where the boundaries of $\Lambda_\epsilon(A)$ and $\Lambda_\epsilon(B)$ intersect, and $\mathrm{sep}_\lambda(A, B)$ is precisely the smallest such value of ϵ, i.e.,

$$\mathrm{sep}_\lambda(A, B) = \min\{\epsilon : \epsilon = \sigma_{\min}(A - zI) = \sigma_{\min}(B - zI) \text{ for some } z \in \mathbf{C}\}. \tag{6}$$

The next key observation is that given any component of $\Lambda_\epsilon(A)$ and any component of $\Lambda_\epsilon(B)$, one of three conditions must hold: they are disjoint, their boundaries intersect, or one is strictly inside the other. Thus, for any given $\epsilon > 0$, at least one of the following three conditions holds:

- $\Lambda_\epsilon(A)$ and $\Lambda_\epsilon(B)$ are disjoint, in which case there does not exist any z satisfying (5).
- The boundaries of $\Lambda_\epsilon(A)$ and $\Lambda_\epsilon(B)$ intersect, in which case there exists z satisfying (5).
- There is a component of $\Lambda_\epsilon(A)$ that lies strictly inside a component of $\Lambda_\epsilon(B)$ or vice versa, in which case there may or may not exist z satisfying (5).

The basic idea of the algorithm is to first determine an upper bound U on $\mathrm{sep}_\lambda(A, B)$ such that, for all $\epsilon \le U$, the third possibility is excluded (we explain how later), and then use a bisection method based on deciding which of the first and second cases hold. Once the third case is excluded, the nonexistence of z satisfying (5) implies that $\Lambda_\epsilon(A)$ and $\Lambda_\epsilon(B)$ are disjoint, so that $\mathrm{sep}_\lambda(A, B) > \epsilon$, while the existence of such a z obviously implies that $\mathrm{sep}_\lambda(A, B) \le \epsilon$.

Figure 1 illustrates the situation for a specific pair A and B. Both are randomly generated complex triangular 10×10 matrices. The real and imaginary parts of the entries of A are generated from the uniform distribution on $[-1, 1]$, while those of B come from the uniform distribution on $[-0.5, 0.5]$. The eigenvalues of A are plotted as crosses and those of B as dots. The four subfigures show the boundaries of the pseudospectra $\Lambda_\epsilon(A)$ (solid curve) and $\Lambda_\epsilon(B)$ (dotted curve) for four different values of ϵ. At the top left, $\epsilon = 0.5\ \mathrm{sep}_\lambda(A, B)$, so $\Lambda_\epsilon(A)$ and $\Lambda_\epsilon(B)$ are disjoint. At the top right, $\epsilon = \mathrm{sep}_\lambda(A, B)$, so the boundaries of $\Lambda_\epsilon(A)$ and $\Lambda_\epsilon(B)$ are tangent to each other at one point, but do not cross. At the bottom left, $\epsilon = 5\ \mathrm{sep}_\lambda(A, B)$, for which the boundaries of $\Lambda_\epsilon(A)$ and $\Lambda_\epsilon(B)$ cross each other. At the bottom right, $\epsilon = 15\ \mathrm{sep}_\lambda(A, B)$, for which $\Lambda_\epsilon(B)$ lies inside $\Lambda_\epsilon(A)$.

Given a value ϵ, how do we determine the points $z = x + iy$, if any, where the boundaries of $\Lambda_\epsilon(A)$ and $\Lambda_\epsilon(B)$ intersect, i.e., (5) holds? Following Byers [Bye88], we observe that $A - (x + iy)I$ has a singular value (not necessarily the least one) equal to ϵ if and only if

$$\begin{bmatrix} \epsilon I & A - (x + iy)I \\ A^* - (x - iy)I & \epsilon I \end{bmatrix}$$

is singular, or equivalently, postmultiplying by the canonical skew symmetric matrix, that the Hamiltonian matrix

$$G(x) = \begin{bmatrix} A - xI & -\epsilon I \\ \epsilon I & -A^* + xI \end{bmatrix} \tag{7}$$

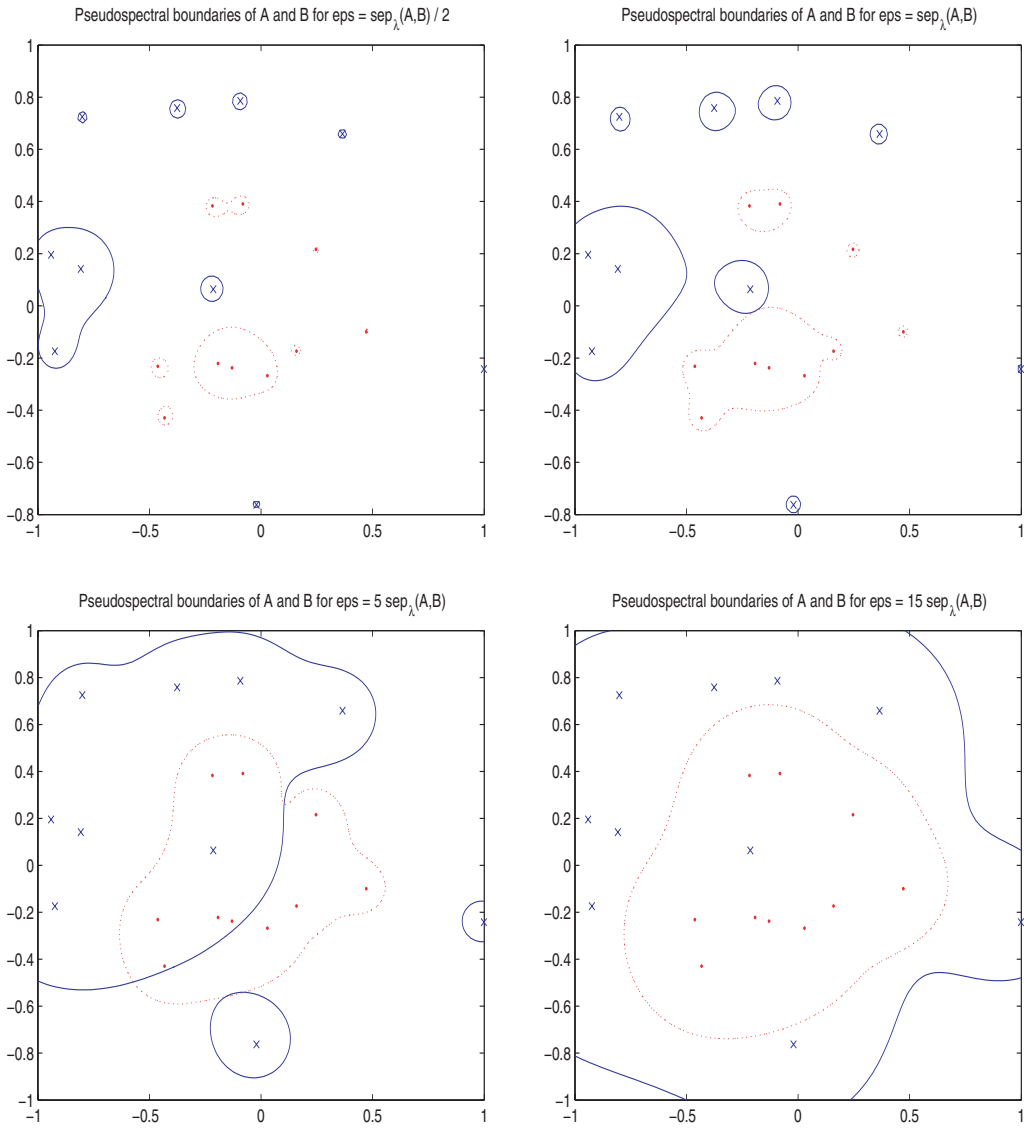


FIG. 1. *The boundaries of the pseudospectra* $\Lambda_\epsilon(A)$ *(solid curve) and* $\Lambda_\epsilon(B)$ *(dotted curve) for* ϵ *equal to half, one, five, and fifteen times* $\mathrm{sep}_\lambda(A,B)$*, respectively.*

has an imaginary eigenvalue iy. Likewise, $B-(x+iy)I$ has a singular value ϵ if and only if the Hamiltonian matrix

$$H(x) = \begin{bmatrix} B - xI & -\epsilon I \\ \epsilon I & -B^* + xI \end{bmatrix} \tag{8}$$

has an imaginary eigenvalue iy. Furthermore, $G(x)$ and $H(x)$ have a common eigenvalue (not necessarily imaginary) if and only if

$$\det\left(I \otimes G(x) - H(x)^T \otimes I\right) = 0. \tag{9}$$

This equation is a generalized eigenvalue problem in x; to see this, write

$$G(x) = G_1 - xG_2, \quad H(x) = H_1 - xH_2,$$

and let

$$K = I \otimes G_1 - H_1^T \otimes I, \quad L = I \otimes G_2 - H_2^T \otimes I, \tag{10}$$

with $K, L \in \mathbf{C}^{4mn \times 4mn}$. Then the solutions x of (9) are the roots of $\det(K - xL)$. Since L is singular, the generalized eigenvalue problem is not trivially convertible to an ordinary one. However, using the assumption that A and B have no common eigenvalue, it can be shown (see section 3) that the pencil $K - xL$ is regular, i.e., its determinant is not identically zero for all x, and of its $4mn$ eigenvalues, half are finite and half are infinite. Under our assumptions, the eigenvalues can be computed in $O(m^3n^3)$ flops. Thus we have the following algorithm to find all solutions of (5), assuming that eigenvalues and singular values can be computed exactly.

ALGORITHM 1.

Input: $A \in \mathbf{C}^{m \times m}$ $B \in \mathbf{C}^{n \times n}$ $\epsilon \in \mathbf{R}$, $\epsilon > 0$

Output: $z \in \mathbf{C}$ (5)

1. $K - xL$ x $\det(K - xL)$ (10)
2. x $G(x)$ $H(x)$ y $G(x)$ $H(x)$ iy (7) (8)
3. (x, y) $z = x + iy$ $A - zI$ $B - zI$ ϵ z (5) (x, y) (5)

An easy mistake to make in implementing this algorithm is to use the conjugate transpose H_1^* and H_2^* in place of the ordinary transpose H_1^T, H_2^T in (10).

Algorithm 1 provides the basis of a bisection method to compute $\mathrm{sep}_\lambda(A, B)$. This requires initialization with lower and upper bounds; either 0 or $\mathrm{sep}(A, B)$ (see section 1) can be used for the initial lower bound. We choose an initial upper bound U for which we can guarantee that, for all $\epsilon \le U$, $\Lambda_\epsilon(A)$ $\Lambda_\epsilon(B)$. With this initialization a bisection method based on Algorithm 1 must converge to $\mathrm{sep}_\lambda(A, B)$. The question remaining then is how to determine a value U that has the desired property.

Let $\mathcal{L}$ denote a line in the complex plane and consider the problem (6) restricted to the line $\mathcal{L}$, i.e., the problem of computing

$$\gamma_{\mathcal{L}} = \min \{\gamma : \gamma = \sigma_{\min}(A - zI) = \sigma_{\min}(B - zI) \text{ for some } z \in \mathcal{L}\}. \tag{11}$$

Now let $\theta \in [0, \pi)$ be fixed and consider the $m + n$ lines parameterized by

$$\mathcal{L}_j = \{z : z = \mu_j + te^{i\theta} \text{ for some } t \in \mathbf{R}\}, \tag{12}$$

where $\mu_j, j = 1, \ldots, m + n$, are the eigenvalues of A and B. Define

$$U = \min_{1 \le j \le m+n} \gamma_{\mathcal{L}_j}. \tag{13}$$

We claim that this value of U has the desired property. If not, then for some $\epsilon \le U$, a pseudospectral component of one of the matrices, say a component C_A of $\Lambda_\epsilon(A)$,

lies strictly inside a pseudospectral component of the other, say a component C_B of $\Lambda_\epsilon(B)$. There must be an eigenvalue of A, say μ_j, lying inside the inner component C_A. The line $\mathcal{L}_j$ passing through μ_j must intersect the boundary of C_A at two or more points. At $z = \mu_j$, $0 = \sigma_{\min}(A - zI) < \sigma_{\min}(B - zI)$, but at the points z where the line crosses the boundary of C_A, we have $\epsilon = \sigma_{\min}(A - zI) > \sigma_{\min}(B - zI)$. Thus, by continuity of $\sigma_{\min}$, $0 < \sigma_{\min}(A - zI) = \sigma_{\min}(B - zI) < \epsilon$ for some z on the line $\mathcal{L}_j$ and strictly contained in C_A. This contradicts the definition of U.

In order to solve (11) on the line $\mathcal{L}_j$ we need to determine all real quantities t and γ for which, setting $z = \mu_j + te^{i\theta}$, we have $\sigma_{\min}(A - zI) = \sigma_{\min}(B - zI) = \gamma$. We want the least such γ. A necessary condition for the two least singular values to equal each other is that

$$M(t) = \begin{bmatrix} 0 & A - (\mu_j + te^{i\theta})I \\ A^* - (\bar{\mu}_j + te^{-i\theta})I & 0 \end{bmatrix} \tag{14}$$

and

$$N(t) = \begin{bmatrix} 0 & B - (\mu_j + te^{i\theta})I \\ B^* - (\bar{\mu}_j + te^{-i\theta})I & 0 \end{bmatrix} \tag{15}$$

have a common eigenvalue, i.e., that

$$\det\left(I \otimes M(t) - N(t)^T \otimes I\right) = 0. \tag{16}$$

As earlier, this is a generalized eigenvalue problem; to see this, write

$$M(t) = M_1 - tM_2, \quad N(x) = N_1 - tN_2$$

and let

$$P = I \otimes M_1 - N_1^T \otimes I, \quad Q = I \otimes M_2 - N_2^T \otimes I. \tag{17}$$

Then the solutions t of (16) are the roots of $\det(P - tQ)$. Whether or not the pencil $P - tQ$ is regular depends on the choice of the angle θ defining the line through the eigenvalue μ_j; see section 3 for details. Provided that θ is chosen correctly, the pencil $P - tQ$ is regular with $2mn$ finite and $2mn$ infinite eigenvalues, and under our assumptions the eigenvalues can be computed in $O(m^3n^3)$ flops. For every finite real eigenvalue t, we set $z = \mu_j + te^{i\theta}$ and check whether $\sigma_{\min}(A - zI) = \sigma_{\min}(B - zI)$; we then set $\gamma_{\mathcal{L}_j}$ to be the smallest such common value. This process is summarized in Algorithm 2.

ALGORITHM 2.

Input: $A \in \mathbf{C}^{m\times m}$ $B \in \mathbf{C}^{n\times n}$ $\theta \in \mathbf{R}$ $j \in \{1, 2, \ldots, m + n\}$

Output: $\gamma_{\mathcal{L}_j}$ (11) (12)

1. [illegible] $P - tQ$ [illegible] t [illegible] $\det(P - tQ)$ (17)
2. [illegible] t [illegible] $z = \mu_j + te^{i\theta}$ [illegible] $\sigma_{\min}(A - zI) = \sigma_{\min}(B - zI)$. [illegible] $\gamma_{\mathcal{L}_j}$ [illegible] ∞ [illegible]

We are now ready to state the complete algorithm. Unfortunately, to exclude the possibility of [illegible] pseudospectral component of A lying strictly inside one of B or vice versa, we must carry out the steps in Algorithm 2 a total of $m + n$ times,[3] making the total cost $O((m + n)m^3n^3)$ flops.

[3]Of course, this can be somewhat reduced if the lines $\mathcal{L}_j$ are not all distinct.

ALGORITHM 3.
Input: $A \in \mathbf{C}^{m\times m}$ $B \in \mathbf{C}^{n\times n}$ $\tau \in \mathbf{R}$ $\tau > 0$
Output: L U $L \le \mathrm{sep}_\lambda(A,B) \le U$ $U - L \le \tau$
1. $j = 1,\ldots,m+n$ $\theta \in [0,2\pi)$ 2 $\gamma_{\mathcal{L}_j}$
2. $L = 0$ $L = \sigma_{\min}(I\otimes A - B^T\otimes I)$ $U = \min_{1\le j\le m+n}\gamma_{\mathcal{L}_j}$ ∞ $\max(\sigma_{\min}(A),\sigma_{\min}(B))$
3. $U - L > \tau$
   (a) $\epsilon = (L+U)/2$ 1 (5)
   (b) $U = \epsilon$ $L = \epsilon$

Under the assumptions that A and B have no common eigenvalue, that the pencils encountered by Algorithm 2 are all regular, and that all eigenvalue and singular value computations are exact, Algorithm 3 is guaranteed to approximate $\mathrm{sep}_\lambda(A,B)$ to any prescribed accuracy.

3. Further details. A MATLAB implementation of Algorithm 3 is freely available.[4] The eigenvalues of the pencils $K - xL$ (see (10)) and $P - tQ$ (see (17)) are computed by calls to the standard MATLAB eigensolver, i.e., by `eig(K,L)` and `eig(P,Q)`, respectively. However, it is of interest for several reasons to consider these generalized eigenvalue problems in more detail.

Let us start with taking a more careful look at the pencil $K - xL$. By definition, x satisfies $\det(K - xL) = 0$ if and only if the matrix equation

$$G(x)T - TH(x) = 0 \tag{18}$$

has a nontrivial solution T, where $G(x)$ and $H(x)$ were defined in (7), (8). Let

$$T = \begin{bmatrix} V & W \\ Y & Z \end{bmatrix}. \tag{19}$$

The eigenvalue parameter x vanishes from the (1,1) and (2,2) blocks of (18) because of cancellation; these blocks reduce to

$$AV - VB = \epsilon(W+Y) \qquad \text{and} \qquad A^*Z - ZB^* = \epsilon(W+Y).$$

These are Sylvester equations defining V and Z in terms of W and Y; furthermore, they are nonsingular (i.e., V and Z are uniquely defined by any W and Y) because of the assumption that A and B do not have a common eigenvalue. Thus we need only find x such that the (1,2) and (2,1) block equations in (18) hold. These equations are

$$AW + WB^* + \epsilon(V - Z) = 2xW$$

and

$$A^*Y + YB - \epsilon(V - Z) = 2xY.$$

Because V and Z depend linearly on W and Y, these equations together reduce to an eigenvalue problem of size $2mn$ with eigenvalue parameter x and eigenvector $[\mathrm{vec}(W);\mathrm{vec}(Y)]$. There are therefore $2mn$ (not necessarily distinct) eigenvalues. This proves that the pencil $K - xL$ is regular with $2mn$ finite and $2mn$ infinite eigenvalues.

[4] http://www.cs.nyu.edu/overton/faculty/software/seplambda

The pencil $P - tQ$ is more complicated. By definition, t satisfies $\det(P - tQ) = 0$ if and only if the matrix equation

$$M(t)T - TN(t) = 0 \tag{20}$$

has a nontrivial solution T, where $M(t)$ and $N(t)$ were defined in (14), (15), and we again partition T by (19). For brevity, let $\hat{A} = A - \mu_j I$ and $\hat{B} = B - \mu_j I$. The (1,1) and (2,2) block equations of (20) are

$$(\hat{A} - te^{i\theta}I)Y = W(\hat{B} - te^{i\theta}I)^* \tag{21}$$

and

$$(\hat{A} - te^{i\theta}I)^*W = Y(\hat{B} - te^{i\theta}I). \tag{22}$$

Adding these equations, the terms involving t cancel and we obtain

$$\hat{A}Y - Y\hat{B} = -\hat{A}^*W + W\hat{B}^*. \tag{23}$$

Because A and B (and therefore $\hat{A}$ and $\hat{B}$) have no common eigenvalue, it follows that Y is uniquely defined in terms of W (or vice versa) by solving a Sylvester equation. Now it also follows from (21) and (22) that

$$(\hat{A} - te^{i\theta}I)(\hat{A} - te^{i\theta}I)^*W = W(\hat{B} - te^{i\theta}I)^*(\hat{B} - te^{i\theta}I),$$

which simplifies to

$$\hat{A}\hat{A}^*W - W\hat{B}^*\hat{B} = t(e^{i\theta}(\hat{A}^*W - W\hat{B}^*) + e^{-i\theta}(\hat{A}W - W\hat{B})). \tag{24}$$

This is a generalized eigenvalue problem in the eigenvalue parameter t and eigenvector $\mathrm{vec}(W)$. It can be reduced to an ordinary eigenvalue problem provided that the linear operator defining the right-hand side in terms of W is invertible. This linear operator is a weighted sum of two nonsingular linear operators, since the equation $\hat{A}W - W\hat{B} = 0$ has only the trivial solution $W = 0$ (since $\hat{A}$ and $\hat{B}$ have no common eigenvalue) and the same is true for the equation $\hat{A}^*W - W\hat{B}^* = 0$. Clearly it is possible to choose θ so that the weighted sum of these two linear operators is also nonsingular; we call this condition the [illegible] condition on θ. Thus as long as the first condition on θ holds, there are mn (not necessarily distinct) eigenvalues t corresponding to the eigenvector $\mathrm{vec}(W)$, and W then uniquely determines Y from (23).

We now turn to the (1,2) and (2,1) block equations of (20). These are

$$(\hat{A} - te^{i\theta}I)Z = V(\hat{B} - te^{i\theta}I) \tag{25}$$

and

$$(\hat{A} - te^{i\theta}I)^*V = Z(\hat{B} - te^{i\theta}I)^*. \tag{26}$$

Adding $e^{-i\theta}$ times (25) to $e^{i\theta}$ times (26) yields

$$(e^{-i\theta}\hat{A})Z - Z(e^{i\theta}\hat{B}^*) = -(e^{i\theta}\hat{A}^*)V + V(e^{-i\theta}\hat{B}), \tag{27}$$

with all terms involving the eigenvalue parameter t cancelling as earlier. To be able to always solve this equation uniquely for Z in terms of V, or vice versa, we need the

following condition to hold: $e^{-i\theta}\hat{A}$ and $e^{i\theta}\hat{B}^*$ have no common eigenvalue. We call this the [illegible] condition on θ. Clearly it is possible to choose θ so that the second condition, as well as the first condition, holds. Finally, it also follows from (25) and (26) that

$$(\hat{A} - te^{i\theta}I)(\hat{A} - te^{i\theta}I)^*V = V(\hat{B} - te^{i\theta}I)(\hat{B} - te^{i\theta}I)^*,$$

which simplifies to

$$\hat{A}\hat{A}^*V - V\hat{B}\hat{B}^* = t(e^{i\theta}(\hat{A}^*V - V\hat{B}^*) + e^{-i\theta}(\hat{A}V - V\hat{B})).$$

This is a generalized eigenvalue problem in the eigenvalue parameter t and eigenvector $\mathrm{vec}(V)$ with exactly the same structure as (24). Provided the first condition on θ holds, this reduces to an ordinary eigenvalue problem, with the same mn eigenvalues t corresponding to the eigenvector $\mathrm{vec}(V)$ that we obtained corresponding to the eigenvector $\mathrm{vec}(W)$ previously. Furthermore, provided the second condition on θ holds, V uniquely defines Z from (27).

It turns out that when A and B are diagonal, the first condition states that θ should not be the angle of the perpendicular bisector of any of the line segments joining an eigenvalue of $\hat{A}$ to an eigenvalue of $\hat{B}$, and the second condition states that θ should not be the angle of any such perpendicular bisector that contains the origin (in fact, this characterization of the second condition does not require A and B to be diagonal). It is the second condition that is relevant to the problem of solving (11). For example, suppose

$$A = \begin{bmatrix} 0 & 0 \\ 0 & -0.1i \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 0 \\ 0 & 0.1i \end{bmatrix},$$

with $\mu_j = 0$, so $\hat{A} = A$, $\hat{B} = B$. Geometrically, it is clear that the corresponding pencil $P - tQ$ [illegible] be singular when $\theta = 0$, because there is a continuum of points z on the real axis where the boundaries of $\Lambda_\epsilon(A)$ and $\Lambda_\epsilon(B)$ intersect for some ϵ. Indeed, the second condition on θ is precisely $\theta \neq 0$, or geometrically, that θ should not be the angle of the perpendicular bisector of the line segment $[-0.1i, 0.1i]$. But what then is the significance of the first condition on θ? To understand this, recall that $\det(P - tQ) = 0$ if and only if the matrices (14) and (15) have a common eigenvalue, or equivalently that $\hat{A} - te^{i\theta}$ and $\hat{B} - te^{i\theta}$ have a common singular value—but this is only a [illegible] condition for these two matrices to have a common [illegible] singular value. Thus, most of the restrictions on θ have nothing to do with pseudospectra, but comprise a technical condition that ensures that the pencil $P - tQ$ is nonsingular. Even the second condition may not be relevant to (11), as we see if we change the (2,2) entries of A and B to $-10i$ and $10i$, respectively, or add a nonzero upper triangular entry to A or B.

Thus, as long as θ is chosen correctly (and choosing it randomly will almost certainly be adequate), the pencil $P - tQ$ is guaranteed to be regular, with $2mn$ finite eigenvalues (mn pairs of double eigenvalues) and $2mn$ infinite eigenvalues. In practice, even when the pencil is singular, as in the example given above, rounding comes to our assistance, so always using $\theta = 0$ seems adequate. Indeed, for randomly generated A with $\|A\| \approx 1$, the algorithm typically approximates $\mathrm{sep}_\lambda(A, A^T) = 0$ to about machine precision, although the basic assumption that A and B have no common eigenvalue is violated.

4. Concluding remarks. We conclude the paper with brief discussions of two key issues: efficiency and numerical stability.

When a QR-based method such as the one invoked by the MATLAB function `eig` is used to compute the eigenvalues of the pencils $K - xL$ and $P - tQ$, the complexity of the algorithm is, as already noted, $O((m+n)m^3n^3)$. This can potentially be reduced by using an iterative method with a shift-and-invert preconditioner based on a Sylvester solver, allowing exploitation of the structure of the generalized eigenvalue problems discussed in the previous section. Since γ real eigenvalues must be found, one might well doubt whether such an approach would work in practice. Nonetheless, a novel divide-and-conquer approach to searching for real eigenvalues, introduced recently in [GMO+06], works very well in the context of computing the distance to uncontrollability, where the issues are similar: the key step is computing all real eigenvalues of a large structured generalized eigenvalue problem. Although there are some inevitable difficulties with the numerical stability of this approach, the complexity drops significantly. For computing the distance to uncontrollability of a matrix pair (A, B), where A is $p \times p$ and B is $p \times q$, with $q \leq p$, the complexity drops from $O(p^6)$ to $O(p^5)$ in the worst case and to $O(p^4)$ on average (both in theory and in practice). For computing $\mathrm{sep}_\lambda(A, B)$, where A and B are both $m \times m$, the analogous drop in complexity would be from $O(m^7)$ to $O(m^6)$ in the worst case and $O(m^5)$ on average, but this has not been implemented.

On the other hand, even using a QR-based algorithm to compute the eigenvalues is not enough to ensure numerical stability of the new algorithm. In order to obtain a numerically stable algorithm, it seems essential to exploit the skew-Hamiltonian structure of the pencils $K - xL$ and $P - tQ$. Assuming $\theta = 0$, the finite eigenvalues of these pencils have skew-Hamiltonian symmetry around the real axis: those that are not real occur in complex conjugate pairs (regardless of whether A and B are real). The MATLAB function `eig` does not exploit this symmetry and hence real eigenvalues often have small imaginary rounding errors, occasionally defeating the test in the code that checks whether they are real and therefore returning invalid lower bounds. Ideally one would like to use a skew-Hamiltonian generalized eigensolver that exploits symmetry and delivers real eigenvalues with no imaginary rounding errors. Likewise, one should use a Hamiltonian eigensolver to compute the eigenvalues of the Hamiltonian matrices $G(x)$ and $H(x)$ in Step 2 of Algorithm 1, delivering imaginary eigenvalues with no real rounding errors. The design of such specialized eigensolvers has been a very active research area in recent years [MW01, BKM04].

In summary, an algorithm to compute sep_λ to arbitrary accuracy has been described, assuming that eigenvalues and singular values can be computed exactly. Since this assumption is very much an idealized one, some interesting questions regarding implementation of the algorithm remain open for future investigation.

Acknowledgments. This paper is dedicated to Jim Varah, who introduced the second author to the world of numerical linear algebra at UBC three decades ago. The second author also warmly thanks Jim Demmel for hosting him during a stimulating and productive sabbatical semester at Berkeley in Fall 2003, and for encouraging this work. Both authors particularly thank Gene Golub for organizing the SVG meeting at Stanford in January 2004, honoring the 60th birthdays of Jim Varah, Alan George, and Mike Saunders, thus providing the stimulus that led to this research and its presentation at that meeting. Finally, the second author thanks Adrian Lewis for discussions on sep_λ in 2001 that led to investigation of upper bounds via optimization.

REFERENCES

[AB05] R. Alam and S. Bora, *On sensitivity of eigenvalues and eigendecompositions of matrices*, Linear Algebra Appl., 396 (2005), pp. 273–301.

[BKM04] P. Benner, D. Kressner, and V. Mehrmann, *Skew-Hamiltonian and Hamiltonian eigenvalue problems: Theory, algorithms and applications*, in Proceedings of the Conference on Applied Mathematics and Scientific Computing, Brijuni, Croatia, 2003. Springer, Dordrecht, The Netherlands, http://www.math.tu-berlin.de/∼kressner/pub/bkm2final.pdf.

[BLO03] J. V. Burke, A. S. Lewis, and M. L. Overton, *Optimization and pseudospectra, with applications to robust stability*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 80–104.

[BLO04] 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.

[BLO05] 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.

[Bye88] R. Byers, *A bisection method for measuring the distance of a stable matrix to the unstable matrices*, SIAM J. Sci. Stat. Comput., 9 (1988), pp. 875–881.

[Dem83] J. W. Demmel, *A Numerical Analyst's Jordan Canonical Form*, Ph.D. thesis, University of California, Berkeley, CA, 1983.

[Dem86] J. W. Demmel, *Computing stable eigendecompositions of matrices*, Linear Algebra Appl., 79 (1986), pp. 163–193.

[Dem87] J. W. Demmel, *A counterexample for two conjectures about stability*, IEEE Trans. Automat. Control, 32 (1987), pp. 340–342.

[ET] M. Embree and L. N. Trefethen, Pseudospectra gateway, http://www.comlab.ox.ac.uk/pseudospectra.

[GMO+06] 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., 2006, to appear.

[Gu00] M. Gu, *New methods for estimating the distance to uncontrollability*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 989–1003.

[HJ91] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[MW01] V. Mehrmann and D. Watkins, *Structure-preserving methods for computing eigenpairs of large sparse skew-Hamiltonian/Hamiltonian pencils*, SIAM J. Sci. Comput., 22 (2000), pp. 1905–1925.

[Ste73] G. W. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.

[TE05] L. N. Trefethen and M. Embree, *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*, Princeton University Press, Princeton, NJ, 2005.

[Var67] J. M. Varah, *The Computation of Bounds for the Invariant Subspaces of a General Matrix Operator*, Ph.D. thesis, Stanford University, Palo Alto, CA, 1967.

[Var79] J. M. Varah, *On the separation of two matrices*, SIAM J. Numer. Anal., 16 (1979), pp. 216–222.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 2, pp. 360–385

A QUADRATICALLY CONVERGENT NEWTON METHOD FOR COMPUTING THE NEAREST CORRELATION MATRIX*

HOUDUO QI† AND DEFENG SUN‡

Abstract. The nearest correlation matrix problem is to find a correlation matrix which is closest to a given symmetric matrix in the Frobenius norm. The well-studied dual approach is to reformulate this problem as an unconstrained continuously differentiable convex optimization problem. Gradient methods and quasi-Newton methods such as BFGS have been used directly to obtain globally convergent methods. Since the objective function in the dual approach is not twice continuously differentiable, these methods converge at best linearly. In this paper, we investigate a Newton-type method for the nearest correlation matrix problem. Based on recent developments on strongly semismooth matrix valued functions, we prove the quadratic convergence of the proposed Newton method. Numerical experiments confirm the fast convergence and the high efficiency of the method.

Key words. correlation matrix, semismooth matrix equation, Newton method, quadratic convergence

AMS subject classifications. 49M45, 90C25, 90C33

DOI. 10.1137/050624509

1. Introduction.

Given a symmetric matrix $G \in \mathcal{S}^n$, computing its nearest correlation matrix, a problem from finance, is recently studied by Higham [25] and is given by

$$\begin{array}{ll} \min & \frac{1}{2}\| G - X \|^2 \\ \text{s.t.} & X_{ii} = 1, \quad i = 1, \ldots, n, \\ & X \in \mathcal{S}^n_+, \end{array} \tag{1}$$

where $\mathcal{S}^n$ and $\mathcal{S}^n_+$ are, respectively, the space of $n \times n$ symmetric matrices and the cone of positive semidefinite matrices in $\mathcal{S}^n$, and $\| \cdot \|$ is the Frobenius norm. It is noted that by introducing auxiliary variables, one may reformulate problem (1) as semidefinite programs or second-order cone programs, which may be solved by the well-developed modern interior point methods. However, when n is reasonably large, the direct use of interior point methods seems infeasible [25].[1] In tackling this difficulty, an alternating projection method of Dykstra [20] was proposed by Higham [25]. The projection method converges at best linearly. The latest study on problem (1) includes a dual approach proposed by Malick [35] and Boyd and Xiao [7]. This dual approach falls within the framework suggested by Rockafellar [43, p. 4] for general convex optimization problems.

Problem (1) is a special case of the following convex optimization problem:

*Received by the editors February 16, 2005; accepted for publication (in revised form) by N. J. Higham January 4, 2006; published electronically May 5, 2006.
http://www.siam.org/journals/simax/28-2/62450.html

†School of Mathematics, The University of Southampton, Highfield, Southampton SO17 1BJ, UK (hdqi@soton.ac.uk). This author's research was partially supported by EPSRC grant EP/D502535/1.

‡Department of Mathematics, National University of Singapore, Singapore 117543, Republic of Singapore (matsundf@nus.edu.sg). This author's research was partially supported by grant R146-000-061-112 of the National University of Singapore.

[1] By using preconditioned conjugate gradient methods to solve the linear system resulting from the interior point method, one may expect the interior point method to work well in practice [48].

$$\begin{array}{ll} \min & \frac{1}{2}\| x^0 - x \|^2 \\ \text{s.t.} & \mathcal{A}x = b \\ & x \in K\,, \end{array} \tag{2}$$

where $K \subseteq \mathcal{X}$ is a closed convex subset in a Hilbert space $\mathcal{X}$ endowed with an inner product $\langle \cdot, \cdot \rangle$ and its induced norm $\| \cdot \|$, $\mathcal{A} : \mathcal{X} \mapsto \mathbb{R}^n$ is a bounded linear operator, $b \in \mathbb{R}^n$ and $x^0 \in \mathcal{X}$ are given data (for problem (1), $\mathcal{X} = \mathcal{S}^n$, $K = \mathcal{S}^n_+$, $b = e$, the vector of all ones, $x^0 = G$, and $\mathcal{A}X = \mathrm{diag}[X]$, the vector formed by all diagonal elements of $X \in \mathcal{S}^n$). Problem (2) is also known as the best approximation from a closed convex set in a Hilbert space. See the recent book by Deutsch [13] and the references therein for details on this topic.

It has now become well known [14] that the (unique) solution x^* of (2) has the representation

$$x^* = \Pi_K(x^0 + \mathcal{A}^* y^*) \tag{3}$$

if and only if the set $\{K, \mathcal{A}^{-1}(b)\}$ has the so-called strong conical hull intersection property (CHIP), where $\Pi_K(\cdot)$ denotes the metric projection operator onto K under the inner product $\langle \cdot, \cdot \rangle$, y^* is a solution of the equation

$$\mathcal{A}\Pi_K(x^0 + \mathcal{A}^* y) = b, \tag{4}$$

and $\mathcal{A}^*$ denotes the adjoint of $\mathcal{A}$ (when $\mathcal{A} = \mathrm{diag}$, $\mathcal{A}^* y = \mathrm{Diag}[y]$, the diagonal matrix whose ith diagonal element is given by y_i). The property CHIP was initially characterized by Chui, Deutsch, and Ward [9] and was refined by Deutsch, Li, and Ward [14] to strong CHIP, which turns out to be a necessary and sufficient condition for the solution of (2) to have representation (3). In practice, however, strong CHIP is often difficult to verify for many interesting cases. Fortunately, there is an easy-to-verify sufficient condition:

$$b \in \mathrm{ri}\ (\mathcal{A}(K))\ . \tag{5}$$

$\mathcal{A}(K)$ is often called the data cone when K is a cone in $\mathcal{X}$ [9] and ri denotes the relative interior. We refer the reader to [2, 3, 5, 6, 10, 15, 36, 37] for related developments.

One well-studied concrete example of problem (2) is the convex best interpolation problem studied in [22, 27, 28, 36], where K is a closed convex cone given by

$$K := \{x \in L_2[0,1] \,|\, x \geq 0 \text{ a.e. on } [0,1]\}.$$

Newton's method for the dual of the convex best interpolation problem has been known to be the most efficient algorithm since [29, 1, 17]. The effectiveness of Newton's method was successfully explained very recently by Dontchev, Qi, and Qi [18, 19], where the authors established the superlinear (quadratic) convergence of Newton's method. The success of Newton's method for solving the convex best interpolation problem motivates us to study Newton's method for matrix nearness problem (1).

Coming to problem (1), we see $b = e$ and $\mathcal{A}(\mathcal{S}^n_+) = \mathbb{R}^n_+$, the nonnegative orthant of $\mathbb{R}^n$. Obviously, $e \in \mathrm{int}\mathbb{R}^n_+ = \mathrm{ri}\ \mathbb{R}^n_+$. Hence, (3) and (4) imply that there exists $y^* \in \mathbb{R}^n$ such that the unique solution X^* of (1) has the representation

$$X^* = (G + \mathcal{A}^* y^*)_+ \tag{6}$$

and y^* is a solution of the equation

$$\mathcal{A}\left(G + \mathcal{A}^* y\right)_+ = b, \quad y \in \mathbb{R}^n, \tag{7}$$

where X_+ denotes the metric projection of X onto $\mathcal{S}^n_+$, i.e., $X_+ := \Pi_{\mathcal{S}^n_+}(X)$. In fact, (7) is just the optimality condition of the following unconstrained and differentiable convex optimization problem [43]:

$$\min_{y \in \mathbb{R}^n} \theta(y) := \frac{1}{2} \| \left(G + \mathcal{A}^* y\right)_+ \|^2 - b^T y. \tag{8}$$

This is the dual problem of (1) studied in [35, 7]. The function $\theta(\cdot)$ is continuously differentiable, and its gradient mapping $\nabla\theta(\cdot)$ is globally Lipschitz continuous with the Lipschitz constant 1. Moreover, since Slater's condition is satisfied, $\theta(\cdot)$ is coercive; i.e., $\theta(y) \to +\infty$ as $\|y\| \to +\infty$ [43]. These nice properties allow one to apply either gradient-type methods or quasi-Newton methods to problem (8) directly [25, 35, 7]. However, since $\theta(\cdot)$ is not twice continuously differentiable, the convergence rate of these methods is at best linear. In this paper, we will show that Newton's method for solving problem (8) can achieve quadratic convergence by using the fact that the metric projection operator $\Pi_{\mathcal{S}^n_+}(\cdot)$ is strongly semismooth [46, 8]. We refer the interested reader to [47] for the strong semismoothness of the metric projection operator over the symmetric cones which include the nonnegative orthant, the second-order cone, and the positive semidefinite cone $\mathcal{S}^n_+$.

The paper is organized as follows. In section 2, we review some basic concepts and results concerning semismooth functions, especially in association with the projection X_+. In section 3, we develop Newton's method and show that it is quadratically convergent. As by-products of our analysis, we prove that the solution y^* is unique for any $G \in \mathcal{S}^n$ and $b > 0$ and is strongly semismooth as a function of G and b. This further implies that the solution X^* is also strongly semismooth as a function of G and b. Section 4 discusses some extensions which cover the W-weighted version of (1), a case with lower bounds, and a nonsymmetric case. We demonstrate that the developed Newton method applies to all those extensions under mild conditions. In section 5, we discuss the implementation issues and report our preliminary numerical results, which show that the Newton method is very efficient compared to existing methods. The conjugate gradient (CG) method is employed to solve the linear system obtained by Newton's method. We conclude our paper in section 6.

We use $\circ$ to denote the Hardmard product of matrices; i.e., for any $B, C \in \mathcal{S}^n$

$$B \circ C = [B_{ij} C_{ij}]^n_{i,j=1}.$$

We let E denote the matrix of all ones in $\mathcal{S}^n$. For subsets α, β of $\{1, 2, \ldots, n\}$, we denote $B_{\alpha\beta}$ as the submatrix of B indexed by α and β. Let e denote the vector of all ones.

2. Preliminaries. In this section, we review some basic concepts such as semismooth functions and generalized Jacobian of Lipschitz functions. These concepts will be used to define Newton's method for solving (7) and play an important role in our convergence analysis. We also review a perturbation result on eigenvalues of symmetric matrices.

Let $\Phi : \mathbb{R}^m \mapsto \mathbb{R}^\ell$ be a (locally) Lipschitz function. According to Redemacher's theorem (see [44, Sect. 9.J] for a proof), Φ is differentiable almost everywhere. We let

$$D_\Phi := \{x \in \mathbb{R}^m | \ \Phi \text{ is differentiable at } x\}.$$

Let $\Phi'(x)$ denote the Jacobian of Φ at $x \in D_\Phi$. The Bouligand subdifferential of Φ at $x \in \mathbb{R}^n$ is then defined by

$$\partial_B \Phi(x) := \left\{ V \in \mathbb{R}^{\ell \times m} \,|\, V \text{ is an accumulation point of } \Phi'(x^k),\ x^k \to x,\ x^k \in D_\Phi \right\}.$$

The generalized Jacobian in the sense of Clarke [11] is the convex hull of $\partial_B \Phi(x)$, i.e.,

$$\partial \Phi(x) = \operatorname{co} \partial_B \Phi(x).$$

Note that $\partial \Phi(x)$ is compact and upper-semicontinuous.

When $\ell = m$, a direct generalization of classical Newton's method for a system of smooth equations to $\Phi(x) = 0$ with a Lipschitz function Φ is given by [32, 42]

$$x^{k+1} = x^k - V_k^{-1} \Phi(x^k), \qquad V_k \in \partial \Phi(x^k), \quad k = 0, 1, 2, \ldots, \tag{9}$$

with x^0 as an initial guess. In general, the above iterative method does not converge. For a counterexample, see Kummer [32]. In extending Kojima and Shindo's condition for superlinear (quadratic) convergence of Newton's method for piecewise smooth equations [30], Kummer [32] proposed a general condition for guaranteeing the superlinear convergence of (9). However, Qi and Sun [42] popularized (9) by showing that the iterate sequence generated by (9) converges superlinearly if Φ belongs to an important subclass of Lipschitz functions—semismooth functions.

We say that Φ is semismooth at x if (i) Φ is directionally differentiable at x and (ii) for any $V \in \partial \Phi(x + h)$,

$$\Phi(x + h) - \Phi(x) - Vh = o(\|h\|).$$

Φ is said to be strongly semismooth at x if Φ is semismooth at x and for any $V \in \partial \Phi(x + h)$,

$$\Phi(x + h) - \Phi(x) - Vh = O(\|h\|^2).$$

The concept of semismoothness was introduced by Mifflin [38] for functionals. In order to study the convergence of (9), Qi and Sun [42] extended the definition of semismoothness to vector-valued functions and established the following convergence result.

THEOREM 2.1 (see [42, Thm. 3.2]). x^* $\Phi(x) = 0$ Φ x^* $V \in \partial \Phi(x^*)$ (9) x^* x^0 x^* Φ x^*

A similar result to the above theorem on the superlinear convergence of (9) can be found in [32, Prop. 3]. Theorem 2.1 gave the rates of convergence of (9) once the starting point x^0 is within the convergence region. The next theorem provides an estimate on how large the region of convergence can be.

THEOREM 2.2 (see [42, Thm. 3.3]). Φ $S := \{x \in \mathbb{R}^m \,|\, \|x - x^0\| \le r\}$ $V \in \partial \Phi(x)$ $x, y \in S$ V

$$\|V^{-1}\| \le \beta, \quad \|V(y - x) - \Phi'(x; y - x)\| \le \gamma \|y - x\|,$$

$$\|\Phi(y) - \Phi(x) - \Phi'(x; y - x)\| \le \delta \|y - x\|,$$

$\beta\|\Phi(x^0)\| \le r(1-\alpha)$ $\alpha := \beta(\gamma+\delta) < 1$ (9) S $\Phi(x) = 0$ S

$$\|x^k - x^*\| \le [\alpha/(1-\alpha)]\|x^k - x^{k-1}\|$$

$k = 1, 2, \ldots$.

Theorem 2.2 is an extension of the classical Newton–Kantorovich convergence theorem of Newton's method for solving smooth equations [40, Sect. 12.6]. Now we return our attention to problem (1). To facilitate our analysis, we define $F : \mathbb{R}^n \mapsto \mathbb{R}^n$ by

$$F(y) := \mathcal{A}(G + \mathcal{A}^* y)_+.$$

Then (7) becomes

$$F(y) = b \tag{10}$$

with $b = e$. It has been proved recently that $(\cdot)_+$ is strongly semismooth everywhere on $\mathcal{S}^n$ [46, 8]. Since the composite of strongly semismooth functions is still strongly semismooth, F is strongly semismooth everywhere on $\mathbb{R}^n$. So, in order to apply Theorem 2.1 to get a quadratically convergent Newton method, we need only to address the nonsingularity of $\partial F(y^*)$. It turns out to be the most difficult part in the analysis of Newton's method for solving (10). We will devote the whole next section to this issue.

We will also need the following perturbation result of Weyl for eigenvalues of symmetric matrices; see [4, p. 63] and [26, p. 367].

LEMMA 2.3. $\lambda_1 \ge \cdots \ge \lambda_n$ $X \in \mathcal{S}^n$ $\mu_1 \ge \cdots \ge \mu_n$ $Y \in \mathcal{S}^n$

$$|\lambda_i - \mu_i| \le \|X - Y\| \qquad \forall i = 1, \ldots, n.$$

3. Newton's method. In this section, we consider the nonsmooth Newton method for (10):

$$y^{k+1} = y^k - V_k^{-1}(F(y^k) - b), \qquad V_k \in \partial F(y^k), \quad k = 0, 1, 2, \ldots . \tag{11}$$

As we briefly discussed in section 2, the core issue for (11) is the nonsingularity of $\partial F(y)$ when y is near y^*, which is a solution of (10). Our main result in this section is that every element in $\partial F(y^*)$ is positive definite. Since F is already known to be strongly semismooth, Theorem 2.1 implies that method (11) is quadratically convergent if the initial point y^0 is sufficiently near y^*.

To facilitate our proofs for the positive definiteness of $\partial F(y^*)$ we need a few more notions. For any given $X \in \mathcal{S}^n$, let $\lambda(X)$ denote the eigenvalue vector of X arranged in the nonincreasing order, i.e., $\lambda_1(X) \ge \lambda_2(X) \ge \cdots \ge \lambda_n(X)$. Let $\mathcal{O}$ denote the set of all orthogonal matrices in $\mathbb{R}^{n\times n}$ and $\mathcal{O}_X$ be the set of orthonormal eigenvectors of X defined by

$$\mathcal{O}_X := \left\{ P \in \mathcal{O} |\ X = P\mathrm{Diag}[\lambda(X)]P^T \right\}.$$

Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function. Then one can define Löwner's function $f : \mathcal{S}^n \to \mathcal{S}^n$ (we adopt the convention of using f to denote both the scalar-valued and matrix-valued functions) by

$$f(X) := P\mathrm{Diag}[f(\lambda_1(X)), f(\lambda_2(X)), \ldots, f(\lambda_n(X))]P^T, \quad P \in \mathcal{O}_X. \tag{12}$$

The study on the matrix-valued function $f(X)$ defined in (12) was initiated by Löwner in his landmark paper [33]. See Donoghue [16] and Bhatia [4] for detailed discussions on (12).

For any $\mu = (\mu_1, \ldots, \mu_n)^T \in \mathbb{R}^n$ such that f is differentiable at $\mu_1, \ldots, \mu_n$, we denote by $f^{[1]}(\mu)$ the $n \times n$ symmetric matrix whose (i,j)th entry is

$$\left(f^{[1]}(\mu)\right)_{ij} = \begin{cases} \dfrac{f(\mu_i) - f(\mu_j)}{\mu_i - \mu_j} & \text{if } \mu_i \neq \mu_j, \\ f'(\mu_i) & \text{if } \mu_i = \mu_j. \end{cases}$$

The matrix $f^{[1]}(\mu)$ is called the first divided difference of f at μ. The following result of Löwner is well known. For a proof, see Donoghue [16, Chap. VIII] or [4, Chap. V.3.3].

LEMMA 3.1. [illegible] $P \in \mathcal{O}$ [illegible] $X = P\mathrm{Diag}[\lambda_1(X), \ldots, \lambda_n(X)]P^T$ [illegible] (a_1, a_2) [illegible] $\mathbb{R}$ [illegible] $\lambda_j(X)$, $j = 1, \ldots, n$ [illegible] f [illegible] [illegible] (a_1, a_2) [illegible] f [illegible] [illegible] X [illegible] [illegible] $H \in \mathcal{S}^n$ [illegible]

$$f'(X)H = P\left(f^{[1]}(\lambda(X)) \circ (P^T H P)\right) P^T. \tag{13}$$

Throughout the remainder of the paper, we let $f(t) = t_+ := \max(0, t)$, $t \in \mathbb{R}$. It is easy to derive from Moreau's theorem on the characterization of the metric projection operator over closed convex cones that (see [24, 50] for a proof)

$$X_+ = f(X) = P\mathrm{Diag}[\max\{\lambda_1(X), 0\}, \max\{\lambda_2(X), 0\}, \ldots, \max\{\lambda_n(X), 0\}]P^T.$$

By using Lemma 3.1 (by considering any continuously differentiable scalar-valued function with value one on an open set containing all the nonnegative eigenvalues of X and zero on an open set containing all negative eigenvalues of X) and the fact that $(\cdot)_+$ is (continuously) differentiable at $X \in \mathcal{S}^n$ if and only if X is nonsingular, we obtain the following useful result.

PROPOSITION 3.2. [illegible] $P \in \mathcal{O}$ [illegible] $X = P\mathrm{Diag}[\lambda_1(X), \ldots, \lambda_n(X)]P^T$ [illegible] $\Pi_{\mathcal{S}^n_+}(\cdot)$ [illegible] [illegible] $X \in \mathcal{S}^n$ [illegible] $\lambda_1(X), \ldots, \lambda_n(X)$ [illegible] $\lambda_i(X) \neq 0$ $i = 1, \ldots, n$ [illegible] $\lambda_i(X) \neq 0$ $i = 1, \ldots, n$ [illegible] $\Pi_{\mathcal{S}^n_+}(\cdot)$ X [illegible] $H \in \mathcal{S}^n$ [illegible] (13) [illegible] $f(t) = t_+$ $t \in \mathbb{R}$

See [8, Props. 4.3, 4.4] for a generalization on Proposition 3.2. We further let

$$C(y) := G + \mathcal{A}^* y \quad \text{and} \quad \lambda(y) := \lambda(C(y)).$$

We define three index sets associated with $\lambda(y)$:

$$\alpha(y) := \{i |\ \lambda_i(y) > 0\}, \quad \beta(y) := \{i |\ \lambda_i(y) = 0\}, \text{ and } \gamma(y) := \{i |\ \lambda_i(y) < 0\}.$$

We also let $\Lambda(y) := \mathrm{Diag}[\lambda(y)]$. When no confusion is involved, we often omit y for brevity. Let y^* be a solution of (4) throughout this section. For simplicity, we let

$$\lambda^* := \lambda(y^*),\ \alpha^* := \alpha(y^*),\ \gamma^* := \gamma(y^*), \text{ and } \Lambda^* := \Lambda(y^*).$$

Now we present our first technical result which is a direct consequence of the positiveness of b.

LEMMA 3.3. [illegible] $b > 0$, (10) [illegible] $\alpha^* \neq \emptyset$ [illegible] $P \in \mathcal{O}_{C(y^*)}$ [illegible]

$$\sum_{\ell \in \alpha^*} P_{i\ell}^2 > 0 \qquad \forall i = 1, \ldots, n.$$

[illegible]. Suppose that $P \in \mathcal{O}_{C(y^*)}$ is arbitrarily given. Then

$$(C(y^*))_+ = P \begin{pmatrix} \Lambda_\alpha^* & & \\ & 0 & \\ & & 0 \end{pmatrix} P^T$$

and (10) implies

$$\mathcal{A}P \begin{pmatrix} \Lambda_\alpha^* & & \\ & 0 & \\ & & 0 \end{pmatrix} P^T = b\,,$$

where Λ_α^* is a diagonal matrix of $|\alpha^*| \times |\alpha^*|$ with its diagonal elements given by $\lambda_i^*, i \in \alpha^*$. The fact that $b \neq 0$ implies that α^* is not empty. Equivalently, we have

$$\left(\sum_{\ell \in \alpha^*} \lambda_\ell^* P_{1\ell}^2, \sum_{\ell \in \alpha^*} \lambda_\ell^* P_{2\ell}^2, \ldots, \sum_{\ell \in \alpha^*} \lambda_\ell^* P_{n\ell}^2 \right) = (b_1, b_2, \ldots, b_n).$$

Since $\lambda_\ell^* > 0$ for all $\ell \in \alpha^*$, the lemma is proved to be true. □

Let

$$\delta^* := \frac{1}{2} \min_{i \in \alpha^* \cup \gamma^*} |\lambda_i^*|$$

and

$$\mathcal{B}(y^*, \delta^*) := \{ y \in \mathbb{R}^n \mid \|y - y^*\| \leq \delta^* \}.$$

Then the perturbation result in Lemma 2.3 implies that for all $y \in \mathcal{B}(y^*, \delta^*)$,

$$|\lambda_i(y) - \lambda_i^*| \leq \|C(y) - C(y^*)\| \leq \|y - y^*\| \leq \delta^* \qquad \forall\, i = 1, \ldots, n.$$

LEMMA 3.4. *F* [illegible] *y* [illegible] *f* [illegible] $C(y)$ [illegible]

$$F'(y)h = \mathcal{A}f'(C(y))H \qquad \forall\, h \in \mathbb{R}^n\,,$$

[illegible] $H := \mathcal{A}^* h = \mathrm{Diag}[h]$ [illegible]

$$f'(C(y))H = P\left(f^{[1]}(\lambda(y)) \circ (P^T H P) \right) P^T \qquad \forall\, P \in \mathcal{O}_{C(y)}.$$

[illegible] $y \in \mathcal{B}(y^*, \delta^*)$ [illegible]

$$\left(f^{[1]}(\lambda(y)) \right)_{ij} = 1 \qquad \forall\, i, j \in \alpha^*$$

[illegible]

$$\left(f^{[1]}(\lambda(y)) \right)_{ij} = 0 \qquad \forall\, i, j \in \gamma^*,$$

$$\left(f^{[1]}(\lambda(y))\right)_{\alpha^*\alpha^*} = E_{\alpha^*\alpha^*}, \qquad \left(f^{[1]}(\lambda(y))\right)_{\gamma^*\gamma^*} = 0_{\gamma^*\gamma^*}. \tag{14}$$

. It is obvious that if f is differentiable at $C(y)$, then F is differentiable at y because it is composed of f with linear transformations.

Suppose f is not differentiable at $C(y)$. Then Proposition 3.2 implies that f is not differentiable at $\lambda_i(y)$ for some $i \in \{1, \ldots, n\}$. The special structure of $f(t) = \max\{0, t\}$ yields that $\lambda_i(y) = 0$. Since $f(t)$ is directionally differentiable and nondecreasing, it holds that

$$f'(x; 1) \geq f'(x; -1) \qquad \forall\, x \in \mathbb{R}.$$

In particular,

$$f'(\lambda_i; 1) = 1 > 0 = f'(\lambda_i; -1).$$

We let $d, \hat{d} \in \mathbb{R}^n$ be defined, respectively, by

$$d_\ell = f'(\lambda_\ell; 1) \;\text{ and }\; \hat{d}_\ell = f'(\lambda_\ell; -1), \quad \ell = 1, \ldots, n.$$

Since $d_i = 1 > \hat{d}_i = 0$, we see that $d \neq \hat{d}$ and $d \geq \hat{d}$. Consider two sequences, respectively, specified by $\{y + te\}_{t>0}$ and $\{y - te\}_{t>0}$. We have

$$C(y + te) = P\mathrm{Diag}[\lambda + te]P^T \;\text{ and }\; C(y - te) = P\mathrm{Diag}[\lambda - te]P^T, \quad P \in \mathcal{O}_{C(y)}.$$

Hence,

$$\lim_{t\downarrow 0} \frac{F(y + te) - F(y)}{t} = \mathcal{A}P\mathrm{Diag}[d]P^T \;\text{ and }\; \lim_{t\downarrow 0} \frac{F(y - te) - F(y)}{-t} = \mathcal{A}P\mathrm{Diag}[\hat{d}]P^T.$$

With a bit of further calculation, we see by noticing $d_\ell \geq \hat{d}_\ell$ for $\ell = 1, \ldots, n$ and $d_i > \hat{d}_i$ that

$$\mathcal{A}P\mathrm{Diag}[d]P^T - \mathcal{A}P\mathrm{Diag}[\hat{d}]P^T = \begin{pmatrix} \sum_{\ell=1}^n (d_\ell - \hat{d}_\ell) P_{1\ell}^2 \\ \vdots \\ \sum_{\ell=1}^n (d_\ell - \hat{d}_\ell) P_{n\ell}^2 \end{pmatrix} \neq 0.$$

This means that

$$\lim_{t\downarrow 0} \frac{F(y + te) - F(y)}{t} \neq \lim_{t\downarrow 0} \frac{F(y - te) - F(y)}{-t},$$

implying that F is not differentiable at y. This establishes the first part of the lemma.

The formula for F' follows just from the chain rule and Proposition 3.2. The relation in (14) follows from the definition of $f^{[1]}$ and the fact that for any $y \in \mathcal{B}(y^*, \delta^*)$, $\lambda_i(y) > 0$ for all $i \in \alpha^*$ and $\lambda_i(y) < 0$ for all $i \in \gamma^*$. □

We now define a collection of matrices in relation to λ^*:

$$\mathcal{M} := \left\{ M \in \mathbb{R}^{n\times n} \mid\; M = \begin{pmatrix} E_{\alpha^*\alpha^*} & E_{\alpha^*\beta^*} & (\tau_{ij})_{\substack{i\in\alpha^* \\ j\in\gamma^*}} \\ E_{\beta^*\alpha^*} & (\omega_{ij})_{\substack{i\in\beta^* \\ j\in\beta^*}} & 0 \\ (\tau_{ji})_{\substack{i\in\alpha^* \\ j\in\gamma^*}} & 0 & 0 \end{pmatrix} \begin{array}{l} \omega_{ij} = \omega_{ji} \in [0, 1], \\ \text{for } i, j \in \beta^*, \\ \tau_{ij} = \lambda_i^*/(\lambda_i^* - \lambda_j^*), \\ \text{for } i \in \alpha^*, j \in \gamma^*. \end{array} \right\}$$

We note that $\mathcal{M}$ is a compact set and $1 > \tau_{ij} > 0$ for any $M \in \mathcal{M}$.

LEMMA 3.5. [illegible] $h \in \mathbb{R}^n$ [illegible]

$$\partial_B F(y^*)h \subseteq \{\mathcal{A}WH : \ W \in \mathcal{W}\},$$

[illegible] $H := \mathcal{A}^*h = \mathrm{Diag}[h]$ [illegible]

$$\mathcal{W} := \left\{W |\ WH = P\left(M \circ (P^T H P)\right)P^T,\ P \in \mathcal{O}_{C(y^*)},\ M \in \mathcal{M} \text{ [illegible] } h \in \mathbb{R}^n\right\}.$$

[illegible]. Let $V \in \partial_B F(y^*)$. By the very definition of $\partial_B F$ we have a sequence $\{y^k\}$ converging to y^* such that F is differentiable at each y^k and $F'(y^k) \to V$. Equivalently, we have

$$\lim_{k\to\infty} F'(y^k)h = Vh \qquad \forall h \in \mathbb{R}^n. \tag{15}$$

Then it follows from Lemma 3.4 that there exists $P^k \in \mathcal{O}_{C(y^k)}$ such that

$$F'(y^k)h = \mathcal{A}f'(C(y^k))H,$$

where $H = \mathcal{A}^*h = \mathrm{Diag}[h]$ and

$$f'(C(y^k))H = P^k\left(f^{[1]}(\lambda(y^k)) \circ ((P^k)^T H P^k)\right)(P^k)^T.$$

Denoting $\lambda^k := \lambda(y^k)$ for simplicity, when $y^k \in \mathcal{B}(y^*, \delta^*)$,

$$\lambda_i^k > 0 \text{ for } i \in \alpha^* \ \text{ and } \ \lambda_i^k < 0 \text{ for } i \in \gamma^*,$$

and λ_i^k for $i \in \beta^*$ could be positive or nonpositive but converges to $\lambda_i^* = 0$. Hence, the definition of $f^{[1]}$ yields

$$\left(f^{[1]}(\lambda^k)\right)_{ij} = \begin{cases} 1, & i,j \in \alpha^*, \\ 0, & i,j, \in \gamma^*, \\ \dfrac{\lambda_i^k - (\lambda_j^k)_+}{\lambda_i^k - \lambda_j^k}, & i \in \alpha^*, j \in \beta^*, \\ \dfrac{\lambda_i^k}{\lambda_i^k - \lambda_j^k}, & i \in \alpha^*, j \in \gamma^*, \end{cases}$$

and $(f^{[1]}(\lambda^k))_{ij} = (f^{[1]}(\lambda^k))_{ji}$ (i.e., it is symmetric). Because $0 \le (f^{[1]}(\lambda^k))_{ij} \le 1$ for all i, j, there exists a sequence (still denoted by $\{y^k\}$ without loss of generality) such that $f^{[1]}(\lambda^k)$ converges to a matrix, say M^*. It is easy to see that $M^* \in \mathcal{M}$. The boundedness of $\{P^k\}$ also implies that there exists a sequence (also denoted by $\{y^k\}$) such that $P^k \to P^*$. Then we have

$$C(y^*) = \lim_{k\to\infty} C(y^k) = \lim_{k\to\infty} P^k \mathrm{Diag}[\lambda^k](P^k)^T = P^* \mathrm{Diag}[\lambda](P^*)^T.$$

Hence, $P^* \in \mathcal{O}_{C(y^*)}$, and consequently we have by (15) that

$$Vh = \lim_{k\to\infty} F'(y^k)h \in \{\mathcal{A}WH : \ W \in \mathcal{W}\} \qquad \forall h \in \mathbb{R}^n.$$

Since $V \in \partial_B F(y^*)$ is arbitrary, we establish our result. □

Now we are ready to prove our main result in this section.

PROPOSITION 3.6. [illegible] $V \in \partial_B F(y^*)$ [illegible] $V \in \partial F(y^*)$ [illegible]. Let $V \in \partial_B F(y^*)$ be arbitrarily chosen. We want to show that for any $0 \neq h \in \mathbb{R}^n$

$$h^T V h > 0.$$

We note that it follows from Lemma 3.5 that there exist $M \in \mathcal{M}$ and $P \in \mathcal{O}_{C(y^*)}$ such that

$$Vh = \mathcal{A}\left(P(M \circ (P^T H P)\right) P^T.$$

Then

$$\begin{aligned}\langle h, Vh\rangle &= \langle \mathcal{A}^* h, P(M \circ (P^T H P))P^T\rangle \\ &= \langle P^T H P, M \circ (P^T H P)\rangle.\end{aligned}$$

Let $\tilde{H} := P^T H P$. Then we have

$$\begin{aligned}\langle h, Vh\rangle &= \langle \tilde{H}, M \circ \tilde{H}\rangle \\ &\geq \sum_{i\in\alpha^*}\left(\sum_{j\in\alpha^*\cup\beta^*} \tilde{H}_{ij}^2 + \sum_{j\in\gamma^*} \tau_{ij}\tilde{H}_{ij}^2\right) \\ &\geq \tau \sum_{i\in\alpha^*}\sum_{j=1}^{n} \tilde{H}_{ij}^2,\end{aligned}$$

where $\tau = \min_{i\in\alpha^*, j\in\gamma^*} \tau_{ij} > 0$. Because V is positive semidefinite, we see that $\langle h, Vh\rangle = 0$ only if

$$\tilde{H}_{ij} = 0 \qquad \forall i \in \alpha^* \text{ and } j \in \{1, \ldots, n\}.$$

The above condition is equivalent to

$$(\tilde{H}_{i1}, \tilde{H}_{i2}, \ldots, \tilde{H}_{in}) = (0, 0, \ldots, 0) \qquad \forall i \in \alpha^*.$$

By recalling that $\tilde{H} = P^T H P$ and $H = \mathrm{Diag}[h]$, we have

$$(\tilde{H}_{i1}, \tilde{H}_{i2}, \ldots, \tilde{H}_{in}) = (h_1 P_{1i}, h_2 P_{2i}, \ldots, h_n P_{ni})P = (0, 0, \ldots, 0)$$

if and only if

$$(h_1 P_{1i}^2, h_2 P_{2i}^2, \ldots, h_n P_{ni}^2) = (0, 0, \ldots, 0) \quad \text{(because } P \text{ is nonsingular)}.$$

Summarizing over $i \in \alpha^*$ in the above relation yields

$$\left(h_1 \sum_{i\in\alpha^*} P_{1i}^2, h_2 \sum_{i\in\alpha^*} P_{2i}^2, \ldots, h_n \sum_{i\in\alpha^*} P_{ni}^2\right) = (0, 0, \ldots, 0).$$

According to Lemma 3.3, the above condition holds if and only if

$$(h_1, h_2, \ldots, h_n) = (0, 0, \ldots, 0),$$

i.e., $h = 0$. This establishes the positive definiteness of V.

Since $\partial_B F(y^*)$ is compact and its every element is positive definite, any convex combination of its elements is also positive definite. That is, every element of $\partial F(y^*)$ is positive definite. □

The first of two important consequences of the above regularity result is on the convergence of Newton's method (11). It is just a direct application of Theorem 2.1, given that we have already known that F is strongly semismooth and every element in $\partial F(y^*)$ is positive definite.

COROLLARY 3.7. [illegible] (11) [illegible] y^0 [illegible] y^*

The second corollary is on the uniqueness of the solution to (10) and its strong semismoothness.

COROLLARY 3.8. [illegible] $G \in \mathcal{S}^n$ [illegible] $0 < b \in \mathbb{R}^n$ [illegible] y^* [illegible] (10) [illegible] y^* [illegible] [illegible] G [illegible] b [illegible] $y^*(G, b)$ [illegible] y^* [illegible] $(G, b) \in \mathcal{S}^n \times \mathbb{R}^n_{++}$ [illegible] X^* [illegible] G [illegible] b [illegible] $(G, b) \in \mathcal{S}^n \times \mathbb{R}^n_{++}$ [illegible]. The proof of Proposition 3.6 is independent of the choice of G and b as long as it belongs to $\mathcal{S}^n \times \mathbb{R}^n_{++}$. Hence, the Clarke inverse theorem says that there is a unique solution $y^*(G, b)$ for any $(G, b) \in \mathcal{S}^n \times \mathbb{R}^n_{++}$. We note that the existence of a solution is guaranteed because $0 < b \in \mathbb{R}^n_{++}$ and $b \in \mathrm{int}\mathcal{A}(\mathcal{S}^n_+)$. The strong semismoothness of y^* follows from a result of Sun [45] on an implicit theorem of strongly semismooth functions. Since X^* is composed of strongly semismooth functions, it is also strongly semismooth with respect to $(G, b) \in \mathcal{S}^n \times \mathbb{R}^n_{++}$. □

4. Extensions.

4.1. The W-weighted version. In practice, the W-weighted version of (1) is very useful [25]:

$$\begin{array}{ll} \min & \frac{1}{2}\|G - X\|_W^2 \\ \text{s.t.} & X_{ii} = 1, \quad i = 1, \ldots, n, \\ & X \succeq 0, \end{array} \tag{16}$$

where $W \in \mathcal{S}^n$ is positive definite and for any $Y \in \mathcal{S}^n$,

$$\|Y\|_W = \|W^{1/2} Y W^{1/2}\|.$$

Let

$$\bar{G} = W^{1/2} G W^{1/2} \quad \text{and} \quad \bar{X} = W^{1/2} X W^{1/2}.$$

Then problem (16) becomes standard in the form of (1):

$$\begin{array}{ll} \min & \frac{1}{2}\|\bar{G} - \bar{X}\|^2 \\ \text{s.t.} & (W^{-1/2} \bar{X} W^{-1/2})_{ii} = 1, \quad i = 1, \ldots, n, \\ & \bar{X} \succeq 0. \end{array}$$

In fact, the constraint $\bar{X} \succeq 0$ should be $W^{-1/2} \bar{X} W^{-1/2} \succeq 0$. It is easy to see that they are equivalent. For simplicity, we drop the bars in the above formulation and

have

$$\begin{array}{ll} \min & \frac{1}{2}\|G-X\|^2 \\ \text{s.t.} & \left(W^{-1/2}XW^{-1/2}\right)_{ii}=1, \quad i=1,\ldots,n, \\ & X \succeq 0. \end{array} \tag{17}$$

Define the linear operator $\mathcal{A}: \mathcal{S}^n \mapsto \mathbb{R}^n$ by

$$(\mathcal{A}X)_i = \left(W^{-1/2}XW^{-1/2}\right)_{ii}, \quad i=1,\ldots,n. \tag{18}$$

The adjoint operator $\mathcal{A}^*: \mathbb{R}^n \mapsto \mathcal{S}^n$ is given by

$$\begin{aligned} \langle \mathcal{A}^* y, X\rangle &= \langle y, \mathcal{A}X\rangle \\ &= \langle y, \mathrm{diag}[W^{-1/2}XW^{-1/2}]\rangle \\ &= \langle \mathrm{Diag}[y], W^{-1/2}XW^{-1/2}\rangle \\ &= \langle W^{-1/2}\mathrm{Diag}[y]W^{-1/2}, X\rangle. \end{aligned}$$

Hence

$$\mathcal{A}^* y = W^{-1/2}\mathrm{Diag}[y]W^{-1/2}. \tag{19}$$

It is easy to see that $e \in \mathrm{int}\mathcal{A}\mathcal{S}^n_+$. With this fact, we once again get (10) with $\mathcal{A}$ and $\mathcal{A}^*$ defined by (18) and (19), respectively. With no difficulty, we can develop parallel results as in Lemmas 3.3–3.5 and in Proposition 3.6. For example, Lemma 3.3 now becomes the following result.

LEMMA 4.1. [illegible] $b > 0$, (10) [illegible] $\mathcal{A}$ [illegible] $\mathcal{A}^*$ [illegible] (18) (19) [illegible] $\alpha^* \neq \emptyset$ [illegible] $P \in \mathcal{O}_{C(y^*)}$ [illegible]

$$\sum_{\ell \in \alpha^*} \hat{P}_{i\ell}^2 > 0 \qquad \forall i = 1,\ldots,n\,,$$

[illegible] $\hat{P} = W^{-1/2}P$

The proof follows just that of Lemma 3.3 and makes use of (18). Lemmas 3.4 and 3.5 remain true with $H = \mathcal{A}^* h = W^{-1/2}\mathrm{Diag}[h]W^{-1/2}$ for $h \in \mathbb{R}^n$. The proof for Proposition 3.6 is also true now with $\tilde{H} = P^T H P$ and H as just defined. Starting from

$$(\tilde{H}_{i1}, \tilde{H}_{i2}, \ldots, \tilde{H}_{in}) = (0, 0, \ldots, 0) \qquad \forall i \in \alpha^*$$

in the proof of Proposition 3.6, we have

$$\left(h_1 \sum_{i\in\alpha^*} \hat{P}_{1i}^2, h_2 \sum_{i\in\alpha^*} \hat{P}_{2i}^2, \ldots, h_n \sum_{i\in\alpha^*} \hat{P}_{ni}^2\right) = (0, 0, \ldots, 0)$$

by noticing

$$\tilde{H} = P^T W^{-1/2}\mathrm{Diag}[h]W^{-1/2}P = \hat{P}\mathrm{Diag}[h]\hat{P}.$$

According to Lemma 4.1, the above condition holds if and only if

$$(h_1, h_2, \ldots, h_n) = (0, 0, \ldots, 0).$$

This proves Proposition 3.6 with $\mathcal{A}$ and $\mathcal{A}^*$ defined by (18) and (19), respectively. Therefore, for the W-weighted version, Newton's method is quadratically convergent.

4.2. The case of lower bounds. The nearest correlation matrix is often rank-deficient [25]. To avoid the ill-conditionedness and to increase the stability, one often requires the matrix to be not less than a positive diagonal matrix. This gives the so-called calibration of correlation matrices, i.e.,

$$\begin{array}{ll} \min & \frac{1}{2}\|G - X\|^2 \\ \text{s.t.} & X \succeq \alpha I, \\ & \mathcal{A}X = e\,, \end{array} \tag{20}$$

where $\alpha \in (0,1)$ and $\mathcal{A}X = \mathrm{diag}[X]$. We will see that it is quite straightforward to apply the generalized Newton method to this case.

First, we note that the following condition is automatically valid:

$$\left\{\mathcal{A}^*y :\ (1-\alpha)y^Te \geq 0,\ y \in \mathbb{R}^n\right\} \cap (-\mathcal{S}^n_+) = \{0\}.$$

This condition corresponds to the condition [37, (2.17)], so that [37, Thm. 2.2] (this theorem considers only the case which corresponds to $G = 0$ in (20); however, it also holds for $G \neq 0$) implies that the unique solution of (20) has the following representation:

$$X^* = (G - \alpha I + \mathcal{A}^*y^*)_+ + \alpha I,$$

where y^* is a solution of the following equation:

$$\mathcal{A}(G - \alpha I + \mathcal{A}^*y)_+ + \alpha\mathcal{A}I = e,$$

which is obviously equivalent to

$$\mathcal{A}(G - \alpha I + \mathcal{A}^*y)_+ = (1-\alpha)e. \tag{21}$$

We now note that this equation actually defines a new problem similar to (1):

$$\begin{array}{ll} \min & \frac{1}{2}\|(G - \alpha I) - X\|^2 \\ \text{s.t.} & \mathcal{A}e = (1-\alpha)e, \\ & X \in \mathcal{S}^n_+. \end{array} \tag{22}$$

Hence, by following the discussion in section 1 and noting that $(1-\alpha)e > 0$, we know that the unique solution of problem (22) has the form

$$X^* = (G - \alpha I + \mathcal{A}^*y^*)_+,$$

where y^* is the unique solution of (21). We note that the uniqueness of y^* follows from Corollary 3.8 applied to (22). Therefore, Newton's method also applies to (21) and is quadratically convergent by Corollary 3.7, and hence solves (20).

A more complicated problem of the [illegible] was also discussed by Malick [35] and is defined by

$$\begin{array}{ll} \min & \frac{1}{2}\|X - \tilde{Q}\|^2 \\ \text{s.t.} & X \succeq \alpha I, \\ & \langle I, X\rangle = \mathrm{tr}(\tilde{Q}), \\ & \langle G_i, X\rangle = \sigma_i^2, \quad i = 1, \ldots, m, \end{array} \tag{23}$$

where $\alpha > 0$, $\tilde{Q}$ is a first estimate of the true covariance matrix Q used in portfolio risk analysis, and σ_i^2 represent "ex-post" volatilities of well-chosen portfolios; $G_i \in \mathcal{S}^n$. We now demonstrate how Newton's method can be applied to this problem.

The feasibility of problem (23) requires

$$\mathrm{tr}(\tilde{Q}) \geq n\alpha.$$

To facilitate our analysis, let

$$b_0 := \mathrm{tr}(\tilde{Q}),\ b_i := \sigma_i^2,\ i = 1, \ldots, m, \quad \text{and} \quad b := (b_0, b_1, \ldots, b_m)^T \in \mathbb{R}^{m+1},$$

$$G_0 := I,\ \mathcal{A} := (G_0, G_1, \ldots, G_m)$$

with

$$\mathcal{A}X := (\langle G_0, X\rangle, \langle G_1, X\rangle, \ldots, \langle G_m, X\rangle)^T \in \mathbb{R}^{m+1}.$$

Suppose that G_i's are positive semidefinite nonzero matrices. Then $\mathrm{tr}(G_i) > 0$ for each i. Let α be chosen such that

$$0 < \alpha < \min\{b_i/\mathrm{tr}(G_i) \mid i = 0, 1, \ldots, m\}. \tag{24}$$

We also assume that for any $y \in \mathbb{R}^{m+1}$ with $y_\ell > 0$ for some $\ell \in \{0, 1, \ldots, m\}$, we have

$$\mathcal{A}^* y := \sum_{i=1}^{m} G_i y_i \not\preceq 0. \tag{25}$$

Conditions (24) and (25) indicate how α and G_i are chosen in problem (23). Under these two conditions, we see that condition (2.17) in [37] is valid for problem (23), i.e.,

$$\{\mathcal{A}^* y :\ y^T(b - \alpha z^0) \geq 0\} \cap (-\mathcal{S}_+^n) = \{0\},$$

where $z^0 := \mathcal{A}I = (\mathrm{tr}(G_0), \mathrm{tr}(G_1), \ldots, \mathrm{tr}(G_m))^T \in \mathbb{R}^{m+1}$. Hence, once again [37, Thm. 2.2] implies that the unique solution of (23) has the representation

$$X^* = (\tilde{Q} - \alpha I + \mathcal{A}^* y^*)_+ + \alpha I\,,$$

where y^* is a solution to the following equation:

$$\mathcal{A}(\tilde{Q} - \alpha I + \mathcal{A}^* y)_+ = b - \alpha z^0. \tag{26}$$

Now the generalized Newton method can be applied to this equation. If we further assume that the matrices $G_i, i = 1, \ldots, m$, are mutually diagonalizable, Newton's method is also quadratically convergent following our results in the last section. To see this, let $P \in \mathcal{O}$ be a matrix such that G_i are simultaneously diagonalizable by P, i.e.,

$$G_i = P\Gamma^i P^T,\quad i = 1, \ldots, m,$$

where each Γ^i is a nonnegative diagonal matrix. Let $\Gamma^0 = I$ and define

$$\mathcal{L} := (\Gamma^0, \Gamma^1, \ldots, \Gamma^m)$$

so that

$$\mathcal{L}X = (\langle \Gamma^0, X\rangle, \langle \Gamma^1, X\rangle, \ldots, \langle \Gamma^m, X\rangle)^T$$

and

$$\mathcal{L}^* y = \sum_{i=0}^{m} \Gamma^i y_i.$$

Then (26) becomes

$$\mathcal{L}(P^T(\tilde{Q} - \alpha I)P + \mathcal{L}^* y)_+ = \tilde{b}\,, \tag{27}$$

where $\tilde{b} := \mathrm{diag}[P^T(b - \alpha z^0)P]$. Since $b - \alpha z^0 > 0$ by the assumed conditions, we see that $\tilde{b} > 0$. Now we note that (27) defines a new problem given by

$$\begin{array}{ll} \min & \dfrac{1}{2}\|P^T(\tilde{Q} - \alpha I)P - X\|^2 \\ \text{s.t.} & \langle \Gamma^i, X\rangle = \tilde{b}_i, \quad i = 0, \ldots, m, \\ & X \in \mathcal{S}^n_+. \end{array}$$

It is easy to repeat the arguments for problem (1) to verify that Newton's method for the above problem is quadratically convergent.

Finally, we note that all the assumptions made so far for problem (23) are automatically satisfied if each $G_i = E_i$, where E_i is the diagonal matrix whose only nonzero element is its ith diagonal element and equals 1.

4.3. The nonsymmetric case. In some applications [31], X may be only required to be positive semidefinite but not necessarily symmetric. Then we have the following matrix nearness problem:

$$\begin{array}{ll} \min & \dfrac{1}{2}\| X - G \|^2 \\ \text{s.t.} & \mathcal{A}X = b, \\ & X \in \mathcal{K}^n\,, \end{array} \tag{28}$$

where $\mathcal{K}^n$ is the cone of $n \times n$ positive semidefinite matrices (not necessarily symmetric)

$$\mathcal{K}^n = \{X \in \mathbb{R}^{n\times n} \,|\, X \text{ is positive semidefinite}\}\,.$$

By assuming the strong CHIP on $\{\mathcal{K}^n, \mathcal{A}^{-1}(b)\}$, we know from section 1 that the unique solution X^* to problem (28) has the representation

$$X^* = \Pi_{\mathcal{K}^n}(G + \mathcal{A}^* y^*) \tag{29}$$

and y^* is a solution of the equation

$$F(y) := \mathcal{A}\Pi_{\mathcal{K}^n}(G + \mathcal{A}^* y) = b, \quad y \in \mathbb{R}^n\,. \tag{30}$$

Next, we derive an explicit formula for computing $\Pi_{\mathcal{K}^n}(X)$ for a given $X \in \mathbb{R}^{n\times n}$. It is easy to see that $\Pi_{\mathcal{K}^n}(X)$ is the unique solution to

$$\begin{array}{ll} \min & \dfrac{1}{2}\| Y - X \|^2 \\ \text{s.t.} & \dfrac{1}{2}(Y + Y^T) \in \mathcal{S}^n_+\,. \end{array} \tag{31}$$

Since the Slater condition for problem (31) holds automatically, $\Pi_{\mathcal{K}^n}(X)$, together with the Lagrange multiplier $\Lambda \in \mathcal{S}^n_+$, satisfies the KKT conditions [34, Chap. 8]

$$\begin{cases} Y - X - \Lambda = 0, \\ \dfrac{1}{2}(Y + Y^T) \in \mathcal{S}^n_+, \quad \Lambda \in \mathcal{S}^n_+, \quad \dfrac{1}{2}(Y + Y^T)\Lambda = 0\,. \end{cases}$$

These conditions can be equivalently written as

$$\begin{cases} Y - X - \Lambda = 0, \\ \Lambda - \Pi_{\mathcal{S}^n_+}[\Lambda - \frac{1}{2}(Y + Y^T)] = 0\,, \end{cases}$$

which imply

$$\Lambda - \frac{1}{2}(Y + Y^T) = -\frac{1}{2}(X + X^T)$$

and

$$\Lambda = \frac{1}{2}\Pi_{\mathcal{S}^n_+}[-(X + X^T)]\,.$$

Hence

$$\Pi_{\mathcal{K}^n}(X) = X + \frac{1}{2}\Pi_{\mathcal{S}^n_+}[-(X + X^T)] = \frac{1}{2}(X - X^T) + \frac{1}{2}\Pi_{\mathcal{S}^n_+}(X + X^T)\,.$$

Therefore, by [46, Thm 4.13], we get the following result.

PROPOSITION 4.2. [illegible] $\Pi_{\mathcal{K}^n}(\cdot)$ [illegible] $X \in \mathbb{R}^{n\times n}$

Proposition 4.2 implies that the function F defined in (30) is strongly semismooth everywhere on $\mathbb{R}^n$. Then, in a similar way as for the symmetric case, we may use our Newton's method to find a solution of $F(y) = b$.

To establish the quadratic convergence of Newton's method, we restrict ourselves to the case that the linear operator $\mathcal{A} : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^n$ is defined by $\mathcal{A}X = \mathrm{diag}[X]$. In this case, the adjoint of $\mathcal{A}$ is $\mathcal{A}^* y = \mathrm{Diag}[y]$ (note that the inner product in $\mathbb{R}^{n\times n}$ is $\langle X, Y\rangle = \mathrm{tr}(X^T Y)$.) Noticing that

$$\mathcal{A}(X - X^T) = 0,$$

we see that the nonsmooth equation (30) becomes

$$F(y) = \frac{1}{2}\mathcal{A}\Pi_{\mathcal{S}^n_+}(C(y) + C^T(y)) = b,$$

where as before we denote $C(y) = G + \mathcal{A}^* y$. In a more explicit form we have

$$F(y) = \mathcal{A}\Pi_{\mathcal{S}^n_+}\left(\frac{1}{2}(G + G^T) + \mathcal{A}^* y\right) = b. \tag{32}$$

This is the nonsmooth equation derived from the following standard problem in the form of (1):

$$\begin{array}{ll} \min & \dfrac{1}{2}\|(G + G^T)/2 - X\|^2 \\ \text{s.t.} & X_{ii} = b_i, \quad i = 1, \ldots, n, \\ & X \in \mathcal{S}^n_+\,. \end{array} \tag{33}$$

Under the condition that $b > 0$, we see from our previous results for the symmetric case similar to (33) that Proposition 3.6 holds for (32). Hence, Newton's method is quadratically convergent for the special case.

5. Numerical results. In numerical experiments, we used the following globalized version of Newton's method for solving the dual problem (8). Recall that for any $y \in \mathbb{R}^n$, $\nabla\theta(y) = F(y) - b$ and $b = e$.

ALGORITHM 5.1. NEWTON'S METHOD.

Step 0. $y^0 \in \mathbb{R}^n$ $\eta \in (0,1)$ $\rho, \sigma \in (0,1/2)$ $k := 0$

Step 1. $V_k \in \partial F(y^k)$ [23] d^k

$$\nabla\theta(y^k) + V_k d = 0 \tag{34}$$

$$\|\nabla\theta(y^k) + V_k d^k\| \le \eta_k \|\nabla\theta(y^k)\|, \tag{35}$$

$\eta_k := \min\{\eta, \|\nabla\theta(y^k)\|\}$ (35)

$$\nabla\theta(y^k)^T d^k \le -\eta_k \|d^k\|^2 \tag{36}$$

$d^k := -B_k^{-1}\nabla\theta(y^k)$ B_k $\mathcal{S}^n$

Step 2. m_k m

$$\theta(y^k + \rho^m d^k) - \theta(y^k) \le \sigma\rho^m \nabla\theta(y^k)^T d^k.$$

$t_k := \rho^{m_k}$ $y^{k+1} := y^k + t_k d^k$

Step 3. k $k+1$ 1

An alternative to calculating the Newton direction is to apply the CG method to the following perturbed Newton equation:

$$\nabla\theta(y^k) + (V_k + \varepsilon_k I)\, d = 0 \quad \text{with} \quad \varepsilon_k > 0.$$

The classical choice of ε_k is the norm of the residue, i.e., $\varepsilon_k = \|F(y^k) - b\|$. Since V_k is always positive semidefinite, the matrix $(V_k + \varepsilon_k I)$ is always positive definite for any $\varepsilon^k > 0$.

We provide a proof for the sake of completeness.

The global convergence analysis of Algorithm 5.1 is quite standard. Since the CG method is used to calculate the Newton direction, it is actually an inexact Newton direction that was used in our implementation. Hence, our local convergence analysis is a bit different from the standard ones. We provide a proof for the sake of completeness.

First, we need the following result due to Facchinei [21, Thm. 3.3 and Remark 3.4]. A similar result was also obtained by Pang and Qi [41].

LEMMA 5.2. k

$$\nabla\theta(y^k)^T d^k \le -\hat{\rho}\|d^k\|^2$$

$\hat{\rho} > 0$ $\mu \in (0,1/2)$ $\bar{k}$ $k \ge \bar{k}$

$$\theta(y^k + d^k) \le \theta(y^k) + \mu\nabla\theta(y^k)^T d^k.$$

THEOREM 5.3. 5.1 $\{\|B_k\|\}$ $\{\|B_k^{-1}\|\}$ $\{y^k\}$ 5.1 y^* $F(y) = b$

. Since for any $k \geq 0$, d^k is always a descent direction of $\theta(\cdot)$ at y^k, Algorithm 5.1 is well defined. Moreover, from the coercive property of θ we know that $\{y^k\}$ is bounded. Then, by employing standard convergence analysis (cf. [12, Thm 6.3.3]), we can conclude that

$$\lim_{k\to\infty} \nabla\theta(y^k) = 0\,,$$

which, together with the convexity of $\theta(\cdot)$ and the boundedness of $\{y^k\}$, implies that $y^k \to y^*$.

Since, by Proposition 3.6, any element $V \in \partial F(y^*)$ is positive definite, it holds that for all k sufficiently large, V_k is positive definite and $\{\|V_k^{-1}\|\}$ is uniformly bounded. Hence, for all k sufficiently large, the CG method can find d^k such that both (35) and (36) are satisfied. This, together with the facts that $\nabla\theta(y^*) = 0$ and $\nabla\theta(\cdot)$ is strongly semismooth at y^*, further implies that for all k sufficiently large,

(37)
$$\begin{aligned}
\|y^k + d^k - y^*\| &= \|y^k + V_k^{-1}[(\nabla\theta(y^k) + V_k d^k) - \nabla\theta(y^k)] - y^*\| \\
&\leq \|y^k - y^* - V_k^{-1}\nabla\theta(y^k)\| + \|V_k^{-1}(\nabla\theta(y^k) + V_k d^k)\| \\
&\leq \|V_k^{-1}\|\|\nabla\theta(y^k) - \nabla\theta(y^*) - V_k(y^k - y^*)\| + \eta_k\|V_k^{-1}\|\|\nabla\theta(y^k)\| \\
&\leq O(\|y^k - y^*\|^2) + \|V_k^{-1}\|\|\nabla\theta(y^k)\|^2 \\
&\leq O(\|y^k - y^*\|^2) + O(\|\nabla\theta(y^k) - \nabla\theta(y^*)\|)^2, \\
&= O(\|y^k - y^*\|^2),
\end{aligned}$$

where in the last equality we used the Lipschitz continuity of $\nabla\theta(\cdot)$. From (37) and the fact that $y^k \to y^*$, we have for all k sufficiently large that

$$y^k - y^* = -d^k + O(\|d^k\|^2) \qquad \text{and} \qquad \|d^k\| \to 0. \tag{38}$$

For each $k \geq 0$, let $r^k := \nabla\theta(y^k) + V_k d^k$. Then for all k sufficiently large,

$$\begin{aligned}
-\nabla\theta(y^k)^T d^k &= \langle d^k, V_k d^k\rangle - \langle d^k, r^k\rangle \\
&\geq \langle d^k, V_k d^k\rangle - \|d^k\|\|r^k\| \\
&\geq \langle d^k, V_k d^k\rangle - \eta_k\|d^k\|\|\nabla\theta(y^k)\| \\
&= \langle d^k, V_k d^k\rangle - \|d^k\|\|\nabla\theta(y^k)\|^2 \\
&\geq \langle d^k, V_k d^k\rangle - \|d^k\|\|y^k - y^*\|^2, \qquad (39)
\end{aligned}$$

which, together with (38) and the uniform positive definiteness of V_k, implies that there exists $\hat{\rho} > 0$ such that for all k sufficiently large,

$$-\nabla\theta(y^k)^T d^k \geq \hat{\rho}\|d^k\|^2.$$

It then follows from Lemma 5.2 that for all k sufficiently large, $t_k = 1$ and

$$y^{k+1} = y^k + d^k\,.$$

The proof is completed by observing (37). □

Next, we discuss several issues regarding the implementation of Algorithm 5.1.

(a) [illegible] In Algorithm 5.1, we need to find a $V \in \partial F(y)$ to form (34). For a given $y \in \mathbb{R}^n$, let $C(y)$ have the following spectral decomposition:

$$C(y) = P\mathrm{Diag}[\lambda(y)]P^T, \quad P \in \mathcal{O}_{C(y)}.$$

Let

$$M_y := \begin{pmatrix} E_{\alpha\alpha} & E_{\alpha\beta} & (\tau_{ij}(y))_{\substack{i\in\alpha\\ j\in\gamma}} \\ E_{\beta\alpha} & 0 & 0 \\ (\tau_{ji}(y))_{\substack{i\in\alpha\\ j\in\gamma}} & 0 & 0 \end{pmatrix}, \qquad \tau_{ij}(y) := \frac{\lambda_i(y)}{\lambda_i(y) - \lambda_j(y)}, \quad i \in \alpha, j \in \gamma.$$

Define the matrix $V_y \in \mathbb{R}^{n\times n}$ by

$$V_y h = \mathcal{A}P\left(M_y \circ (P^T H P)\right) P^T, \quad h \in \mathbb{R}^n, \tag{40}$$

where $H := \mathrm{Diag}[h]$.

PROPOSITION 5.4. [illegible] V_y [illegible] (40) [illegible]

$$V_y \in \partial_B F(y) \subseteq \partial F(y).$$

[illegible]. Recall that the scalar-valued function $f(t) = \max(0, t), t \in \mathbb{R}$. For each $k > 0$, let $t_k := -1/k$. We now consider the sequence $\{z^k\}$ with z^k given by $z^k := y - t_k e = y + (1/k)e$. Then $\lambda(y) - t_k e$ is the spectrum of $C(z^k)$, i.e.,

$$C(z^k) = C(y) - t_k C(e) = P\mathrm{Diag}[\lambda(y) - t_k e]P^T = P\mathrm{Diag}[\lambda(y) + (1/k)e]P^T.$$

Let $\bar{k} > 0$ be sufficiently large such that $1/\bar{k} < \min\{|\lambda_i(y)| \,|\, i \in \alpha \cup \gamma\}$ (recall the definitions of α and γ). Then, for each $k \geq \bar{k}$, the matrix-valued function $f : \mathcal{S}^n \to \mathcal{S}^n$ is differentiable at $C(z^k)$ because $C(z^k)$ is nonsingular and in this case, by Lemma 3.1, f is differentiable at $C(z^k)$ and for any $Z \in \mathcal{S}^n$,

$$f'(C(z^k))Z = P\left(f^{[1]}(\lambda(y) + (1/k)e) \circ (P^T Z P)\right) P^T.$$

Therefore, from Lemma 3.4 we know that for each $k \geq \bar{k}$, F is differentiable at z^k and for any $h \in \mathbb{R}^n$,

$$F'(z^k)h = \mathcal{A}f'(C(z^k))H,$$

where $H := \mathrm{Diag}[h]$. After direct computations we can see that

$$M_y = \lim_{k\to\infty} f^{[1]}(\lambda(y) + (1/k)e).$$

Hence, for each $h \in \mathbb{R}^n$,

$$\lim_{k\to\infty} F'(z^k)h = \mathcal{A}P\left(M_y \circ (P^T H P)\right) P^T,$$

which, together with (40), implies that

$$V_y = \lim_{k\to\infty} F'(z^k).$$

Thus, by the definition of $\partial_B F(y)$, $V_y \in \partial_B F(y)$. The proof is completed by observing that $\partial F(y) = \mathrm{co}\,\partial_B F(y)$. □

We see from Proposition 5.4 that we can obtain an element $V_y \in \partial F(y)$ by the spectral decomposition of $C(y)$. Since we use the CG method to solve (34), we do not need to form V_y explicitly.

(b) We tested the following four classes of problems.

EXAMPLE 5.5. C is a randomly generated $n \times n$ correlation matrix by `gallery ('randcorr',n)` of MATLAB 7.0.1. R is a random $n \times n$ symmetric matrix with $R_{ij} \in [-1, 1]$, $i, j = 1, 2, \ldots, n$. Then we set

$$G = C + \alpha R\,,$$

where $\alpha = 0.01, 0.1, 1.0, 10.0$. We fix $n = 1000$ in our numerical reports. This problem was tested by Higham [25].

EXAMPLE 5.6. G is a randomly generated symmetric matrix as in the first example of Malick [35] with $G_{ij} \in [-1, 1]$ and $G_{ii} = 1.0$, $i, j = 1, 2, \ldots, n$, and $n = 500, 1000, 1500, 2000$.

EXAMPLE 5.7. G is a randomly generated symmetric matrix with $G_{ij} \in [0, 2]$ and $G_{ii} = 1.0$, $i, j = 1, 2, \ldots, n$, and $n = 500, 1000, 1500, 2000$.

EXAMPLE 5.8. G is a randomly generated symmetric matrix as in the second example of Malick [35] with

$$G_{ii} \in [-2.0 \times 10^4, 2.0 \times 10^4],\ i = 1, 2, \ldots, n.$$

We add to G a perturbed $n \times n$ random symmetric matrix with entries in $[-\alpha, \alpha]$, where $\alpha = 0.0, 0.01, 0.1, 1.0$. We report our numerical results for $n = 1000$.

(c) In our numerical experiments, two initial points were used: (i) $b - \mathrm{diag}(G)$; and (ii) $b - \mathrm{diag}(G) + e$. Other initial points may be used. For example, we may start from a positive point, i.e., $y^0 > 0$, such that $C(y^0)$ is positive definite. The performance of Newton's method is similar, as we reported below. We set other parameters as $\eta = 10^{-5}$, $\rho = 0.5$, and $\sigma = 2.0 \times 10^{-4}$. For simplicity, we fix $B_k \equiv I$ for all $k \geq 0$.

(d) For the purpose of comparison, we tested the performance of the BFGS method with the Wolfe line search used by Malick [35] and the alternating projection method employed by Higham [25]. The details of the implementation of the BFGS method can be found in [39, Chap. 8]. As observed by Malick [35, Thm. 5.1], Higham's method is the following standard gradient optimization algorithm applied to (8):

$$y^{k+1} := y^k - \nabla\theta(y^k)\,, \quad k = 0, 1, \ldots,$$

and is therefore called the gradient method. We also tested a hybrid method that combines the BFGS method and Newton's method. The hybrid method, which is called BFGS-N here, starts with the BFGS method and switches to Newton's method when $\|\nabla\theta(y^k)\| \leq 1.0$.

All tests were carried out in MATLAB 7.0.1 running on a PC Pentium IV. In our experiments, our stopping criterion is

$$\|\nabla\theta(y^k)\| \leq 10^{-5}\,.$$

The reason that we chose 10^{-5} instead of 10^{-6} or higher accuracy is because the BFGS method and the gradient method ran into difficulty for a higher accuracy in a few cases. Our numerical results are reported in Tables 1–4, where Init., Iter., Func.,

TABLE 1
Numerical results of Example 5.5.

Init.	Algorithm	α	cputime	Iter.	Func.	Res.
(i)	Newton	0.01	2 m 13 s	1	2	2.6×10^{-7}
		0.1	2 m 58 s	3	4	2.0×10^{-8}
		1.0	3 m 38 s	5	6	2.7×10^{-8}
		10.0	4 m 13 s	7	8	9.9×10^{-8}
	BFGS	0.01	2 m 19 s	2	3	2.3×10^{-7}
		0.1	3 m 03 s	5	6	8.0×10^{-7}
		1.0	6 m 27 s	18	19	9.7×10^{-6}
		10.0	15 m 10 s	53	54	6.4×10^{-6}
	BFGS-N	0.01	2 m 16 s	1	2	7.2×10^{-8}
		0.1	3 m 10 s	4	5	4.9×10^{-11}
		1.0	3 m 50 s	7	8	4.0×10^{-6}
		10.0	6 m 00 s	15	16	2.6×10^{-10}
	Gradient	0.01	2 m 20s	2	3	6.0×10^{-6}
		0.1	4 m 56 s	13	14	6.6×10^{-6}
		1.0	24 m 38 s	107	108	9.3×10^{-6}
		10.0	1 h 57 m 54 s	500	501	8.2×10^{-3}
(ii)	Newton	0.01	0.22 s	2	3	1.4×10^{-6}
		0.1	3 m 12 s	4	5	1.1×10^{-10}
		1.0	3 m 41 s	5	6	4.5×10^{-7}
		10.0	4 m 39 s	7	8	1.2×10^{-7}
	BFGS	0.01	2 m 50 s	3	4	6.9×10^{-8}
		0.1	3 m 25 s	6	7	6.9×10^{-6}
		1.0	8 m 09 s	19	20	6.3×10^{-6}
		10.0	15 m 11 s	53	54	7.9×10^{-6}
	BFGS-N	0.01	2 m 39 s	2	3	4.6×10^{-6}
		0.1	3 m 08 s	4	5	6.3×10^{-7}
		1.0	4 m 16 s	7	8	4.0×10^{-6}
		10.0	6 m 37 s	15	16	2.3×10^{-9}
	Gradient	0.01	02 m 48s	3	4	5.1×10^{-6}
		0.1	5 m 24 s	14	15	6.0×10^{-6}
		1.0	24m 06 s	106	107	9.2×10^{-6}
		10.0	1 h 59 m 53 s	500	501	8.4×10^{-3}

and Res. stand for, respectively, the initial point used, the number of iterations, the number of function evaluations of θ, and the residual $\|\nabla\theta(y^k)\|$ at the final iterate of an algorithm (we set a maximum of 500 iterations). LS failed means that the line search failed (the steplength is too small to proceed) during the computation.

An outstanding observation is that Newton's method took less than 10 iterations for all the problems to reach the reported accuracy and the quadratic convergence was observed. The BFGS method performed quite well for Examples 5.5, 5.6, and 5.8, while there are four line search failures in Example 5.7. Sometimes it took much longer to reach the required accuracy. Numerical results for BFGS-N clearly showed that Newton's method can be used to save a lot of computing time required by the BFGS method. The gradient method is generally outperformed by the BFGS method. Compared to the numerical results reported in [49] on the inexact primal-dual path following interior point methods for the similar tested examples, our proposed Newton method is much faster (4 to 5 times) in terms of the cputime. The main reason is that the proposed Newton method needs fewer iterations and at each iteration it needs only one eigenvalue decomposition instead of two as in the inexact primal-dual path following interior point methods [49].

More specific observations are included in the following remarks.

TABLE 2
Numerical results of Example 5.6.

Init.	Algorithm	n	cputime	Iter.	Func.	Res.
(i)	Newton	500	16.6 s	5	6	1.0×10^{-9}
		1,000	1 m 49 s	5	6	3.3×10^{-8}
		1,500	5 m 44 s	5	6	2.7×10^{-7}
		2,000	12 m 34 s	5	6	1.5×10^{-6}
	BFGS	500	32.1 s	16	17	5.5×10^{-6}
		1,000	4 m 03 s	19	20	5.7×10^{-6}
		1,500	13 m 26 s	20	21	9.1×10^{-6}
		2,000	33 m 10 s	22	23	3.9×10^{-6}
	BFGS-N	500	15.1 s	6	7	4.0×10^{-6}
		1,000	2 m 00 s	7	8	3.6×10^{-6}
		1,500	7 m 44 s	7	8	7.4×10^{-6}
		2,000	17 m 06 s	8	9	1.9×10^{-11}
	Gradient	500	2 m 43 s	76	77	9.2×10^{-6}
		1,000	25 m 26 s	106	107	9.0×10^{-6}
		1,500	1 h 24 m 44 s	126	127	9.5×10^{-6}
		2,000	3 h 41 m 16 s	144	145	1.0×10^{-5}
(ii)	Newton	500	16.4 s	5	6	4.3×10^{-9}
		1,000	1 m 50 s	5	6	9.4×10^{-8}
		1,500	6 m 10 s	5	6	7.0×10^{-7}
		2,000	13 m 38 s	5	6	2.2×10^{-6}
	BFGS	500	32.2 s	17	18	8.1×10^{-6}
		1,000	4 m 14 s	19	20	7.0×10^{-6}
		1,500	15 m 23 s	21	22	4.9×10^{-6}
		2,000	35 m 04 s	22	23	3.9×10^{-6}
	BFGS-N	500	14.8 s	6	7	6.7×10^{-6}
		1,000	2 m 02 s	7	8	3.3×10^{-6}
		1,500	5 m 57 s	7	8	9.5×10^{-6}
		2,000	18 m 35 s	8	9	8.3×10^{-11}
	Gradient	500	2 m 25 s	78	79	9.4×10^{-6}
		1,000	21 m 31 s	105	106	9.0×10^{-6}
		1,500	1 h 46 m 40 s	127	128	9.7×10^{-6}
		2,000	3 h 34 m 59 s	144	145	9.4×10^{-6}

REMARK 5.9. Newton's method takes less cputime and fewer iterations. For all the tested examples, it is observed that the Newton method always took the unit steplength and achieved the quadratic convergence at the last several iterations. Typically, Newton's method was terminated in two or three steps after the residue of the gradient was below 10^{-1} or 10^{-2}.

REMARK 5.10. The major cost in Newton's method includes two parts: (1) the spectral decomposition and (2) the CG method for solving the linear system. In order to form the linear system, we need the computation of the full eigensystem. So it seems that the computing time involved in part (1) is inevitable. The computing time in part (2) may be reduced by making use of the special structure of $\partial_B F(y)$, $y \in \mathbb{R}^n$. We did not explore the latter in our implementation, as we are quite satisfied with the performance of Newton's method.

REMARK 5.11. The major cost in the BFGS method and the gradient is the spectral decomposition. By doing a partial spectral decomposition as outlined in [25], we may be able to save some cputime. We did not exploit this, as we do not know the distributions of the eigenvalues of the optimal correlation matrix.

REMARK 5.12. It can be seen clearly from the numerical results for BFGS-N that Newton's steps reduced the cputime committed by the BFGS method substantially.

TABLE 3
Numerical results of Example 5.7.

Init.	Algorithm	n	cputime	Iter.	Func.	Res.
(i)	Newton	500	34.3 s	8	9	3.7×10^{-9}
		1,000	4 m 55 s	9	10	3.1×10^{-9}
		1,500	14 m 04 s	9	10	4.5×10^{-7}
		2,000	33 m 52 s	9	10	2.6×10^{-6}
	BFGS	500	2 m 46 s	88	89	9.4×10^{-6}
		1,000	LS failed	110	119	2.3×10^{-5}
		1,500	LS failed	111	123	4.7×10^{-5}
		2,000	LS failed	112	129	8.1×10^{-5}
	BFGS-N	500	43.1 s	12	13	1.4×10^{-7}
		1,000	6 m 09 s	15	17	9.8×10^{-10}
		1,500	19 m 03 s	15	17	3.6×10^{-10}
		2,000	1 h 08 m 36 s	20	28	1.1×10^{-7}
	Gradient	500	15 m 53 s	500	501	3.7×10^{-2}
		1,000	2 h 01 m 01 s	500	501	1.3×10^{-1}
		1,500	5 h 25 m 42 s	500	501	2.0×10^{-1}
		2,000	–	–	–	–
(ii)	Newton	500	35.6 s	8	9	1.7×10^{-7}
		1,000	4 m 34 s	9	10	6.1×10^{-8}
		1,500	15 m 37 s	9	10	6.2×10^{-7}
		2,000	40 m 06 s	9	10	3.8×10^{-6}
	BFGS	500	2 m 51 s	89	90	9.3×10^{-6}
		1,000	26 m 01 s	116	118	9.6×10^{-6}
		1,500	LS failed	122	126	2.6×10^{-5}
		2,000	3 h 43 m 33 s	139	140	1.0×10^{-5}
	BFGS-N	500	45.2 s	12	15	2.4×10^{-6}
		1,000	6 m 16 s	15	17	2.6×10^{-9}
		1,500	18 m 55 s	15	17	8.3×10^{-8}
		2,000	50 m 56 s	14	18	7.0×10^{-7}
	Gradient	500	15 m 13 s	500	501	3.7×10^{-2}
		1,000	1 h 54 m 18 s	500	501	1.2×10^{-1}
		1,500	5 h 22 m 08 s	500	501	1.9×10^{-1}
		2,000	–	–	–	–

If one can calculate $F(y)$ much less costly than via the computation of the full eigensystem, then it may be a good choice to start with a method such as the BFGS, which costs less than Newton's method at each step, and then switch to Newton's method when the iterates are close to the solution. In this case, BFGS-N may be an ideal choice.

6. Conclusion. In this paper, a close look at the nearest correlation matrix problem as the best approximation from a convex set in a Hilbert space led us to consider Newton's method. Theoretically, we proved that Newton's method is well defined and is quadratically convergent. Our theoretical results were then extended to such problems as the W-weighted nearest correlation problem, the case with lower bounds, and the nonsymmetric case. Numerically, Newton's method is shown to be extremely efficient, taking less than 10 iterations to solve all the test problems. This research opens the possibility of developing Newton's method for other least-square semidefinite problems. We shall pursue this possibility in our future research.

Acknowledgments. The authors are grateful to Professor N. J. Higham for suggesting the present title and to the referees for their helpful comments.

TABLE 4
Numerical results of Example 5.8.

Init.	Algorithm	α	cputime	Iter.	Func.	Res.
(i)	Newton	0.0	9.4 s	1	2	2.3×10^{-13}
		0.01	1 m 52 s	5	6	1.4×10^{-6}
		0.1	2 m 33 s	6	7	3.9×10^{-7}
		1.0	4 m 19 s	8	9	1.6×10^{-8}
	BFGS	0.0	28.0 s	2	9	4.6×10^{-13}
		0.01	5 m 00 s	23	27	1.4×10^{-6}
		0.1	5 m 23 s	27	29	8.9×10^{-6}
		1.0	9 m 24 s	50	52	9.1×10^{-6}
	BFGS-N	0.0	8.7 s	1	2	1.6×10^{-13}
		0.01	2 m 03 s	5	6	1.4×10^{-6}
		0.1	2 m 25 s	11	12	4.1×10^{-9}
		1.0	6 m 11 s	20	25	2.0×10^{-9}
	Gradient	0.0	27 m 29 s	500	501	1.6×10^{-2}
		0.01	1 h 36 m 35 s	500	501	5.6×10^{-2}
		0.1	1 h 26 m 35 s	500	501	4.0×10^{-1}
		1.0	1 h 51 m 23 s	500	501	4.0×10^{0}
(ii)	Newton	0.0	14.3 s	2	3	1.4×10^{-13}
		0.01	2 m 19 s	6	7	1.3×10^{-6}
		0.1	3 m 08 s	7	8	2.1×10^{-7}
		1.0	4 m 11 s	8	9	1.7×10^{-7}
	BFGS	0.0	32.6 s	3	10	7.2×10^{-11}
		0.01	3 m 47 s	17	20	4.6×10^{-6}
		0.1	5 m 50 s	25	28	6.9×10^{-7}
		1.0	LS failed	60	74	1.1×10^{-5}
	BFGS-N	0.0	12.7 s	2	3	2.7×10^{-13}
		0.01	2 m 06 s	6	7	1.3×10^{-6}
		0.1	2 m 33 s	9	10	2.5×10^{-9}
		1.0	6 m 36 s	21	25	3.4×10^{-7}
	Gradient	0.0	27 m 35 s	500	501	1.6×10^{-2}
		0.01	1 h 25 m 10 s	500	501	5.6×10^{-2}
		0.1	1 h 28 m 51 s	500	501	4.0×10^{-1}
		1.0	1 h 23 m 17 s	500	501	4.0×10^{0}

REFERENCES

[1] L.-E. ANDERSSON AND T. ELFVING, *An algorithm for constrained interpolation*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 1012–1025.

[2] H. H. BAUSCHKE, J. M. BORWEIN, AND W. LI, *Strong conical hull intersection property, bounded linear regularity, Jameson's property* (*G*)*, and error bounds in convex optimization*, Math. Program., 86 (1999), pp. 135–160.

[3] H. H. BAUSCHKE, J. M. BORWEIN, AND P. TSENG, *Bounded linear regularity, strong CHIP, and CHIP are distinct properties*, J. Convex Anal., 7 (2000), pp. 395–412.

[4] R. BHATIA, *Matrix Analysis*, Springer-Verlag, New York, 1997.

[5] J. BORWEIN AND A. S. LEWIS, *Partially finite convex programming* I: *Quasi relative interiors and duality theory,* Math. Programming, 57 (1992), pp. 15–48.

[6] J. BORWEIN AND H. WOLKOWICZ, *A simple constraint qualification in infinite-dimensional programming,* Math. Programming, 35 (1986), pp. 83–96.

[7] S. BOYD AND L. XIAO, *Least-squares covariance matrix adjustment*, SIAM J. Matrix Anal. Appl., 27 (2005), pp. 532–546.

[8] X. CHEN, H. QI, AND P. TSENG, *Analysis of nonsmooth symmetric-matrix-valued functions with applications to semidefinite complementarity problems*, SIAM J. Optim., 13 (2003), pp. 960–985.

[9] C. K. CHUI, F. DEUTSCH, AND J. D. WARD, *Constrained best approximation in Hilbert space*, Constr. Approx., 6 (1990), pp. 35–64.

[10] C. K. CHUI, F. DEUTSCH, AND J. D. WARD, *Constrained best approximation in Hilbert space*

II, J. Approx. Theory, 71 (1992), pp. 213–238.
[11] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley and Sons, New York, 1983.
[12] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1983.
[13] F. Deutsch, *Best Approximation in Inner Product Spaces*, CMS Books Math./Ouvrages Math. SMC 7, Springer-Verlag, New York, 2001.
[14] F. Deutsch, W. Li, and J. D. Ward, *A dual approach to constrained interpolation from a convex subset of Hilbert space*, J. Approx. Theory, 90 (1997), pp. 385–414.
[15] F. Deutsch, W. Li, and J. D. Ward, *Best approximation from the intersection of a closed convex set and a polyhedron in Hilbert space, weak Slater conditions, and the strong conical hull intersection property*, SIAM J. Optim., 10 (1999), pp. 252–268.
[16] W. F. Donoghue, *Monotone Matrix Functions and Analytic Continuation*, Springer-Verlag, New York, 1974.
[17] A. L. Dontchev and B. D. Kalchev, *Duality and well-posedness in convex interpolation*, Numer. Funct. Anal. Optim., 10 (1989), pp. 673–689.
[18] A. L. Dontchev, H.-D. Qi, and L. Qi, *Convergence of Newton's method for convex best interpolation*, Numer. Math., 87 (2001), pp. 435–456.
[19] A. L. Dontchev, H.-D. Qi, and L. Qi, *Quadratic convergence of Newton's method for convex interpolation and smoothing*, Constr. Approx., 19 (2003), pp. 123–143.
[20] R. L. Dykstra, *An algorithm for restricted least squares regression*, J. Amer. Statist. Assoc., 78 (1983), pp. 837–842.
[21] F. Facchinei, *Minimization of SC^1 functions and the Maratos effect*, Oper. Res. Lett., 17 (1995), pp. 131–137.
[22] J. Favard, *Sur l'interpolation*, J. Math. Pures Appl. (9), 19 (1940), pp. 281–306.
[23] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436.
[24] N. J. Higham, *Computing a nearest symmetric positive semidefinite matrix*, Linear Algebra Appl., 103 (1988), pp. 103–118.
[25] N. J. Higham, *Computing the nearest correlation matrix—a problem from finance*, IMA J. Numer. Anal. 22 (2002), pp. 329–343.
[26] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
[27] U. Hornung, *Interpolation by smooth functions under restriction on the derivatives,* J. Approx. Theory, 28 (1980), pp. 227–237.
[28] G. Iliev and W. Pollul, *Convex interpolation by functions with minimal L_p norm* $(1 < p < \infty)$ *of the kth derivative,* in Mathematics and Mathematical Education (Sunny Beach, 1984), Bulg. Akad. Nauk, Sofia, Bulgaria, 1984, pp. 31–42.
[29] L. D. Irvine, S. P. Marin, and P. W. Smith, *Constrained interpolation and smoothing*, Constr. Approx., 2 (1986), pp. 129–151.
[30] M. Kojima and S. Shindo, *Extensions of Newton and quasi-Newton methods to systems of PC^1 equations*, J. Oper. Res. Soc. Japan, 29 (1986), pp. 352–374.
[31] N. Krislock, J. Lang, J. Varah, and D. K. Pai, *Local compliance estimation via positive semidefinite constrained least squares*, IEEE Transactions on Robotics and Automation, 20 (2004), pp. 1007–1011.
[32] B. Kummer, *Newton's method for nondifferentiable functions*, in Advances in Mathematical Optimization, Math. Res. 45, Akademie-Verlag, Berlin, 1988, pp. 114–125.
[33] K. Löwner, *Über monotone matrixfunctionen*, Math. Z., 38 (1934), pp. 177–216.
[34] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley and Sons, New York, 1969.
[35] J. Malick, *A dual approach to semidefinite least-squares problems*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 272–284.
[36] C. A. Micchelli, P. W. Smith, J. Swetits, and J. D. Ward, *Constrained L_p approximation*, Constr. Approx., 1 (1985), pp. 93–102.
[37] C. A. Micchelli and F. I. Utreras, *Smoothing and interpolation in a convex subset of a Hilbert space,* SIAM J. Sci. Statist. Comput., 9 (1988), pp. 728–746.
[38] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972.
[39] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer-Verlag, New York, 1999.
[40] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, SIAM, Philadelphia, 2000.
[41] J. S. Pang and L. Qi, *A globally convergent Newton method for convex SC^1 minimization problems*, J. Optim. Theory Appl., 85 (1995), pp. 633–648.

[42] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–367.
[43] R. T. Rockafellar, *Conjugate Duality and Optimization*, SIAM, Philadelphia, 1974.
[44] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[45] D. Sun, *A further result on an implicit function theorem for locally Lipschitz functions*, Oper. Res. Lett., 28 (2001), pp. 193–198.
[46] D. Sun and J. Sun, *Semismooth matrix valued functions*, Math. Oper. Res., 27 (2002), pp. 150–169.
[47] D. Sun and J. Sun, *Löwner's Operator and Spectral Functions in Euclidean Jordan Algebras,* Technical report, National University of Singapore, Singapore, 2004.
[48] K. C. Toh, *private communication*, 2005.
[49] K. C. Toh, R. H. Tütüncü, and M. J. Todd, *Inexact Primal-Dual Path-Following Algorithms for a Special Class of Convex Quadratic SDP and Related Problems*, Technical report, National University of Singapore, Singapore, 2005.
[50] P. Tseng, *Merit functions for semi-definite complementarity problems*, Math. Programming, 83 (1998), pp. 159–185.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 2, pp. 386–397

AN A PRIORI BOUND FOR AUTOMATED MULTILEVEL SUBSTRUCTURING*

KOLJA ELSSEL† AND HEINRICH VOSS†

Abstract. The automated multilevel substructuring (AMLS) method has been developed to reduce the computational demands of frequency response analysis and has recently been proposed as an alternative to iterative projection methods like those of Lanczos or Jacobi–Davidson for computing a large number of eigenvalues for matrices of very large dimension. Based on Schur complements and modal approximations of submatrices on several levels, AMLS constructs a projected eigenproblem which yields good approximations of eigenvalues at the lower end of the spectrum. Rewriting the original problem as a rational eigenproblem of the same dimension as the projected problem and taking advantage of a minmax characterization for the rational eigenproblem, we derive an a priori bound for the AMLS approximation of eigenvalues.

Key words. eigenvalues, AMLS, substructuring, nonlinear eigenproblem, minmax characterization

AMS subject classifications. 65F15, 65F50

DOI. 10.1137/040616097

1. Introduction.

Over the last few years, a new method for performing frequency response and eigenvalue analysis of complex finite element (FE) structures has been developed by Bennighof and collaborators [2], [3], [4], [5], [9], known as automated multilevel substructuring (AMLS). In AMLS the large FE model is recursively divided into many substructures on several levels based on the sparsity structure of the system matrices. Assuming that the interior degrees of freedom of substructures depend quasi-statically on the interface degrees of freedom, and modeling the deviation from quasi-static dependence in terms of a small number of selected substructure eigenmodes, the size of the FE model is reduced substantially while yielding satisfactory accuracy over a wide frequency range of interest. Recent studies ([11], [9], e.g.) in vibro-acoustic analysis of passenger car bodies, where very large FE models with more than one million degrees of freedom appear and several hundred eigenfrequencies and eigenmodes are needed, have shown that AMLS is considerably faster than Lanczos-type approaches.

We stress the fact that substructuring does not mean that the partitioning is obtained by a domain decomposition of a real structure. It is understood in a purely algebraic sense; i.e., the dissection of the matrices can be derived by applying a graph partitioner like CHACO [7] or METIS [10] to the matrix under consideration. However, because of the pictographic nomenclature of frequency response analysis we will use terms like substructure or eigenmode when recalling the AMLS method.

From a mathematical point of view AMLS is a projection method where the ansatz space is constructed by exploiting Schur complements of submatrices and truncation of spectral representations of subproblems. In this paper we will take advantage of the

*Received by the editors September 30, 2004; accepted for publication (in revised form) by I. S. Dhillon November 29, 2005; published electronically May 5, 2006.
http://www.siam.org/journals/simax/28-2/61609.html

†Institute of Mathematics, Hamburg University of Technology, D-21071 Hamburg, Germany (elssel@tu-harburg.de, voss@tu-harburg.de). The first author gratefully acknowledges financial support of this project by the German Foundation of Research (DFG) within the Graduiertenkolleg "Meerestechnische Konstruktionen."

fact that the original eigenproblem is equivalent to a rational eigenvalue problem of the same dimension as the projected problem in AMLS, which can be interpreted as exact condensation of the original eigenproblem with respect to an appropriate basis. The eigenvalues at the lower end of the spectrum can be characterized as minmax values of a Rayleigh functional of this rational eigenproblem. Hence, comparing the Rayleigh quotient of the projected problem and the Rayleigh functional of the rational problem, we derive an a priori bound for the error of the AMLS method. Following the same lines, the corresponding a priori bound for the static condensation method was already proved in [12].

In a recent paper Yang et al. [14] considered a single level version of AMLS (actually only the component mode synthesis method (CMS), since they did not reduce the number of interface degrees of freedom by modal truncation). The authors obtained a simple heuristic for choosing spectral components from each substructure, suggesting that we drop all eigenpairs (ω, ϕ) of substructures in the reduction process such that

$$\rho_1(\omega) := \frac{\lambda_1}{\omega - \lambda_1} \leq \tau.$$

Here λ_1 is the smallest eigenvalue of the problem under consideration and τ is a given tolerance. By our new a priori bound this omission rule guarantees that the relative error of the smallest eigenvalue of the projected problem is not greater than the tolerance τ. Moreover, for all eigenvalues $\lambda_j \in (0, \tau)$ the relative error of the CMS approximation is less than $\lambda_j/(\omega - \lambda_j)$.

Our presentation is organized as follows. In section 2, we give a brief overview of the AMLS method. We interpret AMLS as a sequence of consecutive CMS steps. In section 3, we provide a variational characterization of nonlinear and nonoverdamped eigenvalue problems. This characterization is exploited in section 4 to derive an a priori bound for the relative errors of the component mode synthesis method. As it is often the case for a priori bounds in general problems, the error bound overestimates the true relative error by one or two orders of magnitude. However, an example demonstrates that the bound cannot be improved without further assumptions. We also provide an error bound for the general multilevel substructuring method. The paper closes with a numerical example in section 5.

2. Substructuring of eigenproblems. We are concerned with the linear eigenvalue problem

$$Kx = \lambda Mx, \tag{2.1}$$

where $K \in \mathbb{R}^{n\times n}$ and $M \in \mathbb{R}^{n\times n}$ are symmetric and positive definite matrices. We recall that the terms structure, substructure, interface, and domain are meant in the algebraic sense to follow.

We first consider the CMS method, which is the essential building block of the AMLS method. Assume that the graph of the matrix $|K| + |M|$ is partitioned into r substructures such that the rows and columns of K can be reordered in the following way,

$$K = \begin{pmatrix} K_{\ell\ell 1} & \dots & O & K_{\ell i 1} \\ \vdots & \ddots & \vdots & \vdots \\ O & \dots & K_{\ell\ell r} & K_{\ell i r} \\ K_{i\ell 1} & \dots & K_{i\ell r} & K_{ii} \end{pmatrix},$$

and that M after reordering has the same block form. Here $K_{\ell\ell j}$, $j = 1, \ldots, r$, is the local stiffness matrix corresponding to the jth substructure, i denotes the set of interface vertices, and $K_{\ell ij}$ describes the interaction of the interface degrees of freedom and the jth substructure.

We distinguish only between local and interface degrees of freedom. Then K and M have the following form:

$$K = \begin{pmatrix} K_{\ell\ell} & K_{\ell i} \\ K_{i\ell} & K_{ii} \end{pmatrix} \quad \text{and} \quad M = \begin{pmatrix} M_{\ell\ell} & M_{\ell i} \\ M_{i\ell} & M_{ii} \end{pmatrix}. \tag{2.2}$$

We transform the matrix K to block diagonal form using block Gaussian elimination; i.e., we apply the congruence transformation

$$P = \begin{pmatrix} I & -K_{\ell\ell}^{-1} K_{\ell i} \\ 0 & I \end{pmatrix}$$

to the pencil (K, M) to obtain the equivalent pencil

$$(P^T K P, P^T M P) = \left(\begin{pmatrix} K_{\ell\ell} & 0 \\ 0 & \tilde{K}_{ii} \end{pmatrix}, \begin{pmatrix} M_{\ell\ell} & \tilde{M}_{\ell i} \\ \tilde{M}_{i\ell} & \tilde{M}_{ii} \end{pmatrix} \right). \tag{2.3}$$

Here $K_{\ell\ell}$ and $M_{\ell\ell}$ stay unchanged, and

$$\begin{aligned} \tilde{K}_{ii} &= K_{ii} - K_{\ell i}^T K_{\ell\ell}^{-1} K_{\ell i} \quad \text{is the Schur complement of } K_{\ell\ell}, \\ \tilde{M}_{\ell i} &= M_{\ell i} - M_{\ell\ell} K_{\ell\ell}^{-1} K_{\ell i} = \tilde{M}_{i\ell}^T, \\ \tilde{M}_{ii} &= M_{ii} - M_{i\ell} K_{\ell\ell}^{-1} K_{\ell i} - K_{i\ell} K_{\ell\ell}^{-1} M_{\ell i} + K_{i\ell} K_{\ell\ell}^{-1} M_{\ell\ell} K_{\ell\ell}^{-1} K_{\ell i}. \end{aligned}$$

We further transform the pencil (2.3), taking advantage of a modal basis for the local degrees of freedom. To this end we consider the eigenvalue problem

$$K_{\ell\ell}\Phi = M_{\ell\ell}\Phi\Omega, \qquad \Phi^T M_{\ell\ell}\Phi = I, \tag{2.4}$$

where Ω is a diagonal matrix containing the eigenvalues. Then applying the congruence transformation $\operatorname{diag}\{\Phi, I\}$ to (2.2) yields the equivalent pencil

$$\left(\begin{pmatrix} \Omega & 0 \\ 0 & \tilde{K}_{ii} \end{pmatrix}, \begin{pmatrix} I & \Phi^T \tilde{M}_{\ell i} \\ \tilde{M}_{i\ell}\Phi & \tilde{M}_{ii} \end{pmatrix} \right). \tag{2.5}$$

In structural dynamics, (2.5) is called Craig–Bampton form of the eigenvalue problem (2.1) corresponding to the partitioning (2.2). In terms of linear algebra it results from block Gaussian elimination to reduce K to block diagonal form, and diagonalization of the block $K_{\ell\ell}$ using a spectral basis.

Selecting some eigenmodes of problem (2.4) (usually the ones associated with eigenvalues below a cut-off threshold; however, in a recent paper Bai and Lia [1] suggested a different choice based on a moment-matching analysis) and dropping the rows and columns in (2.5) corresponding to the other modes, one arrives at the CMS method introduced by Hurty [8] and Craig and Bampton [6]. Hence, if the diagonal matrix Ω_1 contains on its diagonal the eigenvalues to drop and Φ_1 the corresponding eigenvectors, and if Ω_2 and Φ_2 contain the eigenvalues and eigenvectors, respectively, to keep, then the eigenproblem (2.5) can be rewritten as

$$\begin{pmatrix} \Omega_1 & 0 & 0 \\ 0 & \Omega_2 & 0 \\ 0 & 0 & \tilde{K}_{ii} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \lambda \begin{pmatrix} I & 0 & \tilde{M}_{\ell i 1} \\ 0 & I & \tilde{M}_{\ell i 2} \\ \tilde{M}_{i\ell 1} & \tilde{M}_{i\ell 2} & \tilde{M}_{ii} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \tag{2.6}$$

with

$$\tilde{M}_{\ell ij} = \Phi_j^T (M_{\ell i} - M_{\ell\ell} K_{\ell\ell}^{-1} K_{\ell i}) = \tilde{M}_{i\ell j}^T, \quad j = 1, 2,$$

and the CMS approximations to the eigenpairs of (2.1) are obtained from the reduced eigenvalue problem

$$\begin{pmatrix} \Omega_2 & 0 \\ 0 & \tilde{K}_{ii} \end{pmatrix} y = \lambda \begin{pmatrix} I & \tilde{M}_{\ell i 2} \\ \tilde{M}_{i\ell 2} & \tilde{M}_{ii} \end{pmatrix} y. \tag{2.7}$$

AMLS generalizes CMS in the following way. Again the graph of $|K| + |M|$ is partitioned into a small number of subgraphs, but more generally than in CMS these subgraphs in turn are substructured on a number p of levels, yielding a tree topology for the substructures. This induces the following partitioning of the index set $I = \{1, \ldots, n\}$ of degrees of freedom. Let I_1 be the set of indices corresponding to interface degrees of freedom on the coarsest level, and for $j = 2, \ldots, p$ define I_j to be the set of indices of interface degrees of freedom on the jth level which are not contained in I_{j-1}. Finally, let I_{p+1} be the set of interior degrees of freedom on the finest level.

With this notation, AMLS works as follows. Its first step is the CMS method with cut-off frequency τ_1 applied to the finest substructuring; i.e., I_{p+1} is the set of local degrees of freedom, and $\tilde{I}_{p+1} := \cup_{j=1}^{p} I_j$ is the set of interface degrees of freedom. After j steps, $1 \le j \le p-1$, one derives a reduced pencil

$$\left(\begin{pmatrix} \Omega_f & O & O \\ O & K_{\ell\ell}^{(j)} & K_{\ell i}^{(j)} \\ O & K_{i\ell}^{(j)} & K_{ii}^{(j)} \end{pmatrix}, \begin{pmatrix} M_{ff}^{(j)} & M_{f\ell}^{(j)} & M_{fi}^{(j)} \\ M_{\ell f}^{(j)} & M_{\ell\ell}^{(j)} & M_{\ell i}^{(j)} \\ M_{if}^{(j)} & M_{i\ell}^{(j)} & M_{ii}^{(j)} \end{pmatrix} \right), \tag{2.8}$$

where f denotes the degrees of freedom obtained in the spectral reduction in the previous steps, ℓ collects the indices in I_{p+1-j}, and i corresponds to the index set $\cup_{k=1}^{p-j} I_k$ of interface degrees of freedom on levels which are not yet treated. Applying the CMS method to the southeast 2×2 blocks of the matrices, i.e., annihilating the off-diagonal block $K_{\ell i}^{(j)}$ by block Gaussian elimination and reducing the set of ℓ-indices by spectral truncation with cut-off frequency τ_{j+1}, one arrives at the next level.

After p CMS steps we obtain the reduced problem

$$\left(\begin{pmatrix} \Omega_p & O \\ O & K_{\ell\ell}^{(p)} \end{pmatrix}, \begin{pmatrix} M_{ff}^{(p)} & M_{f\ell}^{(p)} \\ K_{\ell f}^{(p)} & M_{\ell\ell}^{(p)} \end{pmatrix} \right), \tag{2.9}$$

and a final spectral truncation of the lower-right blocks with cut-off frequency τ_{p+1} yields the reduction of problem (2.1) by AMLS.

We have chosen this unusual description of AMLS because it is very convenient for deriving the error bound in section 4. Note that this description neglects the algorithmically important fact that all matrices $K_{\ell\ell}^{(j)}$ and $M_{\ell\ell}^{(j)}$ are block diagonal. Hence, the annihilation of the off-diagonal blocks $K_{\ell i}^{(j)}$ and the spectral reduction on each level is quite inexpensive. A matrix and variational analysis of AMLS is contained in [5]; implementation details can be found in [9].

3. A minmax principle for nonlinear eigenproblems. In this section we provide a minmax result for symmetric nonlinear eigenvalue problems, which generalizes the well-known variational characterization of Poincaré for linear problems and

which will be used in section 4. We consider the nonlinear eigenvalue problem

$$T(\lambda)x = 0, \tag{3.1}$$

where $T(\lambda) \in \mathbb{R}^{n\times n}$ is a family of real symmetric matrices for every λ in an open real interval J.

For a linear symmetric and positive definite problem $Kx = \lambda Mx$ all eigenvalues are real. If they are ordered by magnitude regarding their multiplicity $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$, then it is well known that they can be characterized by the minmax principle of Poincaré,

$$\lambda_k = \min_{V\in S_k} \max_{x\in V,\ x\neq 0} \frac{x^T Kx}{x^T Mx}, \quad k = 1, 2, \ldots, n. \tag{3.2}$$

Here S_k denotes the set of all k dimensional subspaces of $\mathbb{R}^n$.

Similar results also hold for certain nonlinear eigenvalue problems (cf. [13]). We assume that for every fixed $x \in \mathbb{R}^n \setminus \{0\}$ the real function $f(\lambda; x) := x^T T(\lambda)x$ is continuously differentiable in J, and that the real equation

$$f(\lambda; x) = 0 \tag{3.3}$$

has at most one solution in J. Then (3.3) implicitly defines a functional p on some subset D of $\mathbb{R}^n \setminus \{0\}$. For a linear problem $T(\lambda) := \lambda M - K$ this is exactly the Rayleigh quotient, and we therefore call p the Rayleigh functional of problem (3.1).

Assume that

$$x^T T'(p(x))x > 0 \quad \text{for every } x \neq 0, \tag{3.4}$$

generalizing the definiteness requirement for M in the linear case.

If

$$\inf_{x\in D}\ p(x) \in J, \tag{3.5}$$

then it follows from the general minmax principle for nonlinear eigenproblems proved in [13] that problem (3.1) has at most n eigenvalues. These eigenvalues $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_m$, $m \le n$, ordered by magnitude satisfy the minmax characterization (cf. Theorems 2.1 and 2.9 in [13])

$$\lambda_k = \min_{V\in S_k, V\subset D\cup\{0\}} \max_{x\in V,\ x\neq 0}\ p(x), \quad k = 1, 2, \ldots, m. \tag{3.6}$$

4. A priori error bounds. We first consider the component mode synthesis method (2.7). If λ is not a diagonal entry of Ω_1, then the first equation of (2.6) yields

$$x_1 = \lambda(\Omega_1 - \lambda I)^{-1}\tilde{M}_{\ell i1}x_3,$$

and λ is an eigenvalue of (2.1) if and only if it is an eigenvalue of the rational eigenproblem

$$T(\lambda)y = 0, \tag{4.1}$$

where

$$T(\lambda) = -\begin{pmatrix}\Omega_2 & 0\\ 0 & \tilde{K}_{ii}\end{pmatrix} + \lambda\begin{pmatrix} I & \tilde{M}_{\ell i2}\\ \tilde{M}_{i\ell 2} & \tilde{M}_{ii}\end{pmatrix} + \lambda^2\begin{pmatrix}0\\ \tilde{M}_{i\ell 1}\end{pmatrix}(\Omega_1 - \lambda I)^{-1}\begin{pmatrix}0 & \tilde{M}_{\ell i1}\end{pmatrix}. \tag{4.2}$$

We denote by

$$\underline{\omega} := \min \operatorname{diag} \Omega_1 \tag{4.3}$$

the smallest eigenvalue of problem (2.4) neglected in the CMS method (which can be replaced by the cut-off threshold). Let $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ denote the eigenvalues of problem (2.1) ordered by magnitude, and let $m \in \mathbb{N}$ such that $\lambda_m < \underline{\omega} \le \lambda_{m+1}$. Then $\lambda_1, \ldots, \lambda_m \in J$ are the eigenvalues of the nonlinear eigenproblem (4.1) in J. We show that these eigenvalues satisfy the minmax principle in section 3.

For

$$f(\lambda; y) := y^T T(\lambda) y \tag{4.4}$$

it follows from the positive definiteness of $\left(\begin{smallmatrix} I & \tilde{M}_{\ell i 2} \\ \tilde{M}_{i\ell 2} & \tilde{M}_{ii} \end{smallmatrix}\right)$ that

$$\frac{\partial}{\partial \lambda} f(\lambda; y) = y^T \begin{pmatrix} I & \tilde{M}_{\ell i 2} \\ \tilde{M}_{i\ell 2} & \tilde{M}_{ii} \end{pmatrix} y + \sum_{\omega_j \ge \underline{\omega}} \frac{(2\lambda\omega_j - \lambda^2) a_j^2}{(\omega_j - \lambda)^2} > 0 \tag{4.5}$$

for every $y \in \mathbb{R}^\nu \setminus \{0\}$. Here ν denotes the dimension of the reduced problem (2.7), and $a := (0 \quad \tilde{M}_{\ell i 1}) y$.

Hence, due to the monotonicity of $f(\lambda; y)$ for every $y \in \mathbb{R}^\nu \setminus \{0\}$ the real equation $f(\lambda; y) = 0$ has at most one solution $p(y) \in J$, and condition (3.4) holds.

If $y := \left(\begin{smallmatrix} x_2 \\ x_3 \end{smallmatrix}\right) \in \mathbb{R}^\nu$ (where we have used the same partitioning of y as in section 2), then it easily seen that for $x_1 := \lambda(\Omega_1 - \lambda I)^{-1} \tilde{M}_{i\ell 1} x_3$ it holds that

$$(x_2^T, x_3^T) T(\lambda) \begin{pmatrix} x_2 \\ x_3 \end{pmatrix} = (x_1^T, x_2^T, x_3^T) \begin{pmatrix} \lambda I - \Omega_1 & O & \lambda \tilde{M}_{\ell i 1} \\ O & \lambda I - \Omega_2 & \lambda \tilde{M}_{\ell i 2} \\ \lambda \tilde{M}_{i\ell 1} & \lambda \tilde{M}_{i\ell 2} & \lambda \tilde{M}_{ii} - \tilde{K}_{ii} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

Therefore, $\left(\begin{smallmatrix} x_2 \\ x_3 \end{smallmatrix}\right) \in D$ if and only if the Rayleigh quotient R of the linear eigenproblem (2.6) at $x := (x_1^T, x_2^T, x_3^T)^T$ is contained in J, and $p(y) = R(x)$. In particular,

$$\inf_{y \in D} \; p(y) = \inf_{x \in \mathbb{R}^n,\, x \ne 0} R(x) = \lambda_1 \in J,$$

and the eigenvalues $\lambda_1, \ldots, \lambda_m$ of problem (4.1) satisfy the minmax characterization (3.6).

The eigenvalues $\tilde{\lambda}_1 \le \tilde{\lambda}_2 \le \cdots \le \tilde{\lambda}_\nu$ of the reduced problem (2.7) are minmax values of the Rayleigh quotient $\rho(x)$ corresponding to problem (2.7). Comparing p and ρ on appropriate subspaces of $\mathbb{R}^\nu$, we arrive at the following a priori bound for the relative errors of the CMS approximations $\tilde{\lambda}_j$ to λ_j.

THEOREM 4.1. [illegible] $K, M \in \mathbb{R}^{n \times n}$ [illegible] $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ [illegible] (2.1) [illegible] $\tilde{\lambda}_1 \le \tilde{\lambda}_2 \le \cdots \le \tilde{\lambda}_\nu$ [illegible] (2.7) [illegible] (2.1) [illegible] $|K| + |M|$ [illegible] $\underline{\omega}$ [illegible] $(0, \underline{\omega})$ [illegible] m [illegible] $\lambda_1, \ldots, \lambda_m$ [illegible] (2.1) [illegible]

$$0 \le \frac{\tilde{\lambda}_j - \lambda_j}{\lambda_j} \le \frac{\lambda_j}{\underline{\omega} - \lambda_j} \le \frac{\tilde{\lambda}_j}{\underline{\omega} - \tilde{\lambda}_j}, \quad j = 1, \ldots, m. \tag{4.6}$$

The left inequality, i.e., $\lambda_j \le \tilde{\lambda}_j$, is trivial since CMS is a projection method. The right inequality follows from the monotonicity of the function $\lambda \mapsto \lambda/(\underline{\omega}-\lambda)$.

To prove the inequality in the middle, denote by $V \in S_j$, $V \setminus \{0\} \subset D$ the j-dimensional subspace of $\mathbb{R}^\nu$ such that

$$\lambda_j = \max_{y\in V,\ y\neq 0}\ p(y).$$

Then $p(y) \le \lambda_j$ for every $y \in V$, $y \neq 0$, and therefore it follows from the monotonicity of the function $f(\lambda; y)$ with respect to λ that

$$-y^T\begin{pmatrix}\Omega_2 & 0\\ 0 & \tilde{K}_{ii}\end{pmatrix}y+\lambda_j y^T\begin{pmatrix}I & \tilde{M}_{\ell i2}\\ \tilde{M}_{i\ell 2} & \tilde{M}_{ii}\end{pmatrix}y+\lambda_j^2 y^T\begin{pmatrix}0\\ \tilde{M}_{i\ell 1}\end{pmatrix}(\Omega_1-\lambda_j I)^{-1}\begin{pmatrix}0 & \tilde{M}_{\ell i1}\end{pmatrix}y \ge 0.$$

Hence, for every $y \in V$, $y \neq 0$ one obtains

$$\lambda_j \ge \frac{y^T\begin{pmatrix}\Omega_2 & 0\\ 0 & \tilde{K}_{ii}\end{pmatrix}y}{y^T\begin{pmatrix}I & \tilde{M}_{\ell i2}\\ \tilde{M}_{i\ell 2} & \tilde{M}_{ii}\end{pmatrix}y} - \lambda_j^2\frac{y^T\begin{pmatrix}0\\ \tilde{M}_{i\ell 1}\end{pmatrix}(\Omega_1-\lambda_j I)^{-1}\begin{pmatrix}0 & \tilde{M}_{\ell i1}\end{pmatrix}y}{y^T\begin{pmatrix}I & \tilde{M}_{\ell i2}\\ \tilde{M}_{i\ell 2} & \tilde{M}_{ii}\end{pmatrix}y}.$$

In particular, for $\hat{y} \in V$ such that $\rho(\hat{y}) = \max_{y\in V,\ y\neq 0}\ \rho(y)$ we have

$$\begin{aligned}\lambda_j &\ge \max_{y\in V,\ y\neq 0}\ \rho(y) - \lambda_j^2\frac{\hat{y}^T\begin{pmatrix}0\\ \tilde{M}_{i\ell 1}\end{pmatrix}(\Omega_1-\lambda_j I)^{-1}\begin{pmatrix}0 & \tilde{M}_{\ell i1}\end{pmatrix}\hat{y}}{\hat{y}^T\begin{pmatrix}I & \tilde{M}_{\ell i2}\\ \tilde{M}_{i\ell 2} & \tilde{M}_{ii}\end{pmatrix}\hat{y}}\\ &\ge \min_{\dim W=j}\ \max_{y\in W,\ y\neq 0}\ \rho(y) - \lambda_j^2\frac{\hat{y}^T\begin{pmatrix}0\\ \tilde{M}_{i\ell 1}\end{pmatrix}(\Omega_1-\lambda_j I)^{-1}\begin{pmatrix}0 & \tilde{M}_{\ell i1}\end{pmatrix}\hat{y}}{\hat{y}^T\begin{pmatrix}I & \tilde{M}_{\ell i2}\\ \tilde{M}_{i\ell 2} & \tilde{M}_{ii}\end{pmatrix}\hat{y}}\\ &\ge \tilde{\lambda}_j - \frac{\lambda_j^2}{\underline{\omega}-\lambda_j}\max_{y\in\mathbb{R}^n, y\neq 0}\frac{y^T\begin{pmatrix}0\\ \tilde{M}_{i\ell 1}\end{pmatrix}\begin{pmatrix}0 & \tilde{M}_{\ell i1}\end{pmatrix}y}{y^T\begin{pmatrix}I & \tilde{M}_{\ell i2}\\ \tilde{M}_{i\ell 2} & \tilde{M}_{ii}\end{pmatrix}y}. \qquad (4.7)\end{aligned}$$

From the positive definiteness of the transformed mass matrix

$$\begin{pmatrix}I & 0 & \tilde{M}_{\ell i1}\\ 0 & I & \tilde{M}_{\ell i2}\\ \tilde{M}_{i\ell 1} & \tilde{M}_{i\ell 2} & \tilde{M}_{ii}\end{pmatrix}$$

it follows that the Schur complement

$$\begin{pmatrix}I & \tilde{M}_{\ell i2}\\ \tilde{M}_{i\ell 2} & \tilde{M}_{ii}\end{pmatrix} - \begin{pmatrix}0\\ \tilde{M}_{i\ell 1}\end{pmatrix}\begin{pmatrix}0 & \tilde{M}_{\ell i1}\end{pmatrix}$$

is positive definite as well. Thus,

$$\max_{y\in\mathbb{R}^n, y\neq 0}\frac{y^T\begin{pmatrix}0\\ \tilde{M}_{i\ell 1}\end{pmatrix}\begin{pmatrix}0 & \tilde{M}_{\ell i1}\end{pmatrix}y}{y^T\begin{pmatrix}I & \tilde{M}_{\ell i2}\\ \tilde{M}_{i\ell 2} & \tilde{M}_{ii}\end{pmatrix}y} \le 1,$$

and (4.7) yields

$$\lambda_j \geq \tilde{\lambda}_j - \frac{\lambda_j^2}{\underline{\omega} - \lambda_j}, \tag{4.8}$$

which completes the proof. □

Numerical examples (cf. section 5) demonstrate that the error bound in (4.6) overestimates the true relative error of CMS by one or two orders of magnitude. The following example demonstrates that the bound cannot be improved without further assumptions.

Consider the eigenvalue problem

$$\begin{pmatrix} \omega & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \lambda \begin{pmatrix} 1 & 0 & m \\ 0 & 1 & 0 \\ m & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \tag{4.9}$$

where $\omega > 1$ and $m \in (0, 1)$. Its eigenvalues are

$$\lambda_2 = 1 \quad \text{and} \quad \lambda_{1/3} = \frac{1}{2(1-m^2)} \left(\omega + 1 \mp \sqrt{(\omega+1)^2 - 4\omega(1-m^2)} \right).$$

Let x_1 and x_2 be the local degrees of freedom, and x_3 be the interface variable. With cut-off frequency ω the reduced problem is

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_2 \\ x_3 \end{pmatrix} = \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_2 \\ x_3 \end{pmatrix},$$

and its minimum eigenvalue is $\tilde{\lambda}_1 = 1$.

Letting $m \to 1-0$, l'Hospital's rule yields that the left-hand side of (4.6) converges to

$$\begin{aligned} \lim_{m \to 1-0} \frac{\tilde{\lambda}_1 - \lambda_1}{\lambda_1} &= \lim_{m \to 1-0} \frac{1}{\lambda_1} - 1 = \lim_{m \to 1-0} \frac{2(1-m^2)}{\omega + 1 - \sqrt{(\omega+1)^2 - 4\omega(1-m^2)}} - 1 \\ &= \lim_{m \to 1-0} \frac{-2}{-0.5((\omega+1)^2 - 4\omega(1-m^2))^{-1/2} 4\omega} - 1 = \frac{1}{\omega}, \end{aligned}$$

and the right-hand side also converges to

$$\lim_{m \to 1-0} \frac{\lambda_1}{\omega - \lambda_1} = \frac{1}{\omega \lim_{m \to 1-0} \frac{1}{\lambda_1} - 1} = \frac{1}{\omega}.$$

Based on accuracy considerations and an a priori error bound for the smallest eigenvalue (which, however, usually cannot be evaluated, since it depends on unknown quantities like a bound for the components of $\tilde{M}_{\ell i 1} \tilde{x}_3$, where $\tilde{x}_3$ is the interface portion of an eigenvector of (2.5) or the minimal distance of neglected diagonal entries of Ω_1 belonging to the same substructure) Yang et al. [14] suggested that one neglect all eigenmodes (ω_j, ϕ_j) in (2.6) for which

$$\frac{\lambda_1}{\omega_j - \lambda_1} < \tau,$$

where $\tau \ll 1$ is a small quantity. Theorem 4.1 guarantees that with this choice the relative error of the CMS approximation $\tilde{\lambda}_1$ to the smallest eigenvalues λ_1 is less than τ.

Since AMLS can be understood as a sequence of p consecutive CMS steps and a terminating spectral truncation, it is clear how to obtain an a priori bound for the general AMLS method. Every reduction step in which a quasi-static/modal representation is obtained, and the dimension reduced by spectral truncation, is identical to a CMS step utilizing the substructuring of the next level.

Hence, if $\lambda_j^{(\nu)}$ denotes the eigenvalues of the reduced eigenvalue problem corresponding to the νth level ordered by magnitude, then it holds by (4.8) that

$$\lambda_j^{(\nu)} \le \lambda_j^{(\nu-1)} \left(1 + \frac{\lambda_j^{(\nu-1)}}{\omega_\nu - \lambda_j^{(\nu-1)}}\right), \quad \nu = 1, 2, \dots, p+1, \tag{4.10}$$

where on the νth level eigenvalues exceeding ω_ν are neglected. Here, $\lambda_j^{(0)} := \lambda_j$, and $\lambda_j^{(p+1)}$ denote the eigenvalues of the projected eigenproblem of AMLS with p levels of substructuring.

Thus, it follows for all $\lambda_j \le \min_{\nu=1,\dots,p} \omega_\nu$ that

$$\lambda_j^{(p+1)} \le \lambda_j \prod_{\nu=0}^{p} \left(1 + \frac{\lambda_j^{(\nu)}}{\omega_{\nu+1} - \lambda_j^{(\nu)}}\right), \tag{4.11}$$

and we have proved the following result.

THEOREM 4.2. K M λ_j $j = 1, \dots, n$ 4.1 $|K| + |M|$ p $\tilde{\lambda}_1^{(\nu)} \le \tilde{\lambda}_2^{(\nu)} \le \cdots$ ω_ν ν $m \in \mathbb{N}$

$$\lambda_m < \min_{\nu=0,\dots,p} \omega_\nu \le \lambda_{m+1},$$

$$\frac{\tilde{\lambda}_j - \lambda_j}{\lambda_j} \le \prod_{\nu=0}^{p} \left(1 + \frac{\lambda_j^{(\nu)}}{\omega_\nu - \lambda_j^{(\nu)}}\right) - 1, \quad j = 1, \dots, m. \tag{4.12}$$

Since the final problem is a projection of each of the intermediate eigenproblems in the AMLS reduction, it follows from the minmax characterization that $\lambda_j^{(\nu)} \le \tilde{\lambda}_j$ for $\nu = 0, \dots, p$. Therefore the a priori bound (4.12) can be replaced by the computable bound

$$\frac{\tilde{\lambda}_j - \lambda_j}{\lambda_j} \le \prod_{\nu=0}^{p} \left(1 + \frac{\tilde{\lambda}_j}{\omega_\nu - \tilde{\lambda}_j}\right) - 1, \quad j = 1, \dots, m. \tag{4.13}$$

5. Numerical experiments. To verify the quality of our a priori bounds we considered the problem of determining the 50 smallest eigenvalues of the FE model of a container ship with 35262 degrees of freedom (cf. Figure 1).

To apply the CMS method and the single-level version of AMLS we partitioned the FEM model into ten substructures, as shown in Figure 2. This substructuring by hand, suggested by the geometry of the ship, yielded a much smaller number of interface degrees of freedom than automatic graph partitioners, which try to construct a partition where the substructures have nearly equal size. For instance, our model

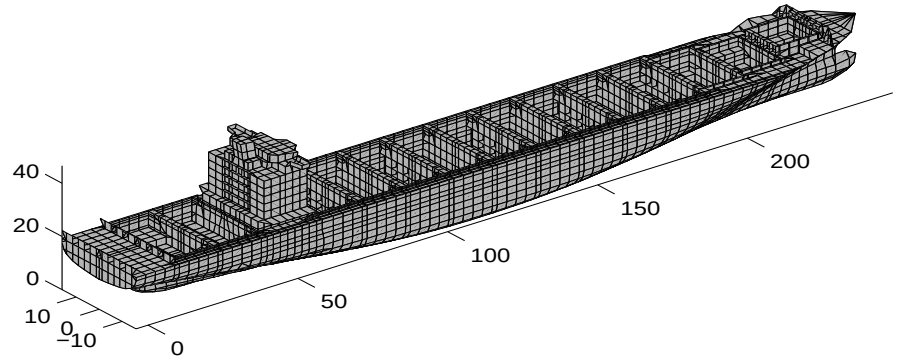


FIG. 1. *FE model of a container ship.*

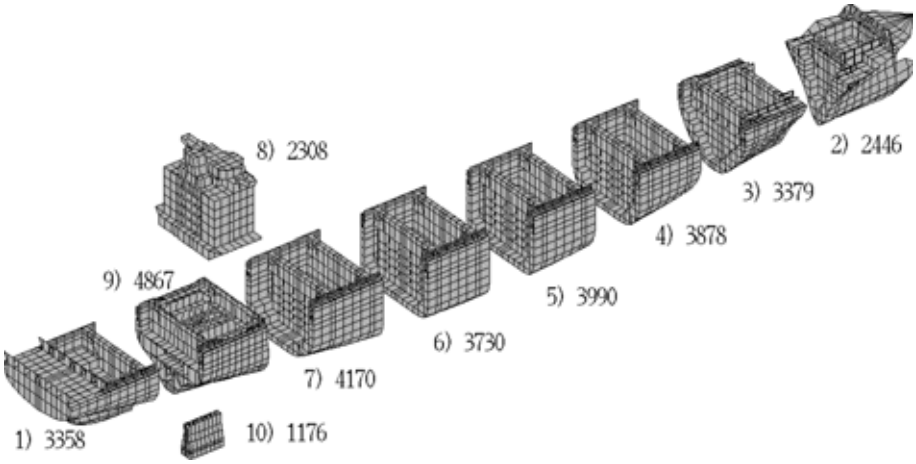


FIG. 2. *Substructuring.*

ends up with 1960 degrees of freedom on the interfaces, whereas CHACO [7] ends up with a substructuring into ten substructures with 4985 interface degrees of freedom.

We solved the eigenproblem by the CMS method using a cut-off bound of 20,000 (about 10 times the largest wanted eigenvalue $\lambda_{50} \approx 2183$). 329 eigenvalues of the substructure problems were less than our threshold, and the dimension of the resulting projected problem was 2289. Figure 3 shows the relative errors for the smallest 50 eigenvalues (lower crosses) and the error bounds by Theorem 4.1 (upper crosses). We reduced the interface degrees of freedom as well with the same cut-off bound 20,000. This reduced the dimension of the projected eigenproblem to 436. The relative errors (lower circles) and bounds by Theorem 4.2 with $p = 1$ (upper circles) are also shown in Figure 3.

We substructured the FE model using METIS with four levels of substructuring. Neglecting eigenvalues exceeding 20,000 and 40,000 on all levels, AMLS produced a projected eigenvalue problem of dimension 451 and 911, respectively. The relative errors and the bounds are shown in Figure 4, where the lower and upper crosses

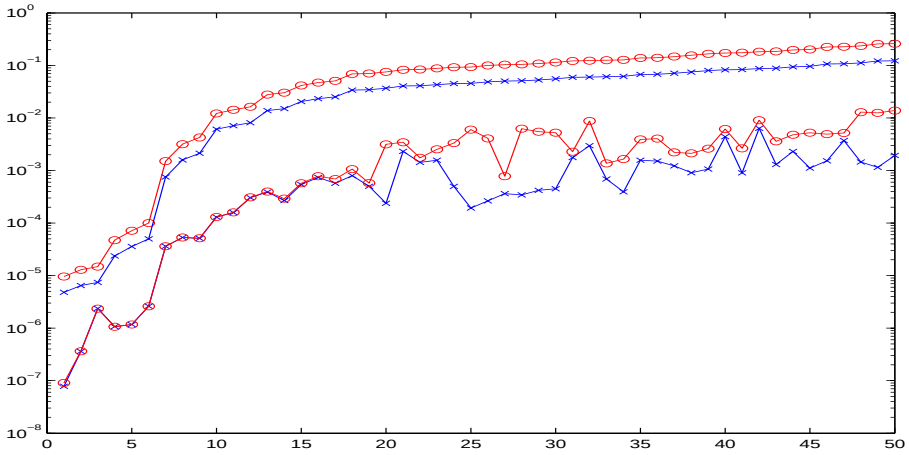


FIG. 3. *Errors and bounds for CMS and single level AMLS. See text for further explanation.*

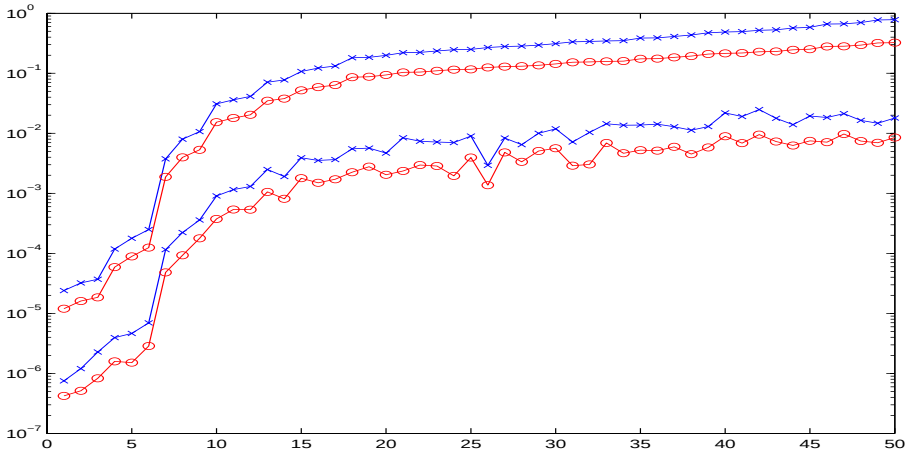


FIG. 4. *Errors and bounds for AMLS. See text for further explanation.*

correspond to the threshold 20,000, and the lower and upper circles to 40,000.

Acknowledgments. Thanks are due to Christian Cabos, Germanischer Lloyd, who provided us with the finite element model of the container ship.

REFERENCES

[1] Z. BAI AND B.-S. LIA, *Towards an optimal substructuring method for model reduction*, in Proceedings of PARA'04, Lyngby, Denmark, 2004, Lecture Notes in Comput. Sci. 3732, Springer-Verlag, Berlin, 2005, pp. 276–285.

[2] J. K. BENNIGHOF AND M. F. KAPLAN, *Frequency sweep analysis using multi-level substructuring, global modes and iteration*, in Proceedings of the 39th Annual AIAA Structural Dynamics and Materials Conference, Long Beach, CA, 1998.

[3] J. K. BENNIGHOF, M. F. KAPLAN, M. B. MULLER, AND M. KIM, *Meeting the NVH computational challenge: Automated multi-level substructuring*, in Proceedings of the 18th International Modal Analysis Conference, San Antonio, TX, 2000, pp. 909–915.

[4] J. K. Bennighof and C. K. Kim, *An adaptive multi-level substructuring method for efficient modeling of complex structures*, in Proceedings of the 33rd Annual AIAA Structural Dynamics and Materials Conference, Dallas, TX, 1992, AIAA, Reston, VA, 1992, pp. 1631–1639.

[5] J. K. Bennighof and R. B. Lehoucq, *An automated multilevel substructuring method for eigenspace computation in linear elastodynamics*, SIAM J. Sci. Comput., 25 (2004), pp. 2084–2106.

[6] R. R. Craig, Jr., and M. C. C. Bampton, *Coupling of substructures for dynamic analysis*, J. AIAA, 6 (1968), pp. 1313–1319.

[7] B. Hendrickson and R. Leland, *The Chaco User's Guide: Version* 2.0., Technical Report SAND94-2692, Sandia National Laboratories, Albuquerque, NM, 1994.

[8] W. C. Hurty, *Vibration of structure systems by component-mode synthesis*, J. Engrg. Mech. Div., ASCE, 86 (1960), pp. 51–69.

[9] M. F. Kaplan, *Implementation of Automated Multilevel Substructuring for Frequency Response Analysis of Structures*, Ph.D. thesis, Department of Aerospace Engineering & Engineering Mechanics, University of Texas at Austin, Austin, TX, 2001.

[10] G. Karypis and V. Kumar, *METIS. A software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices, Version* 4.0, Technical report, University of Minnesota, Minneapolis, MN, 1998.

[11] A. Kropp and D. Heiserer, *Efficient broadband vibro-acoustic analysis of passenger car bodies using an FE-based component mode synthesis approach*, J. Comput. Acoustics, 11 (2003), pp. 139–157.

[12] H. Voss, *An error bound for eigenvalue analysis by nodal condensation*, in Numerical Treatment of Eigenvalue Problems, Internat. Ser. Numer. Math., J. Albrecht, L. Collatz, and W. Velte, eds., Birkhäuser, Basel, 1984, Vol. 3, pp. 205–214.

[13] H. Voss and B. Werner, *A minimax principle for nonlinear eigenvalue problems with applications to nonoverdamped systems*, Math. Methods Appl. Sci., 4 (1982), pp. 415–424.

[14] C. Yang, W. Gao, Z. Bai, X. S. Li, L.-Q. Lee, P. Husbands, and E. Ng, *An algebraic sub-structuring method for large-scale eigenvalue calculations*, SIAM J. Sci. Comput., 27 (2005), pp. 873–892.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 2, pp. 398–424

THE SNAP-BACK PIVOTING METHOD FOR SYMMETRIC BANDED INDEFINITE MATRICES*

DROR IRONY† AND SIVAN TOLEDO†

Abstract. The four existing stable factorization methods for symmetric indefinite matrices suffer serious defects when applied to banded matrices. Partial pivoting (row or column exchanges) maintains a band structure in the reduced matrix and the factors, but destroys symmetry completely once an off-diagonal pivot is used. Two-by-two block pivoting and Gaussian reduction to tridiagonal (Aasen's algorithm) maintain symmetry at all times, but quickly destroy the band structure in the reduced matrices. Orthogonal reductions to tridiagonal maintain both symmetry and the band structure, but are too expensive for linear-equation solvers.

We propose a new pivoting method, which we call *snap-back* pivoting. When applied to banded symmetric matrices, it maintains the band structure (like partial pivoting does), it keeps the reduced matrix symmetric (like 2-by-2 pivoting and reductions to tridiagonal), and it is fast.

Snap-back pivoting reduces the matrix to a diagonal form using a sequence of elementary elimination steps, most of which are applied symmetrically from the left and from the right (but some are applied unsymmetrically).

In snap-back pivoting, if the next diagonal element is too small, the next pivoting step might be unsymmetric, leading to asymmetry in the next row and column of the factors. But the reduced matrix snaps back to symmetry once the next step is completed.

Key words. symmetric-indefinite matrices, pivoting, banded matrices, matrix factorizations, element growth

AMS subject classifications. 15A06, 15A23, 65F05

DOI. 10.1137/040610106

1. Introduction.

We propose a new method for the direct solution of a linear system of equations $Ax = b$ where A is an n-by-n banded symmetric indefinite matrix with half bandwidth m. The method performs $O(nm^2)$ work. Our method reduces A to a diagonal matrix by eliminating one or two rows and columns in each step. After each elimination step, the reduced matrix is a banded symmetric matrix with half bandwidth at most $2m$. Although at each step the reduced matrix is symmetric, the factors corresponding to steps in which we eliminate two columns together may not be symmetric. The algorithm requires a mixture of symmetric and unsymmetric data. The element growth in the reduced matrices is bounded by 4^{n-1}. Elements of the factors are bounded by 3 or by the elements of the reduced matrices. Our method achieves these goals using an intricate elimination scheme that employs both Gaussian row and column operations and Givens rotations.

Our method reduces the matrix to a diagonal one or two rows/columns at a time. If the next diagonal element is large, an ordinary symmetric Gaussian elimination step will reduce the next row and column. Such a step adds a column to the left factor and its transpose to the right factor. Since a symmetric matrix is subtracted from the symmetric trailing submatrix, it remains symmetric. If the next diagonal

*Received by the editors June 16, 2004; accepted for publication (in revised form) by N. J. Higham November 30, 2005; published electronically May 26, 2006. This research was supported in part by an IBM Faculty Partnership Award, by grants 572/00 and 848/04 from the Israel Science Foundation (founded by the Israel Academy of Sciences and Humanities), and by grant 2002261 from the United-States-Israel Binational Science Foundation.

http://www.siam.org/journals/simax/28-2/61010.html

†School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel (irony@tau.ac.il, stoledo@tau.ac.il).

element is too small, we use a more complex elimination step. During that step, the trailing submatrix becomes unsymmetric, but it snaps back to symmetry at the end of the step. That is why we call our method [illegible] pivoting. Such a complex elimination step contributes one row to the right factor that is not a transpose of a column in the left factor. Therefore, the amount of asymmetry in the factors depends on the number of these complex elimination steps.

The paper is organized as follows. The next section surveys existing factorization methods that can be applied to banded symmetric matrices. Section 3 describes the new method; the presentation is not specific to banded matrices. Section 4 shows how to adapt the general method to preserve the band structure. Section 5 bounds the element growth in the reduced matrices and in the factors. Section 6 presents the results of experiments that we conducted to assess the performance and stability of our method. We present our conclusions from this research and list open problems in section 7.

2. Related work. Several existing direct factorization methods can reduce a banded symmetric matrix to a diagonal, tridiagonal, or block-diagonal form. If the matrix is indefinite, either some form of pivoting or an orthogonal reduction must be employed to ensure stability.

Gaussian elimination with partial pivoting (GEPP) maintains a band structure in the reduced matrices, but not symmetry. The first off-diagonal pivot that is chosen completely destroys symmetry in the reduced matrix, so from that row/column until the end of the matrix, both the reduced matrix and the factors become unsymmetric. The loss of symmetry results in doubling the computational and storage costs. On the other hand, GEPP maintains a band structure in the reduced matrix and in the factors. No matter how many off-diagonal pivots are chosen, the reduced matrix and the right upper-triangular factor have half bandwidth at most $2m$. The left lower triangular factor, and also the lower triangular part of the reduced matrices, maintain half bandwidth m.

Two other factorization techniques employ pivoting but maintain symmetry in the factors and in the reduced matrices. One technique, which is used in several algorithms [4, 6, 7, 8, 14], uses 2-by-2 block pivots. In other words, it reduces the matrix to a block diagonal form with 1-by-1 and 2-by-2 blocks. This technique usually cannot be applied to banded matrices, because every 2-by-2 pivot might increase the half-bandwidth by $m - 2$. Therefore, the half bandwidth can quickly expand. This bandwidth expansion can lead to catastrophic increase in the computational and storage requirements of the algorithm. Still, in some special cases this technique can be applied to banded matrices. Bunch and Kaufman showed that if $m \leq 2$, then one of the variants of their algorithm (variant D) maintains the band structure [4]. They also showed that the number of 2-by-2 pivots is bounded by the minimum of the number of positive and the number of negative eigenvalues of A. This observation led Jones and Patrick [14] to propose that the Bunch–Kaufman algorithm can be used when A has very few negative or very few positive eigenvalues; in such cases, the bandwidth expansion might still be preferable to GEPP. A combination of 1-by-1 and 2-by-2 pivoting steps is also used in multifrontal factorization algorithms for general sparse matrices [6, 7, 8], but without any a priori bound on the fill that pivoting might cause.

In the 2-by-2 pivoting methods of Bunch and Kaufman, the element growth in the reduced matrices is bounded by 2.57^{n-1}, but the magnitude of elements in the lower-triangular factor is not bounded by the magnitude of elements in the reduced matrices. (In most other factorization methods, the elements of the factors are bounded by 1

or by the elements of the reduced matrices.) This fact, along with some disappointing practical experience, led some to question the reliability of these methods [2]. However, the methods were found to be formally backward stable [11, 12].

Another pivoting symmetric factorization technique reduced A to a tridiagonal form using a sequence of 1-by-1 but off-diagonal pivots. The resulting tridiagonal matrix is subsequently factored using GEPP, but the cost of that step is usually insignificant. This technique was initially proposed for dense matrices by Parlett and Reid [17], and later improved by Aasen [1]. The trouble with these methods is that they can also quickly expand the bandwidth, so they have never been applied to banded matrices.

Another way to reduce a banded symmetric matrix to a tridiagonal form is to employ orthonormal transformations. This technique is used in eigensolvers, which can only use orthonormal transformations. Algorithms that use this technique [3, 5, 15, 16, 18, 19] annihilate in every step one row and column, or even just one element. The annihilation creates fill in the form of a bulge, which expands the band. To avoid gradual band expansion, the algorithm then "chases" the bulge down the matrix, so that the matrix regains the original half bandwidth m before the next annihilation step. The trouble with these approaches is that chasing the bulge is expensive, so the total computational cost of these algorithm is proportional to n^2m. By contrast, the cost of GEPP, as well as the cost of our algorithm, is only nm^2. For $m \ll n$, the difference can be enormous.

From the band-preservation viewpoint, the difference between 2-by-2 block algorithms and the orthonormal algorithms lies not in the elementary transformations that are used, but in whether the bulge is chased or not. In the 2-by-2 block algorithms of Bunch and Kaufman, as well as in the Parlett–Reid and Aasen algorithms, the bulge is not chased, so the band expands without a bound. In the orthonormal reductions to tridiagonal, the bulge is chased, so the bandwidth remains bounded, but at an unacceptable cost to linear-equation solvers. (The use of orthonormal transformations also causes the bulge to always appear, whereas in the other factorization algorithms the size of the bulge depend on the choice of pivots, but that is not the main point here.) We believe that a bulge-chasing variant of the other algorithms can also be developed, but that its cost will still be proportional to n^2m, just like the orthonormal reductions.

3. Snap-back pivoting. This section presents the new pivoting strategy that we proposed, called [illegible] pivoting. We mostly ignore the issue of the bandwidth of the matrix in this section and focus instead on the mathematical and algorithmic ideas. Bandwidth issues do narrow down some of the algorithmic design space that we explore; the text explicitly highlights these cases. The next section explains how to apply snap-back pivoting to banded matrices.

Let A be a symmetric n-by-n nonsingular matrix. We show how to perform a sequence of elementary elimination steps that reduce the matrix to a diagonal form. The sequence is [illegible]: the trailing uneliminated submatrix is always symmetric. Most, but not all, of the elementary transformations are applied symmetrically from the left and from the right. We use three classes of elementary transformations: Gauss transforms, Givens rotations, and permutations. Each elimination step, which often consists of applying multiple elementary transformations, eliminates the offdiagonals in either one row and column or in two. The first row and column are always eliminated; in some cases, another row and columns, not necessarily the second, are also eliminated.

Our method uses three kinds of elimination steps. The next subsection provides a high-level overview of the possible elimination steps and the one that follows specifies the transformation matrices in detail. To further clarify the algorithm, we have produced a Matlab implementation (of the banded variant, described below). This implementation is freely available.[1] It stops after every elimination step, shows the nonzero structure of the reduced matrix, and prints the class of transformation that it just applied. It also produces transformation matrices that correspond exactly to the notation in this section.

3.1. An overview of the elimination process. We begin the presentation with a high-level overview of the structure of the algorithm.

Elimination steps of the first kind. When a_{11} is nonzero (to avoid growth it will need to be not only nonzero, but large; we ignore numerical issues for now), we can eliminate the offdiagonals in the first row and column using ordinary symmetric Gaussian elimination,

$$A = \begin{bmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{bmatrix} \longrightarrow \begin{bmatrix} \times & & & & \\ & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & \times & \times & \times & \times \end{bmatrix} = L^{-1}AL^{-T}.$$

In this notation, $\times$ symbols denote nonzero elements in a symmetric matrix (that is, if a_{ij} is denoted by a $\times$, then $a_{ij} = a_{ji}$, and they are not necessarily zero). Elements that are blank are zeros. The matrix product on the right means that we transform the matrix on the left, A, to the matrix on the right by multiplying A by a lower triangular matrix L^{-1} and its transpose. We also use this kind of elimination step, with L an identity matrix, when the first row and column are identically zero. We defer the exact specification of the transformation matrices to the next subsection; here L is an elementary Gauss transform.

Elimination steps of the second kind. When a_{11} is zero (or simply too small), our method uses a more elaborate sequence of elementary transformations. We begin with a series of either Givens or Gauss transforms that eliminate all the nonzeros in the first column except for the last,

$$A = \begin{bmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{bmatrix} \longrightarrow \begin{bmatrix} \times & & & & \times \\ & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{bmatrix} = Y^{-1}AY^{-T}.$$

Next, we eliminate element $(n, 1)$ in the reduced matrix, which we shall show later is always nonzero, using a Givens rotation that transforms the first and last rows. If element $\left[Y^{-1}AY^{-T}\right]_{11} = 0$, the Givens rotation is essentially a row exchange, so we

[1] The code is available from http://www.tau.ac.il/∼stoledo/research.html.

get the following form:

$$\begin{bmatrix} \times & & & & \times \\ & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{bmatrix} \longrightarrow \begin{bmatrix} \times & \otimes & \otimes & \otimes & \otimes \\ & \times & \times & \times & \otimes \\ & \times & \times & \times & \otimes \\ & \times & \times & \times & \otimes \\ & & & & \times \end{bmatrix} = G^{-1}Y^{-1}AY^{-T},$$

where

$$G = \begin{bmatrix} 0 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ -1 & \cdots & 0 \end{bmatrix}.$$

The symbol $\otimes$ denotes a nonsymmetric nonzero, i.e., an a_{ij} that might be nonzero and might be different from a_{ji}. Here, when a_{ij} is denoted by a $\otimes$, $a_{ji} = 0$, so it is left blank. When both a_{ij} and a_{ji} might be nonzeros and different from each other, we denote one by $\otimes$ and the other by $\oplus$ to emphasize the lack of symmetry, as in the next equation. If $\left[Y^{-1}AY^{-T}\right]_{11} \neq 0$, then we get another form, which turns out to be simpler,

$$\begin{bmatrix} \times & & & & \times \\ & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{bmatrix} \longrightarrow \begin{bmatrix} \times & \otimes & \otimes & \otimes & \otimes \\ & \times & \times & \times & \otimes \\ & \times & \times & \times & \otimes \\ & \times & \times & \times & \otimes \\ & \oplus & \oplus & \oplus & \times \end{bmatrix} = G^{-1}Y^{-1}AY^{-T}.$$

This is a simpler case than the previous one, because the last row and the last column, while not a transpose of each other, are symmetric up to a scaling factor (and up to the first element, which is zero in the row but not zero in the column). In either case, we now eliminate the offdiagonals in the first row, which have just filled. We use an ordinary sequence of column operations (a Gauss transform applied from the right),

$$\begin{bmatrix} \times & \otimes & \otimes & \otimes & \otimes \\ & \times & \times & \times & \otimes \\ & \times & \times & \times & \otimes \\ & \times & \times & \times & \otimes \\ & \oplus & \oplus & \oplus & \times \end{bmatrix} \longrightarrow \begin{bmatrix} \times & & & & \\ & \times & \times & \times & \otimes \\ & \times & \times & \times & \otimes \\ & \times & \times & \times & \otimes \\ & \oplus & \oplus & \oplus & \times \end{bmatrix} = G^{-1}Y^{-1}AY^{-T}U^{-1}$$

(the offdiagonals in the last row might be zero). If $\left[Y^{-1}AY^{-T}\right]_{11} \neq 0$, then the last row and column are symmetric up to a scaling, so we scale the reduced matrix back to symmetry,

$$\begin{bmatrix} \times & & & & \\ & \times & \times & \times & \otimes \\ & \times & \times & \times & \otimes \\ & \times & \times & \times & \otimes \\ & \oplus & \oplus & \oplus & \times \end{bmatrix} \longrightarrow \begin{bmatrix} \times & & & & \\ & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & \times & \times & \times & \times \end{bmatrix} = S^{-1}G^{-1}Y^{-1}AY^{-T}U^{-1}.$$

If $\left[Y^{-1}AY^{-T}\right]_{11}$ was zero, then scaling does not symmetrize the matrix, so instead of scaling we switch to an elimination step of the third kind. To achieve numerical stability we also switch to an elimination of the third kind if $\left[Y^{-1}AY^{-T}\right]_{11}$ was not zero, but it was so small that after the multiplication by G^{-1}, the (n, n) element was larger than all the other elements in the last row.

Elimination steps of the third kind. When we switch from an elimination of the second kind to an elimination of the third kind, the reduced matrix has the form

$$\begin{bmatrix} \times & & & & \\ & \times & \times & \times & \otimes \\ & \times & \times & \times & \otimes \\ & \times & \times & \times & \otimes \\ & \oplus & \oplus & \oplus & \times \end{bmatrix} = G^{-1}Y^{-1}AY^{-T}U^{-1},$$

where the offdiagonals in the last row might be either all zeros (this is the only case in exact arithmetic), or else they are identical up to a scaling factor to the last column. We treat both cases, although some of the transformations in the identically-zero case can be skipped. We begin by permuting the last row and column to be the second; this is a bandwidth issue, and if the matrix is not banded, we could continue with the last row and column in place. We now have

$$\begin{bmatrix} \times & & & & \\ & \times & \times & \times & \otimes \\ & \times & \times & \times & \otimes \\ & \times & \times & \times & \otimes \\ & \oplus & \oplus & \oplus & \times \end{bmatrix} \longrightarrow \begin{bmatrix} \times & & & & \\ & \times & \oplus & \oplus & \oplus \\ & \otimes & \times & \times & \times \\ & \otimes & \times & \times & \times \\ & \otimes & \times & \times & \times \end{bmatrix} = P^T G^{-1}Y^{-1}AY^{-T}U^{-1}P.$$

We now use a sequence of symmetric Gauss or Givens transforms to eliminate all but the last element in the second row and column. Because the second row and column are scaled copies of each other, the same transformations applied to both the rows and the columns eliminate the nonzeros in both,

$$\begin{bmatrix} \times & & & & \\ & \times & \oplus & \oplus & \oplus \\ & \otimes & \times & \times & \times \\ & \otimes & \times & \times & \times \\ & \otimes & \times & \times & \times \end{bmatrix} \longrightarrow \begin{bmatrix} \times & & & & \\ & \times & & & \oplus \\ & & \times & \times & \times \\ & & \times & \times & \times \\ & \otimes & \times & \times & \times \end{bmatrix}$$

$$= X^{-1}P^T G^{-1}Y^{-1}AY^{-T}U^{-1}PX^{-T}.$$

The last step is to eliminate the offdiagonal nonzero in the second row and column. We eliminate them using two elementary unsymmetric Gauss transforms. We show later that this is always possible and stable,

$$\begin{bmatrix} \times & & & & \\ & \times & & & \oplus \\ & & \times & \times & \times \\ & & \times & \times & \times \\ & \otimes & \times & \times & \times \end{bmatrix} \longrightarrow \begin{bmatrix} \times & & & & \\ & \times & & & \\ & & \times & \times & \times \\ & & \times & \times & \times \\ & & \times & \times & \times \end{bmatrix}$$

$$= K^{-1}X^{-1}P^T G^{-1}Y^{-1}AY^{-T}U^{-1}PX^{-T}\hat{K}^{-T}.$$

This concludes the overview of our new elimination method.

3.2. Specification of the transformations. We now specify exactly each one of the transformations that are involved in the new elimination method. More specifically, we show how to construct the matrices L, Y, G, U, S, P, X, K, and $\hat{K}$. Some of them can be constructed in many different ways; we present the different options, but focus on the ones that guarantee stability and maintain the bandwidth of the matrix.

The matrix L. When a_{11} is large in absolute value relative to the rest of the first column, the column can be eliminated using a conventional Gauss transform

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{bmatrix}$$
$$= \begin{bmatrix} 1 & & & \\ l_{2,1} & 1 & & \\ \vdots & & \ddots & \\ l_{n,1} & & & 1 \end{bmatrix} \begin{bmatrix} a_{1,1} & 0 & \cdots & 0 \\ 0 & a^{(1)}_{2,2} & \cdots & a^{(1)}_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a^{(1)}_{n,2} & \cdots & a^{(1)}_{n,n} \end{bmatrix} \begin{bmatrix} 1 & l_{2,1} & \cdots & l_{n,1} \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix},$$

where $l_{i,1} = a_{i,1}/a_{1,1}$. If all the elements in the first row and column are zeros, then we simply set L to be the identity matrix.

The matrix Y. The task of Y is to zero rows $2, \ldots, n-1$ in the first column of A, even when a_{11} is zero or small. There are many choices for Y. For example, we could define Y to be a product of a permutation matrix that exchanges row n with the row whose first element is largest in absolute value, and a Gauss transform that uses the last row to zero the first element in the other rows. For example, if the largest element in the first column is in row 2,

$$Y^{-1} = \begin{bmatrix} 1 & & & & & 0 \\ & 1 & & & & y_{n,2} \\ & & 1 & & & y_{n,3} \\ & & & 1 & & \vdots \\ & & & & 1 & y_{n,n-1} \\ & & & & & 1 \end{bmatrix} \begin{bmatrix} 1 & & & & & \\ & 0 & & & & 1 \\ & & 1 & & & \\ & & & \ddots & & \\ & & & & 1 & \\ & 1 & & & & 0 \end{bmatrix}.$$

However, this would ruin the banded structure, so we resort to a slightly more complex transformation. Our strategy is to annihilate the nonzeros in the first column sequentially from top to bottom, and, in particular, to annihilate nonzero in position $(i, 1)$ using a row operation involving only rows i and $i+1$. That row operation can be either a Givens rotation or a Gauss transform possibly preceded by a row exchange, to ensure that the first element in row $i+1$ is larger or equal in absolute value to the first element in row i. Our implementation uses a Givens rotation. The matrix Y has the following structure:

$$Y^{-1} = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & 1 & & & \\ & & & & y^{(n-1)}_{n-1,n-1} & y^{(n-1)}_{n-1,n} \\ & & & & y^{(n-1)}_{n,n-1} & y^{(n-1)}_{n,n} \end{bmatrix} \cdots \begin{bmatrix} 1 & & & & & \\ & y^{(2)}_{2,2} & y^{(2)}_{2,3} & & & \\ & y^{(2)}_{3,2} & y^{(2)}_{3,3} & & & \\ & & & 1 & & \\ & & & & \ddots & \\ & & & & & 1 \end{bmatrix}.$$

Each 2-by-2 block is either a Givens rotation, or (if we use Gauss transforms) a product of a row exchange, perhaps identity, and a row operation on one row. After the application of Y^{-1} and Y^{-T} to A, the $(n, 1)$ element of the reduced matrix cannot

be zero: if it is, then all the subdiagonal elements of A in the first column were zeros, and we would have eliminated the first column using an elimination step of the first kind.

The matrix G. The role of G is to annihilate the $(n,1)$ element in the reduced matrix. If we are using G, then that element is larger than the $(1,1)$ element, otherwise we would have used an elimination step of the first kind. Therefore, there are essentially two options for G^{-1}: an exchange of rows 1 and n, followed by a row operation that reduces row n using row 1, or a Givens transform on rows 1 and n. We always use a Givens transform, but the other choice is valid as well.

The matrix U. After the application of G^{-1}, the $(1,1)$ element of the reduced matrix is always nonzero, although it is not necessarily large relative to the rest of row 1. We use a matrix U^{-1}, an elementary Gauss transform applied from the right to annihilate elements $(1,2),(1,3),\ldots,(1,n)$ using column operations with the first column. Because only the first element in the first column is nonzero, the trailing submatrix is not modified. This will prove to be important when we analyze the numerical stability of the algorithm, since U is potentially ill-conditioned. Formally,

$$U^{-1} = \begin{bmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,n} \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}.$$

The matrix S. The job of S is simply to scale the last row so that it becomes identical to the last column, so it is a diagonal matrix with 1s on the diagonal, except for the last element, which is nonzero but different from 1.

The matrix P. If an appropriate S does not exist or if it would be too ill-conditioned, the algorithm switches to an elimination step of the third kind. The matrix P now moves the last row and column to the second position, so that they can be eliminated. If there are no band issues, we could simply use a single row and column exchange. But to maintain the band, we do not exchange rows 2 and n. Instead, we use a cyclic permutation that puts row n in row 2 and shifts rows $2,\ldots,n-1$ one row down each,

$$P^{-1} = \begin{bmatrix} 1 & & & & \\ & 0 & & & 1 \\ & 1 & 0 & & \\ & & \ddots & \ddots & \\ & & & 1 & 0 \end{bmatrix}.$$

The matrix X. The matrix X has the same structure as Y, except that it operates on the second column, not the first. Although row 2 and column 2 in the reduced matrix are not transposes of each other, they are transposes up to a scaling except for the diagonal element. Since X and X^T do not use the diagonal element, they reduce the matrix to the desired form when applied symmetrically.

The matrices K and $\hat{K}$. These two matrices use the $(2,2)$ element in the reduced matrix to annihilate the $(n,2)$ and $(2,n)$ elements in the reduced matrix. These elements are different, so we use two different Gauss transforms from the left and the right.

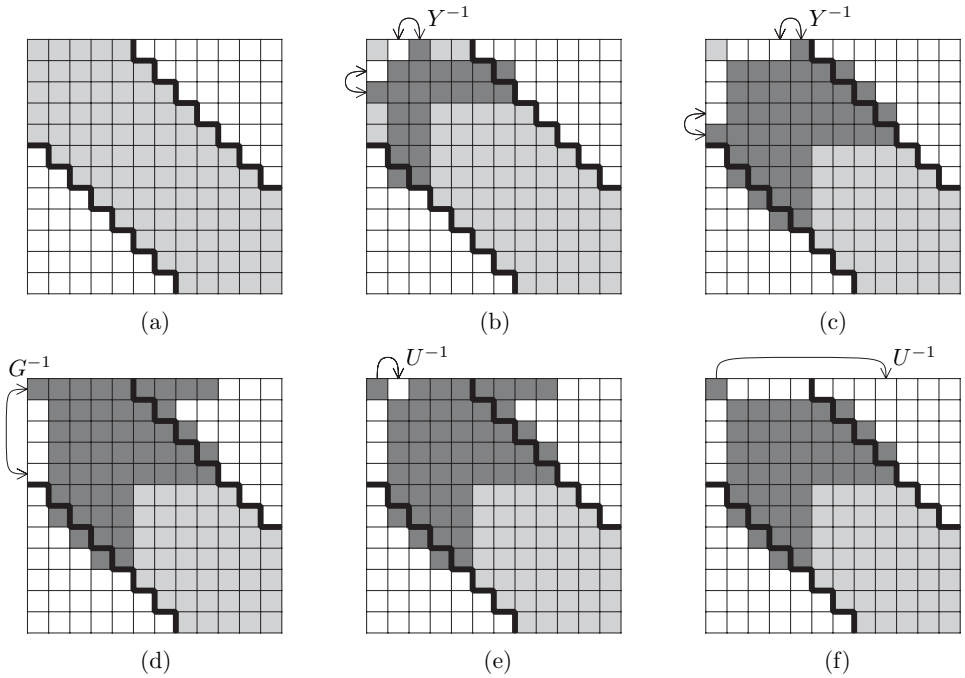


FIG. 1. *An elimination step of the second kind.*

4. Applying the new method to banded matrices. When A is banded, we need to modify the elimination algorithm to avoid increasing the band too much. To simplify the description of the algorithm, we assume that no exact cancellation occurs; but the algorithm itself copes easily with exact cancellations. It turns out that with a careful selection of transformations, the bandwidth of the reduced matrix grows, but not by much. We denote the half-bandwidth of A by m, so all but $2m+1$ of A's diagonals are zero. An elimination step of the first kind does not fill the matrix at all, so it does not require any special attention, except to ensure that no computation on zeros occurs. Elimination steps of the second and third kinds require more care.

Figures 1 and 2 illustrate elimination steps of the second and third kinds. They depict nonzero elements in gray; light gray signifies that an element has not been modified in this elimination step, and dark gray signifies that an element has been modified. The original profile of the matrix is indicated by a heavy black border. The arrows show which column/row are scaled and subtracted from which other column/row to annihilate a nonzero. The letters indicate the transformation matrix that performs the elimination.

When an elimination step of the second kind is applied to a banded matrix, we must construct transformation Y in a way that A does not fill too much. Specifically, we eliminate element i in the first row and column using columns/rows i and $i+1$, but we stop if elements $i+2,\ldots,n$ are all zeros. We do not "roll" the last nonzero in the row/column down to the end of the row/column. This is illustrated in Figure 1 (top row). The rest of the step proceeds without modification. As shown in the figure, an elimination step of the second kind can increase the half-bandwidth of the matrix, but only by one, and only in rows and columns $2,\ldots,m'$, where $m'+1$ is the number of nonzeros in row/column 1 (m' might be larger than m if the bandwidth increased

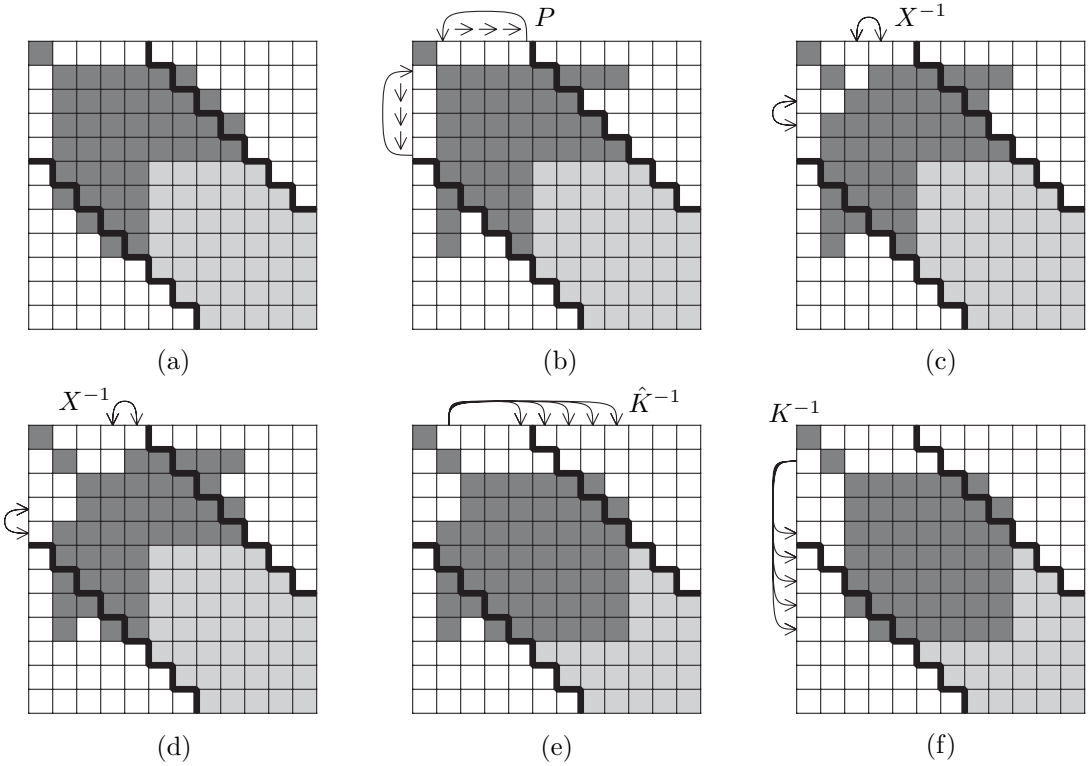


FIG. 2. *An elimination step of the third kind. The top left illustration describes the matrix after all the operations of a second-kind elimination step have been applied.*

in previous elimination steps).

Applying an elimination step of the third kind to a banded matrix is more involved. After the permutation P has been applied, the second row and column are eliminated. The elimination of these nonzeros can destroy the band structure if done carelessly. We can eliminate all but one of the nonzeros using a series X of Givens transforms, as we have done (using Y) in the beginning of the second-kind elimination. We can also use column operations to annihilate the second row using its diagonal element as a pivot, and then row operations to annihilate the second column; these are the transformations K and $\hat{K}$ that we described above. In the dense case, we use K and $\hat{K}$ only to annihilate a single nonzero in the row and a single nonzero in the column. But in the banded case, we need to restrict X to the annihilation of just $m' - 2$ nonzeros near the diagonal, and use K and $\hat{K}$ to annihilate the rest. As we shall see later, this does not introduce growth in the reduced matrix, since the diagonal element in row 2 is larger than the offdiagonal nonzeros.

The strategy that we use is shown in Figure 2. It is easy to see that if we use Givens transforms to annihilate all but the last element in the second row and column, the band will grow by one in rows/columns higher than m', which we try to avoid. On the other hand, if we use only diagonal Gauss transforms, an entire bulge outside the band would fill, which we also seek to avoid. Therefore, we use Givens transforms, which only increase the band locally by one, to annihilate $m' - 2$ nonzeros in the first row and column, and Gauss transforms, which introduce no fill at all, to annihilate

the rest.

4.1. Bound on half-bandwidth. We now prove that this strategy ensures that the half-bandwidth of the reduced matrix is bounded by $2m-1$. We do this in two steps. We first define a variant of our algorithm, which we call ALG2, and show that the fill in our algorithm is a subset of the fill created by ALG2. This variant always uses reductions of the second type; it may be unstable, so it is not useful in practice. We only use it to bound the bandwidth of the reduced matrix.

In the context of this section we ignore any numerical considerations. Our analysis deals only with the structural aspects of our algorithm (and of the variant ALG2). In particular, we assume that A has no zeros inside its band and we ignore zeros that may appear during the reduction process.

We start with a few definitions. A reduction step in our algorithm starts with a symmetric banded matrix $A^{(k)}$, where k denotes the size of the matrix (in the first step, $A^{(n)} = A$), and outputs a reduced symmetric matrix $A^{(k-s)}$ of size $(k-s)$-by-$(k-s)$. The number s of eliminated rows and columns may be 1 or 2. This notation is borrowed from [4]. We denote by $a_{ij}^{(k)}$ both the ij element of $A^{(k)}$ at the start and during the reduction step. The [illegible] of a banded matrix B in column (row) j, defined as the minimum $r' \geq 0$ for which $b_{ij} = 0$ for all $i > j + r'$, is denoted by r_B^j. In the case of the input matrix A, all the columns (rows) have the same local half bandwidth m.

We now define the variant algorithm, ALG2. It is similar to our actual algorithm, which we denote by ALG1, except that it only uses reduction steps of the second type, and that it eliminates the offdiagonal nonzeros in the first column only using Givens rotations. (ALG1 uses Gauss transforms, perhaps with row exchanges.) ALG2 eliminates one row and column in each reduction step, so it uses exactly $n-1$ reduction steps. We denote by $\widetilde{A}^{(k)}$ the reduced matrix we get after applying the first $n-k$ reduction steps of ALG2 on A, where $0 < k \leq n$. Note that $\widetilde{A}^{(k)}$ is k-by-k.

The following theorem states the main result of this section.

THEOREM 4.1. [illegible] (ALG1) [illegible] $r^i_{A^{(k)}} < 2m$ [illegible] $0 < k \leq n$ $1 \leq i \leq k$

This theorem is an immediate corollary of Lemmas 4.3 and 4.4, which we state and prove below. But we first state and prove a technical lemma.

LEMMA 4.2. [illegible] ALG2, $r^{i+1}_{\widetilde{A}^{(k)}} \leq r^i_{\widetilde{A}^{(k)}} \leq r^{i+1}_{\widetilde{A}^{(k)}} + 1$ [illegible] $1 < k \leq n$ [illegible] $1 \leq i < k$ [illegible]. We prove the lemma by induction on k. The lemma holds for $A = \widetilde{A}^{(n)}$. We assume that the lemma holds for $\widetilde{A}^{(k)}$. During the $n-k+1$th reduction step, each of the columns $2, \ldots, r^1_{\widetilde{A}^{(k)}}$ gets the structure of the column to its right. This is because

$$r^i_{\widetilde{A}^{(k)}} \leq r^{i+1}_{\widetilde{A}^{(k)}} + 1$$

by the induction assumption. Each of the rows $2, \ldots, r^1_{\widetilde{A}^{(k)}}$ behaves in a symmetric way. For j such that $1 \leq j < r^1_{\widetilde{A}^{(k)}} - 1$, we have (see Figure 3)

$$r^j_{\widetilde{A}^{(k-1)}} = r^{j+2}_{\widetilde{A}^{(k)}} + 1$$

and

$$r^{j+1}_{\widetilde{A}^{(k-1)}} = r^{j+3}_{\widetilde{A}^{(k)}} + 1.$$

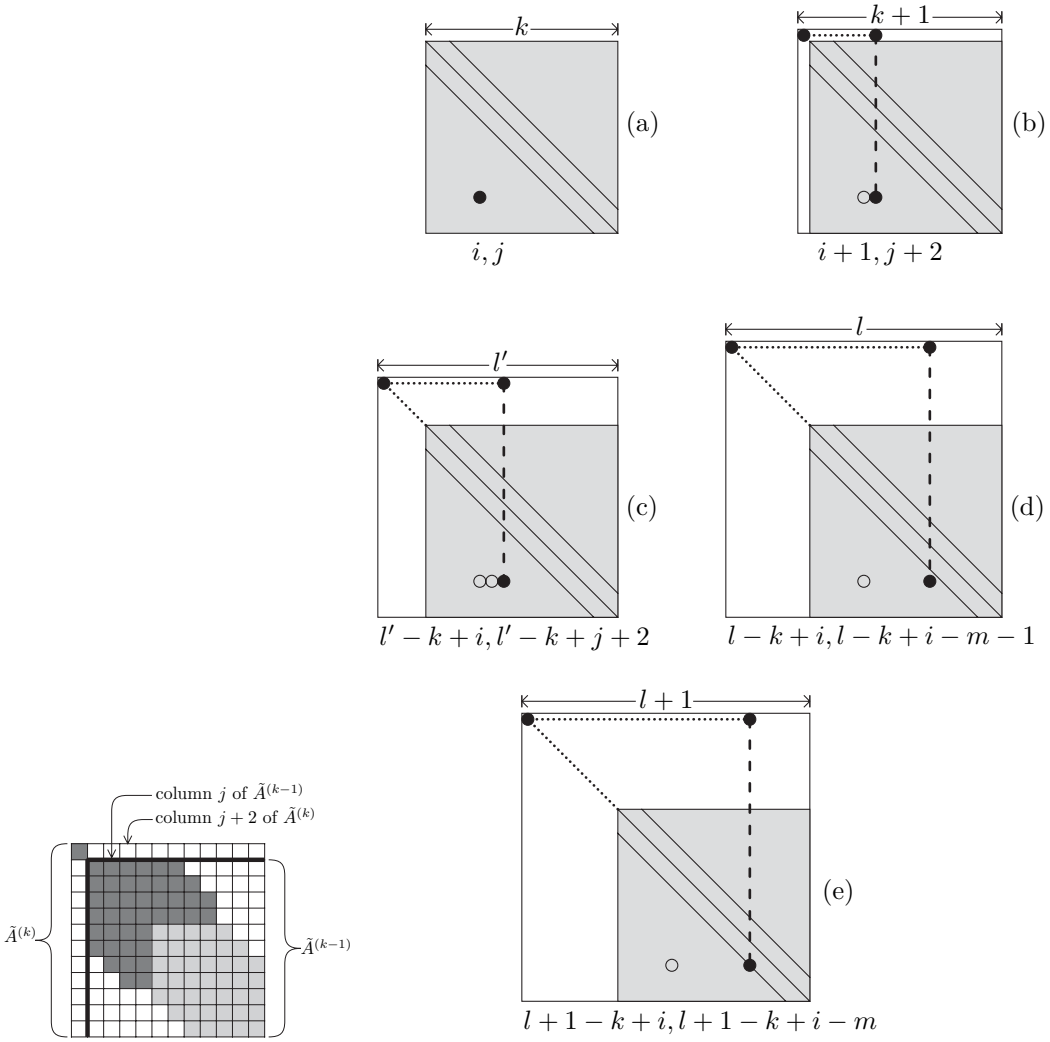


FIG. 3. *Left: An illustration of the proof of Lemma* 4.2. *Right: An illustration of the main steps in the proof of Lemma* 4.4. *The submatrix depicted in gray is always k-by-k. The row and column indices of the black element are shown below each matrix (the black element within the gray submatrix).*

From this and the induction assumption,

$$r^{j+1}_{\tilde{A}^{(k-1)}} \le r^{j}_{\tilde{A}^{(k-1)}} \le r^{j+1}_{\tilde{A}^{(k-1)}} + 1,$$

where $1 \le j < r^{1}_{\tilde{A}^{(k)}} - 1$. We also have

$$r^{j}_{\tilde{A}^{(k-1)}} = r^{j+2}_{\tilde{A}^{(k)}} + 1 = r^{j+1}_{\tilde{A}^{(k-1)}} + 1$$

for $j = r^{1}_{\tilde{A}^{(k)}} - 1$, which means that

$$r^{j+1}_{\tilde{A}^{(k-1)}} \le r^{j}_{\tilde{A}^{(k-1)}} \le r^{j+1}_{\tilde{A}^{(k-1)}} + 1$$

is true also for such j. For $r^1_{\widetilde{A}^{(k)}} - 1 < j < k - 1$ the lemma is immediate from the induction assumption. □

LEMMA 4.3. η_M M $A^{(k)}$ 1 A $\widetilde{A}^{(k)}$ 2 A

$$\eta_{A^{(k)}} \subseteq \eta_{\widetilde{A}^{(k)}}.$$

. Let $A^{(k)}$ be the reduced matrix we get from the lth reduction step of ALG1, when applied to A.

The proof is immediate for $l = 0$ (and so $k = n$) when no reduction step has been applied to A yet.

Otherwise, the output of the previous reduction step is either $A^{(k+1)}$ or $A^{(k+2)}$. If the output of reduction step $l - 1$ is $A^{(k+1)}$, then by the induction assumption we have

$$\eta_{A^{(k+1)}} \subseteq \eta_{\widetilde{A}^{(k+1)}}.$$

Moreover, it is guaranteed that either Gaussian elimination or a reduction of the second kind is applied to $A^{(k+1)}$ in order to get $A^{(k)}$. Due to the facts that the reduction applied on $\widetilde{A}^{(k+1)}$ by ALG2 is of the second kind and that the Gaussian elimination, if applied by ALG1 to $A^{(k+1)}$, does not increase the band, we have

$$\eta_{A^{(k)}} \subseteq \eta_{\widetilde{A}^{(k)}}.$$

If the output of the $l-1$th reduction step is $\mathrm{A}^{(k+2)}$, then a reduction of the third kind is applied to $A^{(k+2)}$ in order to get $A^{(k)}$. The operations whose effect on the structure of $A^{(k)}$ we need to consider are applying Y^{-1}, bringing column and row $r^{(1)}_{A^{k+2}} + 1$ to the second position, applying X^{-1}, and finally applying a nonsymmetric Gauss reduction. Let B be the $k + 1 \times k + 1$ right bottom block of the matrix we get after applying Y^{-1} to $A^{(k+2)}$. We have

$$\eta_B \subseteq \eta_{\widetilde{A}^{(k+1)}} \tag{4.1}$$

due to similar arguments to those by which we have shown that $\eta_{A^{(k)}} \subseteq \eta_{\widetilde{A}^{(k)}}$, in case the output of the $l - 1$th reduction step is $\mathrm{A}^{(k+1)}$. Let C be the matrix we get after bringing the $r^{(1)}_{A^{(k+2)}}$ column and row of B to the first column and row, respectively. During this operation, each of the columns (rows) $1, \ldots, r^{(1)}_{A^{(k+2)}} - 1$ of B is moved one location to the right (down). From the combination of (4.1) with Lemma 4.2 we have that the structures of columns (rows) $2, ldots, r^{(1)}_{A^{(k+2)}}$ in C are contained in the structures of these columns (rows) in $\widetilde{A}^{(k+1)}$, respectively. We get

$$\eta_{C_{[k]}} \subseteq \eta_{\widetilde{A}^{(k+1)}_{[k]}}, \tag{4.2}$$

where $M_{[k]}$ denotes the k-by-k right bottom block of the matrix M. The next structure-relevant part of a third-kind reduction is X^{-1}. The last element elimination performed during applying X^{-1} on C in the first column is done by the $r^1_{A^{(k+2)}}$th row. The last element elimination performed during the Y^{-1} part in the ALG2 reduction step of $\widetilde{A}^{(k+1)}$ in the first column is done by the $r^1_{\widetilde{A}^{(k+1)}} + 1$th row. For these two indices, we have

$$r^1_{A^{(k+2)}} \le r^1_{\widetilde{A}^{(k+2)}} \le r^2_{\widetilde{A}^{(k+2)}} + 1 \le r^1_{\widetilde{A}^{(k+1)}} + 1 \tag{4.3}$$

by the induction assumption, Lemma 4.2, and the trivial fact $r^2_{\widetilde{A}^{(k+2)}} \leq r^1_{\widetilde{A}^{(k+1)}}$. Denote the matrix we get by applying X^{-1} on C by D. By combining (4.3) with (4.2) we get that the structure of $D_{[k]}$ is contained in the structure of $\widetilde{A}^{(k)}$. The last operation performed during a reduction of the third kind is a nonsymmetric Gauss reduction. The effect on the fill that this operation may have on D is limited to elements d_{ij}, where

$$r^1_{A^{(k+2)}} \leq i, j \leq r^1_C + 1.$$

The fact that $\widetilde{A}^{(k+1)}$ has a nonzero element in entry $(r^1_C + 1, r^1_{A^{(k+2)}})$ (because B has a nonzero there, and due to (4.1)) and Lemma 4.2 bring us to the conclusion that the elements of $\widetilde{A}^{(k)}$ in entries i, j, where $r^1_{A^{(k+2)}} - 1 \leq i, j \leq r^1_C$ are all nonzeros. We got

$$\eta_{A^{(k)}} \subseteq \eta_{\widetilde{A}^{(k)}}. \qquad \square$$

We now prove the following lemma on ALG2.

LEMMA 4.4. [illegible] $r^i_{\widetilde{A}^{(k)}} < 2m$ [illegible] $0 < k \leq n$ $1 \leq i \leq k$

[illegible]. Let us focus in a nonzero element located at i, j such that $i - j$ is maximal over all elements that appear during the factorization. More formally, i and j are such that $i - j = max_{i',j',t}\{i' - j' : \widetilde{a}^{(t)}_{i'j'} \neq 0\}$, $i > j$. Assume, without loss of generality, that this nonzero is the first to fill the criterion and that the reduced matrix $F = \widetilde{A}^{(k)}$ is the first in which the nonzero appears (see Figure 3:A). The appearance of the nonzero in the jth column of the ith row of $\widetilde{A}^{(k)}$ implies that the $i+1, j+2$ element in $\widetilde{A}^{(k+1)}$ is nonzero (Figure 3:B). This, in turn, implies that there is some $l' > k+1$ s.t. $\widetilde{a}^{(l')}_{l'-k+i,l'-k+j+2} \neq 0$ (Figure 3:C). If we continue similarly, we get to the observation that there is some $l > k+i-j-m-2$ s.t. $\widetilde{a}^{(l)}_{l-k+i,l-k+i-m-1} \neq 0$ (Figure 3:D). Assume, without loss of generality, that l is maximal, which means that $\widetilde{a}^{(l+1)}_{l+1-k+i,l+1-k+i-m-1} = 0$ (Figure 3:E). This implies that $r^1_{\widetilde{A}^{(l+1)}} \geq 2i - j - 2m - 1$. On the other hand, $r^1_{\widetilde{A}^{(l+1)}} < i - j$. We get $2i - j - 2m - 1 < i - j$, i.e., $i < 2m + 1$ and therefore $i - j < 2m$. This proves the lemma. $\square$

4.2. Storing the factors. We store the input matrix and the factors in a format that is quite similar to LAPACK's banded storage format. The code expects to receive only the upper triangle of A. The representation of the matrix A will be stored in a two-dimensional array, with n columns and with a number of rows related to the half-bandwidth m. On input, each column of A will be stored in the corresponding column of the representation. Each row of the representation will correspond to a diagonal of A, with the main diagonal given as the first row of the representation, the first superdiagonal of A as the second row of the representation, and so forth. Rows of A will correspond to diagonals of the representation, filling the first m rows of the representation. As the algorithm runs, rows of A are overwritten with elements of the factors.

The factors of A require more storage than A itself. We accomodate the factors by requiring that the row dimension of the representation array be larger than the half-bandwidth m. This storage format is similar to the input/output format of LAPACK's DGBTRF routine, which implements Gaussian elimination with partial pivoting, where reduced rows also expand.

Strictly speaking, the leading dimension of the array should be at least eight times the half-bandwidth, to allow for storage of the output. Since most matrices do not

require that much storage to factor, the code can be called with a smaller leading dimension. In particular, we have never encountered a matrix that requires more than six times the half-bandwidth, so all the runs reported later in this paper use a leading dimension that is six times the half-bandwidth.

The details of how we pack the representation of the factors into the array are not important and are omitted. Only three aspects of the data structure are important. First, the fact that rows of A are not packed, but stored with spaces in between them, allows the algorithm to reduce the remaining rows in each elimination step. That is, the algorithm is essentially a right-looking algorithm. A fully packed storage would not have allowed a right-looking algorithm. Second, the coefficients of the transformations that eliminate a row are stored in the memory that was used to store the row of A and additional memory locations that are statically preallocated for that row in the input/output array. Third, the representation of each Givens rotation is stored in one word using the scheme of [20].

We note that it is possible to reduce the storage requirements of the algorithm using dynamic storage allocation for the factors, rather than using static preallocated areas. The static allocation leaves some unused storage.

5. Numerical stability: Bounding the factors. In this section we show that the growth during the algorithm is bounded from above by 4^{n-1}, and that the factors are bounded. The growth factor ρ_A is the ratio of the largest element in absolute value in the reduced matrices to the largest element in A,

$$\rho_A = \frac{\max_{i,j,k} \left| a_{i,j}^{(k)} \right|}{\max_{i,j} |a_{i,j}|}.$$

A bound of this form, even an exponential bound like the one that we prove here, is often associated with backward stable elimination algorithms. In particular, the growth in GEPP is bounded by 2^{n-1}, the growth in Bunch and Kaufman's algorithm is bounded by 2.57^{n-1}, and the growth in Aasen's algorithm is bounded by 4^{n-2}. In practice, such growth factors are rarely encountered; the growth in practice is usually small. Researchers have shown that when the growth is small, these algorithms are backward stable [11, 12, 21, 22]. Since large growth factors are rare, these algorithms are stable in practice [9, 13]. Although we do not show that a similar implication holds for our algorithm, we believe that one does.

Moreover, the fact that a bound on the growth exists at all, even if the bound is exponential, is usually a reflection of a sound numerical design of the pivoting strategy. This observation does hold for our algorithm. The careful design of the snap-back pivoting strategy ensures that the elements of the reduced matrix can grow by at most a factor of 4 in every step.

Growth is measured on the reduced matrices, the intermediate matrices after some elimination steps have been carried out. In some algorithms, such as GEPP and Aasen, the entries in the factors are also bounded. In GEPP, for example, the magnitude of the entries of the lower triangular factor are bounded by 1 and the magnitudes in the upper triangular factor are bounded by the magnitude of elements in reduced matrices. Bounded factors are a stronger property than bounded growth. In particular, in some cases, bounded factors imply backward stability [2, Appendix B]. We show in this section that the factors in our algorithm are bounded, although this does not formally imply backward stability. We note that in the algorithms of Bunch and Kaufman, the factors are not bounded: they can have large entries even when the reduced matrices remain small (see, for example, [13, p. 219]).

Before we analyze growth, we should define formally the conditions by which we choose the next type of reduction step to perform. If possible, Gaussian elimination is applied. The condition for applying this reduction is the same as the condition that appears in the Bunch–Kaufman algorithm. This condition involves a constant α with a value between 0 and 1 and the values $\gamma_1^{(k)} = \max_{1<l\le k} \mid a_{l,1}^{(k)} \mid$ and $\gamma_t^{(k)} = \max_{1<l\le k} \mid a_{l,t}^{(k)} \mid$, where t is the index of the row of the entry with the maximum magnitude in the first column. Using this notation, we use a reduction of the first kind if one of the following conditions is satisfied:

1. $\mid a_{1,1}^{(k)} \mid > \alpha \cdot \gamma_1^{(k)}$, or
2. $\mid a_{1,1}^{(k)} \mid \cdot \gamma_t^{(k)} > \alpha(\gamma_1^{(k)})^2$ (optional; see below).

Otherwise, a reduction of one of the other two types is applied. Condition 2 is optional in the sense that if we use it, growth in the reduced matrices is still bounded, but the factors may grow; if we only use a reduction of the first type when condition 1 is satisfied, then both the reduced matrices and the factors are bounded.

The two other types of reductions start with the same sequences of elementary transformations. The decision about the chosen type is taken only after this shared sequence of transformations terminates. If the element $a_{r^1_{A^{(k)}}+1, r^1_{A^{(k)}}+1}$ is less than or equal (a threshold may be used) to all the other elements in its row, then the second type of reduction is chosen and S^{-1} is applied. Otherwise, the reduction proceeds as a third type reduction.

We now analyze the growth.

Consider first symmetric Gauss transformations. The analysis of this case is the same as in [4, 2]. For an element $a_{ij}^{(k)}$ in the input matrix $A^{(k)}$ and an element $a_{ij}^{(k-1)}$ in the reduced matrix $A^{(k-1)}$ we have

$$a_{ij}^{(k-1)} = a_{ij}^{(k)} - \frac{a_{i1}^{(k)} a_{1j}^{(k)}}{a_{11}^{(k)}}, \quad i > 1, \ \ j > 1. \tag{5.1}$$

Let μ be the magnitude of the largest entry in $A^{(k)}$ and μ' the maximum magnitude of any entry in the reduced matrix $A^{(k-1)}$. If $a_{11}^{(k)} > \alpha \cdot \max_{1<l\le k}(a_{l,1}^{(k)})$, then from (5.1) we get

$$\mu' \le \mu + \frac{\mid a_{i1}^{(k)} a_{1j}^{(k)} \mid}{\mid a_{11}^{(k)} \mid} \le \mu \left(1 + \frac{1}{\alpha}\right). \tag{5.2}$$

If $\mid a_{11}^{(k)} \mid \cdot \gamma_t^{(k)} \ge \alpha \cdot (\gamma_1^{(k)})^2$, then using (5.1)

$$\mu' \le \mu + \frac{(\gamma_1^{(k)})^2}{\mid a_{11}^{(k)} \mid} \le \mu + \frac{\gamma_t^{(k)}}{\alpha} \le \mu \left(1 + \frac{1}{\alpha}\right). \tag{5.3}$$

Now, consider the second and third kinds of transformations. In order to analyze the growth for these kinds of transformations, we look at the elementary transformations from which these two complex kinds are made up. We may divide the elementary transformations into four categories. The first category includes the transformation that annihilates an element in the first or second column (row) by its nearby element in the same column (row). In our algorithm, this transformation is applied as a part of an elimination of a full column. During such an elimination, the first annihilated

element is the one directly under the diagonal. The next elementary transformation annihilates the element located two entries under the diagonal, and so on. If an element to be annihilated is greater than the element under it, the rows of these two elements and the symmetric columns are swapped first. The second category of elementary transformations is the Gauss transformations, and the third includes the Givens transformation that annihilates one nondiagonal element by a diagonal element, both in the first column. The fourth category includes the one-row-scaling diagonal transformation S^{-1}, in which all of the diagonal is 1's, except, maybe, for one diagonal element, whose magnitude is smaller than 1.

The last three categories of transformations with the conditions for applying them trivially imply an upper bound of 2 on the growth the transformations induce on the matrix they are applied on. We now focus on the first category. We show it has an upper bound of 4 on the growth it induces. Assume that a sequence of Q elementary transformations of the first category, as described above, is applied on a matrix $B^{(0)}$, with a maximum magnitude element ν, in order to eliminate its first column. Denote the matrix we get after applying $q \le Q$ such elementary transformations by $B^{(q)}$, and its i, j element by $b_{ij}^{(q)}$.

LEMMA 5.1. A. , $i \ge q+2$, $j \ge q+2$, . , $| b_{ij}^{(q)} | \le \nu$

B. , $1 \le i < q+2$, $j \ge q+2$, . $i \ge q+2$, $1 \le j < q+2$, . $| b_{ij}^{(q)} | \le 2 \cdot \nu$

C. , $1 \le i < q+2$, $1 \le j < q+2$, . $| b_{ij}^{(q)} | \le 4 \cdot \nu$

, , . The lemma is trivially correct for $q = 0$.

Assume the lemma is correct for some $q \le Q-1$.

The only rows that may change due to the $q+1$th left elementary transformation are the $q+2$ and $q+3$ rows, and similarly, the $q+2$ and $q+3$ columns are the only columns that may change due to the $q+1$th right elementary transformation.

Elements $b_{i,q+3}^{(q+1)}$ and $b_{q+3,i}^{(q+1)}$, where $i > q+3$ are equal either to elements $b_{i,q+3}^{(q)}$ and $b_{q+3,i}^{(q)}$, where $i > q+3$ or to elements $b_{i,q+2}^{(q)}$ and $b_{q+2,i}^{(q)}$, where $i > q+3$, respectively, depends on whether or not a swap is part of the $q+1$th elementary transformation. Similarly, element $b_{q+3,q+3}^{(q+1)}$ equals either to $b_{q+3,q+3}^{(q)}$ or to $b_{q+2,q+2}^{(q)}$. By the induction assumption we have $| b_{ij}^{(q+1)} | \le \nu$ for $i \ge q+3$ and $j \ge q+3$. By that we proved A.

We also have $| b_{ij}^{(q+1)} | \le 2\nu$ for $i \ge q+3$ and $j = q+2$ and for $i = q+2$ and $j \ge q+3$, because each of these elements is computed as a sum of two elements in $B^{(q)}$, which are at most ν by the induction assumption. In addition, from the induction assumption, $| b_{ij}^{(q+1)} | \le 2\nu$ for $i \ge q+3$ and $1 \le j < q+2$ and for $1 \le i < q+2$ and $j \ge q+3$. This completes the proof of B.

In order to prove C it is enough to show that $| b_{ij}^{(q+1)} | \le 4 \cdot \nu$ for $i = q+2$ and $1 < j < q+3$ (and for $1 < i < q+3$ and $j = q+2$). When $i \ne j$ this bound holds because each of the elements is computed as the sum of two elements in $B^{(q)}$, which are at most 2ν by the induction assumption. If $i = j = q+2$ then $b_{i,j}^{(q+1)}$, is the sum of the four elements $b_{i,j}^{(q)}$, $b_{i+1,j}^{(q)}$, $b_{i,j+1}^{(q)}$, and $b_{i+1,j+1}^{(q)}$. By the induction assumption, each of these elements is at most ν. This completes the proof. □

A similar lemma holds for the case where the second column and row are eliminated using a sequence of transformations of the first category. Such a case may be interpreted simply as eliminating the first row and column of a matrix C, which is $B^{(0)}$ without its first row and column and with some permutation on its columns/rows.

This extraction of the lemma bounds the growth induced by the second sequence of transformations of the first category applied during the third type of reduction.

From the above upper bounds on growth during the first type of reduction and on growth induced by the four categories of elementary transformations, we can now conclude an upper bound on the growth during the full reduction step. In case the first type of elimination is chosen, the growth is bounded by $1 + \frac{1}{\alpha}$. In case the second or third type is applied, we have from Lemma 5.1 that the sequences of the first category of elementary transformations (Y^{-1}or X^{-1}) causes each a growth bounded by 4. The other elementary transformations may cause growth only to elements that didn't grow above the old maximum element in the matrix during the last applied Y^{-1} (or X^{-1}) transformation. It means that at the end of a second type reduction step, the growth is bounded by 4. It means also that this is the case just before applying X^{-1}. The growth caused by the X^{-1} transformation itself and the elementary transformation that follows, is again bounded by 4 .This brings us to total growth bounded by 16 for the third type of reduction.

We now can use a similar approach to the one used by the Bunch–Kaufman algorithm to determine the value α. We have growth bounded by 16 for a double column elimination done during the third type of reduction, and bounded by 4 for a single column elimination done during the second type. It means that the elimination of a column during the second or third reduction types induces a growth of 4. In order to minimize the bound that we have on a general reduction step, we need to find α such that

$$\mu\left(1 + \frac{1}{\alpha}\right) \leq 4\mu.$$

Therefore, we need to ensure that $\alpha \geq 1/3$; by setting $\alpha = 1/3$, we maximize the opportunity to use elimination steps of the first type, which is the cheapest type.

In total we have an upper bound of 4^{n-1} on the growth factor during the factorization.

We now prove a bound on the entries of the factors.

LEMMA 5.2. $| a_{1,1}^{(k)} | > \alpha \cdot \gamma_1^{(k)}$ $0 < \alpha \leq 1$ α^{-1} . We show that elements of the nine matrices L, Y, G, U, S, P, X, K, and $\hat{K}$ are all bounded. L is a Gauss transform in which the pivot is smaller than the elements in its row and column by at most a factor of α, so the elements of L are bounded by α^{-1}. K is a similar transform, except that the pivot is larger or equal to the rest of its row. Therefore, the elements of K are bounded by $1 \leq \alpha^{-1}$. G represents a single Givens rotation, so its entries are bounded by 1. Y and X represent either a sequence of Givens rotations, or a sequence of offdiagonal Gauss transforms in which the pivot is larger than the element that it annihilates. If we use Givens rotations, Y and X are orthonormal so their entries are bounded by 1. If we use Gauss transforms, Y and X are row permutations of unit lower-triangular matrices with subdiagonal elements bounded by 1. The matrix U represents a series of column operations that annihilate a row in a reduced matrix. Thus, U is an upper triangular matrix, with one row that contains a copy of the annihilated row in the reduced matrix, and with a unit diagonal elsewhere. Therefore, the elements of U are

elements of a reduced matrix and 1's, so they are bounded. The matrix $\hat{K}$ is similar, except that it represents row operations. P is a permutation matrix, so its entries are bounded by 1. S is a transformation that scales a single row [illegible] (S^{-1} scales a row up), so it is a diagonal matrix, with diagonal entries that are 1 except for one entry, which is smaller than 1. □

The qualification that the factors are bounded with respect to partially-eliminated reduced matrices, not with respect to reduced matrices after the full elimination of some rows and columns, is not a significant one. Our previous analysis of the growth in the reduced matrices covers partially-eliminated reduced matrices as well.

6. Implementation and results. This section describes our implementation of the algorithm and presents results of experiments that investigate the behavior and performance of the algorithm.

6.1. Implementation and benchmark codes. We have implemented the algorithm in the C language[2] using LAPACK-style interfaces. The implementation includes two externally-visible routines, one to factor a symmetric banded matrix and another to solve a linear system using a previously-computed factorization.

The threshold value α that our implementation uses is $\alpha = 1/3$, the value that minimizes worst-case element growth. The implementation uses Givens rotations in the Y and X factors.

We tested this implementation against three other codes: LAPACK's banded LU with partial pivoting (DGBTRF), LAPACK's banded LU with partial pivoting but without blocking (DGBTF2; this is an internal LAPACK routine), and the symmetric band reduction, an orthogonal factorization code for banded symmetric matrices [3]. Our code is not blocked. That is, it does not partition the matrix into blocks to achieve high cache efficiency. Thus, from the algorithmic point of view, it is most appropriate to compare it to other nonblocked factorization codes, such as DGBTF2. From the practical point of view, it is also important to know how our implementation compares to the best existing factorization code, which is DGBTRF. (This comparison, however, does not reveal much about our algorithm, since it is difficult to separate the algorithmic and cache issues that influence the performance of DGBTRF.)

6.2. Test environment. We conducted the experiments on two different machines. Some of the experiments were conducted on a 3.2 GHz Pentium 4 computer running Linux with 2 GB of main memory. We compiled our code using GCC version 4.0.0. The version of LAPACK that we used was compiled using GCC 2.91 from C sources that were generated from the Fortran sources using `f2c`. We linked our code as well as LAPACK with an implementation of the BLAS by Kazushige Goto, version 0.99 for Pentium 4 (Coppermine). We measured this LAPACK/BLAS combination against Intel's Math Kernel Library (MKL) version 7.2.1 and found that the performance of the banded GEPP subroutines in the two implementations was similar. Therefore, the performance that we report does not appear to be affected by the use of the Fortran-to-C translation or of GCC. We used this LAPACK/BLAS combination because it delivered better performance for our algorithm than MKL.

Other experiments were performed on a dual 3 GHz AMD Opteron machine with 8 GB of main memory. On this machine we used GCC 3.4.2 and Goto's BLAS version 0.97 for 64-bit Opterons (we do not report detailed timing data on this machine, only accuracy data).

[2]The code is available from http://www.tau.ac.il/∼stoledo/research.html.

TABLE 1
Test matrices from the Gould–Scott study.

#	name	n	m	#	name	n	m	#	name	n	m
1	linverse	11999	4	12	crystk02	13965	821	23	bcsstk39	46772	817
2	spmsrtls	29995	4	13	mario001	38434	356	24	dawson5	51537	867
3	dtoc	24993	5	14	aug3d	24300	759	25	bcsstk35	30237	1764
4	dixmaanl	60000	7	15	crystk03	24696	1034	26	vibrobox	12328	4535
5	sit100	10262	396	16	bratu3d	27792	945	27	helm3d01	32226	2593
6	tuma2	12992	322	17	cont-201	80595	403	28	ncvxbqp1	50000	1682
7	stokes64	12546	384	18	ncvxqp1	12111	2692	29	k1_san	67759	1574
8	stokes64s	12546	384	19	aug3dcqp	35543	994	30	olesnik0	88263	1214
9	aug2d	29008	198	20	bcsstk37	25503	1427	31	cont-300	180895	603
10	aug2dc	30200	202	21	ncvxqp9	16554	2216	32	copter2	55476	2304
11	tuma1	22967	483	22	stokes128	49666	768	33	qa8fk	66127	2016

TABLE 2
Test matrices from John Betts.

#	name	n	m	#	name	n	m	#	name	n	m
1	traj02	1665	457	4	traj15a	1999	1882	7	traj15d	1999	1882
2	traj06a	1665	466	5	traj15b	1999	1882				
3	traj06b	1665	466	6	traj15c	1999	1882				

We used two different machines because we had the MKL performance of GEPP as a baseline only on Intel Pentium 4 machines, but 32-bit Pentium 4 machines could not factor some of the larger test matrices. The 64-bit Opteron allowed us to factor large matrices.

6.3. Reliability, stability, and accuracy. We performed several experiments in order to assess the reliability, stability, and accuracy of the algorithm. These were all conducted on the Opteron machine. We used three families of test matrices for these experiments. One family consists of the 61 matrices that Gould and Scott used to evaluate sparse symmetric indefinite direct solvers [10]. We reordered the rows and columns using reverse Cuthill–McKee ordering, to reduce their bandwidth. Of the 61 matrices, our algorithm ran out of memory on 28 on a machine with 8 GB of memory, leaving for our experiment the 33 matrices listed in Table 1. Another family consisted of 7 matrices that were collected by Roger Grimes from John Betts of Boeing, listed in Table 2. They are KKT matrices extracted out of a sparse nonlinear optimization package called SOCS. These problems were among those that led to the discovery that the Bunch–Kaufman algorithm can create large entries in the lower-triangular factor. Solving linear systems with these factors resulted in very-low-accuracy solutions, which led to convergence failures in the nonlinear optimization algorithm.

Figure 4 shows the growth and the norm of the residual on the Gould–Scott matrices. These runs were performed with $\alpha = 1/3$. The right-hand-side b was generated by multiplying A by a random solution vector x. The figure shows the growth in the reduced matrices in our algorithm and the residual in both our algorithm and LAPACK's GEPP. Our code failed on one matrix due to overflows in the solve phase (GEPP overflowed on 4), and produced an unacceptable residual on only one matrix. That matrix suffered from large growth. The data suggests that very large growth leads to backward instability in the algorithm. The data also suggests that problems in the solve phase can occur even when there is no growth. On the other

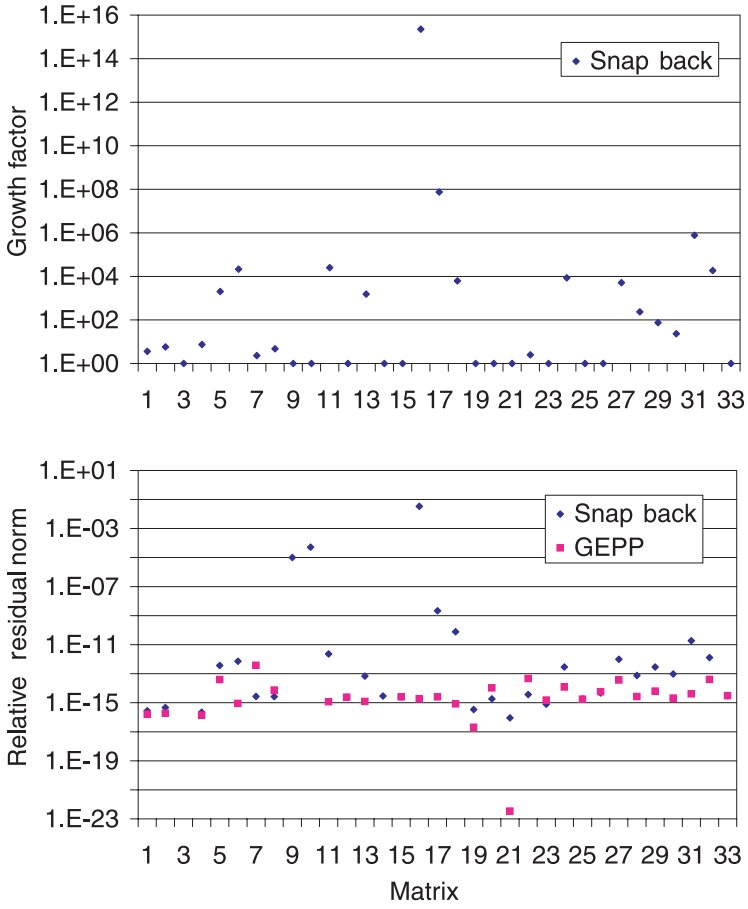


FIG. 4. *Growth in the reduced matrices in our algorithm and the residual in both our algorithm and GEPP on the* 33 *matrices from the Gould–Scott study.*

hand, our algorithm suffers from fewer problems in the solve phase than GEPP. On the 7 Betts matrices the residual was always at least a factor of 10^{12} smaller than the right-hand side; we did not detect any problems.

We also conducted experiments to assess the behavior of the algorithm under various α-thresholds. For this experiment we selected two of the Betts matrices and two particularly difficult matrices from the Gould–Scott collection, CONT-300 and BRATU3D. These two matrices caused large growth under $\alpha = 1/3$; CONT-300 caused instability, but BRATU3D did not. Figure 5 shows the norms of the forward errors, normalized with respect to $\|x\|$, as a function of α. The results show that increasing α can improve significantly the accuracy of the algorithm. This is particularly pronounced in the difficult Gould–Scott matrices. The accuracy appears to be roughly monotone with α. We note that a small α improves performance, as we shall see later, but even with a large α our theoretical results concerning growth and bandwidth hold.

6.4. Performance. Next, we describe several experiments that evaluate the performance of our algorithm. These experiments were all carried out on the Pentium 4 machine.

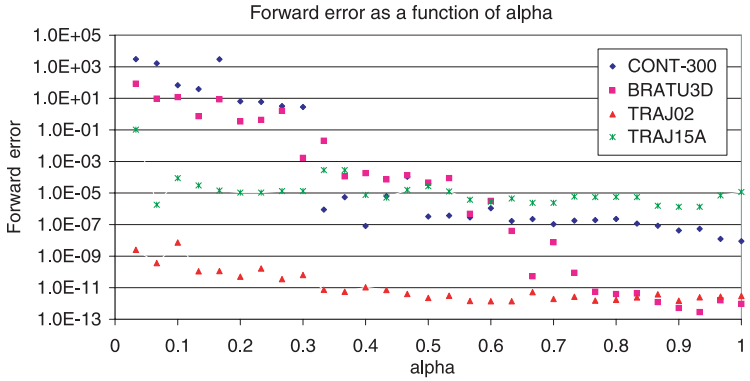


FIG. 5. *Forward accuracy as a function of α on four matrices.*

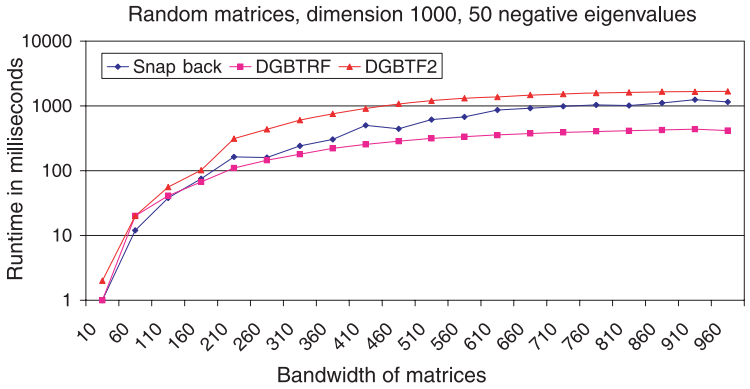


FIG. 6. *Performance as a function of the bandwidth.*

Figure 6 shows the performance of our algorithm relative to LAPACK's implementations of GEPP. The experiment was performed on random 1000-by-1000 matrices with 50 negative eigenvalues. We generated the matrices as follows. We start with a real normally distributed random nonsymmetric matrix T, and compute its QR factorization, $T = QR$, where Q is unitary and R is upper triangular. We then select 1000 uniform random values w_i between 0 and 25. From the w_i's we construct a diagonal matrix Λ with diagonal elements $\lambda_i = (-1)^{[i \le \nu]} 2^{w_i}$, where $[i \le \nu]$ is 1 when $i \le \nu$ and 0 otherwise. Next, we compute $Q\Lambda Q^T$, which has eigenvalues λ_i, exactly ν of which are negative and the rest positive. Finally, we apply the SBR library [3] to $Q\Lambda Q^T$ with the target bandwidth as input. This unitarily reduces $Q\Lambda Q^T$ to a banded symmetric matrix A with the same spectrum as $Q\Lambda Q^T$ and Λ. By the construction of the λ_i's, A is ill conditioned but far from numerical rank deficiency in double-precision IEEE 754 arithmetic.

The results show that our algorithm is always faster than LAPACK's subroutine DGBTF2, which is not blocked for cache efficiency. When the bandwidth is small (e.g., 60 in this experiment), our algorithm is also faster than DGBTRF, which partitions the matrix for cache efficiency. For matrices with higher bandwidth, DGBTRF is faster than our algorithm.

Figure 7 presents similar results on real-world, nonrandom matrices, S3RMT3M1

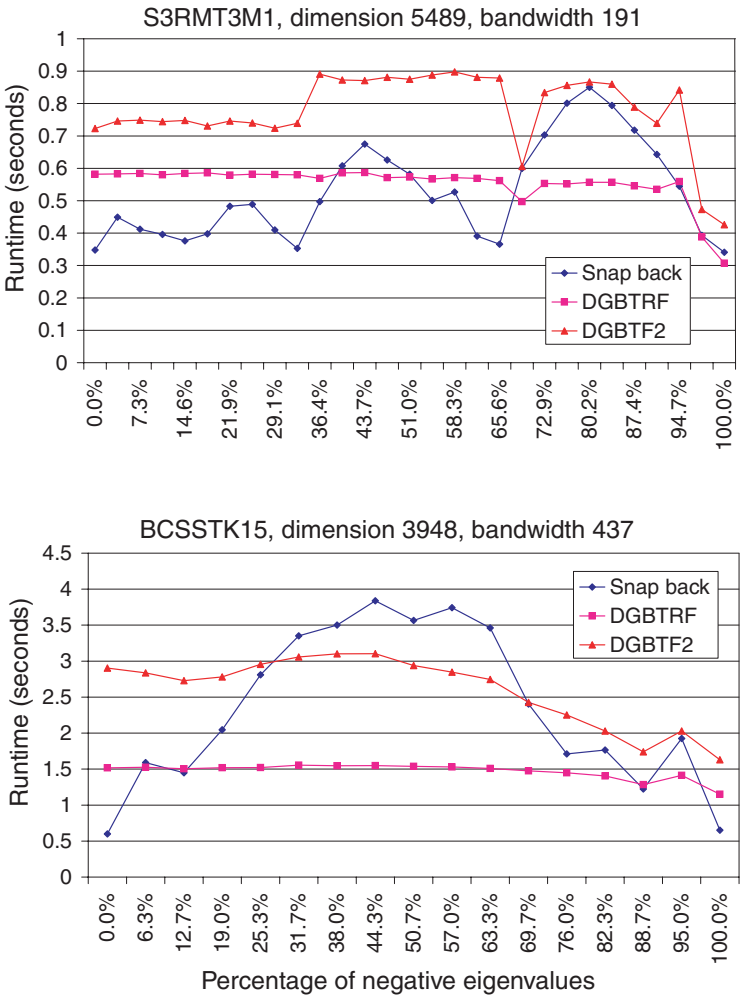


FIG. 7. *Performance on two real-world matrices, shifted to produce different numbers of negative eigenvalues.*

and BCSSTK15. Both are available from public sparse-matrix collections.[3] The figure plots the performance of the three algorithms on shifted versions of the two matrices. All the shifts are half-way between eigenvalues, so the matrices are relatively well conditioned. On S3RMT3M1, which has a relatively narrow bandwidth, our algorithm outperforms both GEPP implementations at many inertia values, and is always at least as fast as DGBTF2. On BCSSTK15, which has a much larger bandwidth, our algorithm is sometimes faster and sometimes slower than DGBTF2 but usually slower than DGBTRF. On this matrix our algorithm exhibits another behavior: higher performance at the ends of the inertia axis than in the middle. We further explore this behavior below.

In another experiment on real-world matrices, we compared the performance of our algorithm to that of DGBTF2 and DGBTRF on the 33 matrices from the Gould–

[3]For example, from http://math.nist.gov/MatrixMarket/.

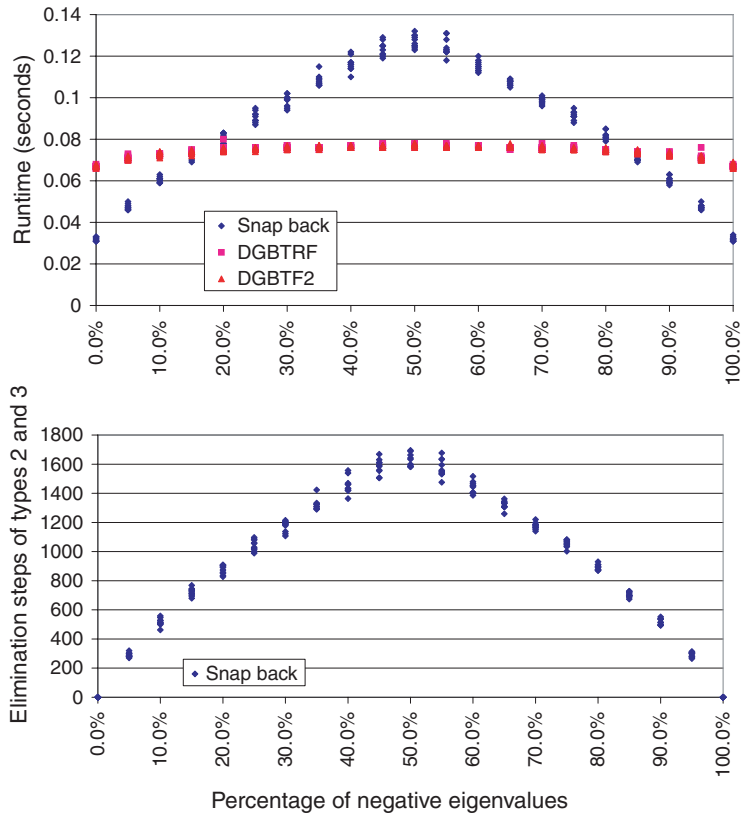


FIG. 8. *Running times of our algorithm on random* 5000*-by-*5000 *matrices with bandwidth* 50*, and the number of type* 2 *and type* 3 *eliminations steps (combined). Each dot corresponds to one matrix.*

Scott study, on the Opteron machine. Most of these matrices have high bandwidths. Our algorithm delivered similar performance to DGBTF2, but DGBTRF was much faster than our algorithm. We omit the detailed results of this experiment.

Figure 8 presents the results of another experiment. For this experiment, we used random 5000-by-5000 matrices with bandwidth 50 and with varying inertia, 10 matrices for each inertia value. We used the Pentium 4 machine for the experiment. The data shown in the top graph of Figure 8 shows that the algorithm runs fastest when most of the eigenvalues have the same sign, and it degrades as the ratio of positive to negative eigenvalues nears 1. This is consistent with the results on some real-world matrices, such as BCSSTK15. The algorithm does not always exhibit this behavior, as the results on S3RMT3M1 show, but it often does.

The bottom graph in Figure 8 explains this phenomenon. The graph shows that near the middle of the inertia axis, the number of type-two and -three elimination steps grows. These elimination steps are more expensive than type-one steps, and they produce more fill, so a larger number of these steps slows down the algorithm. It appears that at least in some cases, the number of type-two and -three steps depends on the inertia.

6.5. Comparisons to other symmetric-indefinite factorization codes. In addition for the tests introduced in this section, we have performed some tests on the SBR library. We have found that the performances of SBR and of our algorithm are almost incomparable: SBR is about three orders of magnitude (about 1000 times) slower.

We were not able to obtain the code of Jones and Patrick [14]. Therefore, we did not compare our algorithm to theirs. We believe that when matrices have only few negative (or few positive) eigenvalues, the behavior of our algorithm and of Jones and Patrick's Bunch–Kaufman code are similar. However, Jones and Patrick's algorithm is not designed for general symmetric banded matrices; it may suffer catastrophic fill when the ratio of positive to negative eigenvalues is near 1, whereas our algorithm degrades smoothly toward this case, and the degradation is never catastrophic (because the bandwidth of the reduced matrices is bounded).

7. Conclusions. The algorithm that we propose in this paper is the first banded symmetric direct solver that exploits symmetry and achieves an $O(nm^2)$ running time and an $O(nm)$ storage requirement.

The reduced matrices in our algorithm may fill somewhat, but they remain banded. This behavior is similar to that of Gaussian elimination with partial pivoting (GEPP) when applied to a banded matrix. However, representation of the factors in our algorithm can require more memory than the representation of the GEPP factors. Also, our algorithm represents the factors in a product form, consisting of several elementary transformation matrices per elimination step. This representation does not allow our algorithm to be easily blocked for cache efficiency, so for large m and highly-indefinite matrices, our algorithm can be slower than the blocked version of GEPP.

When most of the diagonal elements are large enough to be used as 1-by-1 pivots, our algorithm performs much less work than GEPP. The performance of the algorithm seems to be related to the ratio of the numbers of positive and negative eigenvalues. When there are only a few negative (or only a few positive) eigenvalues, our algorithm uses mostly cheap symmetric Gaussian reduction steps. When there are many eigenvalues of both signs, the algorithm must resort to more expensive Givens and unsymmetric Gaussian reduction steps, so its performance degrades. But even when m is large and A is highly indefinite, our new algorithm is competitive with the unblocked version of GEPP. For small m, our algorithm outperforms even the blocked version of GEPP.

The new algorithm is reliable when implemented in floating-point arithmetic. More specifically, we believe that the algorithm is backward stable, but we formally show only a weaker result: that the element growth is bounded by 4^{n-1}. That is, the entries of the reduced matrices are at most a factor of 4^{n-1} larger in absolute value than the entries of A. In most of the existing elimination algorithms, including GEPP, Bunch–Kaufman and Aasen, a result of this type holds (2 in GEPP, 2.57 in Bunch–Kaufman and 4 in Aasen) and can be used to show backward stability. The existence of such a bound on the growth, in both existing algorithms and in our new algorithm, reflects a careful numerical design whose goal is to avoid catastrophic cancellation. In particular, the growth bound that we show in section 5 holds thanks to an intricate elimination strategy that is designed to simultaneously avoid growth, maintain symmetry, and maintain the band structure. Also, the elements of the factors in our method are bounded in magnitude by the maximum of 1 and the elements of the reduced matrices. We note that the backward stability of the Aasen and Bunch–

Kaufman algorithms, which were proposed in 1971 and 1977, respectively, was only formally proved by Higham in [11, 12].

Numerical experiments show that in practice, growth is almost always much smaller than the worst-case bound, and residuals are small, suggesting that the algorithm is backward stable. Furthermore, the accuracy (forward errors) of our algorithm is similar to that of GEPP even on notoriously difficult matrices. This suggests that the accuracy problems in Bunch–Kaufman, reported in [2], are not present in our algorithm.

This paper leaves a few interesting questions open.

- Can this algorithm be blocked for cache efficiency? That is, can this algorithm be accelerated by exploiting fast level-3 BLAS subroutines?
- Is this algorithm backward stable? Our bound on the element growth, the boundedness of the factors, and our numerical experiments suggest that it is, but we have not formally proved backward stability.
- Can a similar elimination scheme be developed for symmetric indefinite matrices with a general sparsity pattern?

REFERENCES

[1] J. O. Aasen, *On the reduction of a symmetric matrix to tridiagonal form*, Nordisk Tidskr. Informationsbehandling (BIT), 11 (1971), pp. 233–242.

[2] C. Ashcraft, R. G. Grimes, and J. G. Lewis, *Accurate symmetric indefinite linear equation solvers*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 513–561.

[3] C. H. Bischof, B. Lang, and X. Sun, *A framework for symmetric band reduction*, ACM Trans. Math. Software, 26 (2000), pp. 581–601.

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

[5] I. A. Cavers, *A hybrid tridiagonalization algorithm for symmetric sparse matrices*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1363–1380.

[6] I. S. Duff and J. K. Reid, *MA27 – A Set of Fortran Subroutines for Solving Sparse Symmetric Sets of Linear Equations*, Tech. report AERE R10533, AERE Harwell Laboratory, London, UK, 1982.

[7] I. S. Duff and J. K. Reid, *The multifrontal solution of indefinite sparse symmetric linear systems*, ACM Trans. Math. Software, 9 (1983), pp. 302–325.

[8] I. S. Duff and J. K. Reid, *MA47, a Fortran Code for Direct Solution of Indefinite Sparse Symmetric Linear Systems*, Tech. report RAL-95-001, Rutherford Appleton Laboratory, Didcot, Oxon, UK, 1995.

[9] L. Fox, H. D. Huskey, and J. H. Wilkinson, *Notes on the solution of algebraic linear simultaneous equations*, Quart. J. Mech. Appl. Math., 1 (1948), pp. 149–173.

[10] N. I. M. Gould and J. A. Scott, *Complete Results from a Numerical Evaluation of hsl Packages for the Direct-Solution of Large Sparse, Symmetric Linear Systems of Equations*, Tech. report, Numerical Analysis Internal Report 2003-2, Rutherford Appleton Laboratory, 2003. Available online from http://www.numerical.rl.ac.uk/reports/reports.shtml.

[11] N. J. Higham, *Stability of Aasen's method*, manuscript, 1997.

[12] N. J. Higham, *Stability of the diagonal pivoting method with partial pivoting*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 52–65.

[13] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., Society for Industrial and Applied Mathematics, Philadelphia, 2002.

[14] M. T. Jones and M. L. Patrick, *Bunch-Kaufman factorization for real symmetric indefinite banded matrices*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 553–559.

[15] B. Lang, *A parallel algorithm for reducing symmetric banded matrices to tridiagonal form*, SIAM J. Sci. Comput., 14 (1993), pp. 1320–1338.

[16] K. Murata and K. Horikoshi, *A new method for the tridiagonalization of the symmetric band matrix*, Information Processing in Japan, 15 (1975), pp. 108–112.

[17] B. N. Parlett and J. K. Reid, *On the solution of a system of linear equations whose matrix is symmetric but not definite*, BIT, 10 (1970), pp. 386–397.

[18] H. Rutishauser, *On Jacobi rotation patterns*, Proc. Sympos. Appl. Math. 15, Amer. Math. Soc., Providence, RI, 1963, pp. 219–239.
[19] H. R. Schwarz, *Tridiagonalization of a symmetric band matrix*, Numer. Math., 12 (1968), pp. 231–241.
[20] G. W. Stewart, *The economical storage of plane rotations*, Numer. Math., 25 (1976), pp. 137–138.
[21] J. H. Wilkinson, *Error analysis of direct methods of matrix inversion*, J. Assoc. Comput. Math., 8 (1961), pp. 281–330.
[22] J. H. Wilkinson, *Rounding Errors in Algebraic Processes*, Prentice-Hall, Englewood Cliffs, NJ, 1963.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 2, pp. 425–445

FINDING A GLOBAL OPTIMAL SOLUTION FOR A QUADRATICALLY CONSTRAINED FRACTIONAL QUADRATIC PROBLEM WITH APPLICATIONS TO THE REGULARIZED TOTAL LEAST SQUARES*

AMIR BECK†, AHARON BEN-TAL†, AND MARC TEBOULLE‡

Abstract. We consider the problem of minimizing a fractional quadratic problem involving the ratio of two indefinite quadratic functions, subject to a two-sided quadratic form constraint. This formulation is motivated by the so-called regularized total least squares (RTLS) problem. A key difficulty with this problem is its nonconvexity, and all current known methods to solve it are guaranteed only to converge to a point satisfying first order necessary optimality conditions. We prove that a global optimal solution to this problem can be found by solving a sequence of very simple convex minimization problems parameterized by a single parameter. As a result, we derive an efficient algorithm that produces an ϵ-global optimal solution in a computational effort of $O(n^3 \log \epsilon^{-1})$. The algorithm is tested on problems arising from the inverse Laplace transform and image deblurring. Comparison to other well-known RTLS solvers illustrates the attractiveness of our new method.

Key words. regularized total least squares, fractional programming, nonconvex quadratic optimization, convex programming

AMS subject classifications. 65F20, 90C20, 90C32

DOI. 10.1137/040616851

1. Introduction.

In this paper we consider the problem of minimizing a fractional quadratic function subject to a quadratic constraint:

$$\min_{\mathbf{x}\in\mathcal{F}} \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}, \tag{1}$$

where

$$f_i(\mathbf{x}) = \mathbf{x}^T\mathbf{A}_i\mathbf{x} - 2\mathbf{b}_i^T\mathbf{x} + c_i, \quad i = 1, 2, \tag{2}$$

$\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{n\times n}$ are symmetric matrices, $\mathbf{b}_1, \mathbf{b}_2 \in \mathbb{R}^n$, $c_1, c_2 \in \mathbb{R}$, and $0 \le L < U$. We do not assume that $\mathbf{A}_1$ and $\mathbf{A}_2$ are positive semidefinite, and the only assumption required for the problem to be well defined is that $f_2(\mathbf{x})$ is bounded away from zero. We will discuss two cases of the feasible set $\mathcal{F}$:

$$\mathcal{F}_1 = \{\mathbf{x} \in \mathbb{R}^n : L^2 \le \mathbf{x}^T\mathbf{T}\mathbf{x} \le U^2\},$$

where $\mathbf{T}$ is a positive definite matrix and $U > L \ge 0$, and

$$\mathcal{F}_2 = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}^T\mathbf{B}\mathbf{x} \le U^2\},$$

where $\mathbf{B}$ is a positive semidefinite matrix and $U > 0$.

*Received by the editors October 13, 2004; accepted for publication (in revised form) by P. C. Hansen November 29, 2005; published electronically May 26, 2006.
http://www.siam.org/journals/simax/28-2/61685.html

†MINERVA Optimization Center, Department of Industrial Engineering and Management Technion, Israel Institute of Technology, Haifa 3200, Israel (becka@ie.technion.ac.il, abental@ie.technion.ac.il). The research of the second author was partially supported by BSF grant 2002038.

‡School of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv 69978, Israel (teboulle@post.tau.ac.il). The research of this author was partially supported by BSF grant 2002010.

The major difficulty associated with problem (1) is the nonconvexity of the objective function and in the case of $\mathcal{F}_1$ also the nonconvexity of the feasible set.

The main motivation for considering problem (1) comes from the so-called regularized total least squares (RTLS) problem. Many problems in data fitting and estimation give rise to an overdetermined system of linear equations $\mathbf{Ax} \approx \mathbf{b}$, where both the matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and the vector $\mathbf{b} \in \mathbb{R}^m$ are contaminated by noise. The total least squares (TLS) approach to this problem [9, 10, 15] is to seek a perturbation matrix $\mathbf{E} \in \mathbb{R}^{m\times n}$ and a perturbation vector $\mathbf{r} \in \mathbb{R}^m$ that minimize $\|\mathbf{E}\|^2 + \|\mathbf{r}\|^2$ subject to the consistency equation $(\mathbf{A}+\mathbf{E})\mathbf{x} = \mathbf{b}+\mathbf{r}$ (here and elsewhere in this paper a matrix norm is always the Frobenius norm and a vector norm is the Euclidean one). The TLS approach was extensively used in a variety of scientific disciplines such as signal processing, automatic control, statistics, physics, economic, biology, and medicine (see, e.g., [15] and the references therein). The TLS problem has essentially an explicit solution, expressed by the singular value decomposition of the augmented matrix $(\mathbf{A}, \mathbf{b})$.

Regularization of the TLS solution is required in the case where $\mathbf{A}$ is nearly rank deficient. Such problems arise, for example, from the discretization of ill-posed problems such as integral equations of the first kind (see, e.g., [8, 13] and the references therein). In these problems the TLS solution can be physically meaningless, and thus regularization is employed in order to stabilize the solution.

Regularization of the TLS solution was addressed by several approaches: truncation methods [5, 13], Tikhonov regularization [8], and recently by introducing a quadratic constraint [20, 11, 8]. All the above methods are still trapped in the nonconvexity of the problem and thus are not guaranteed to converge to a global optimum. At best, they are proven to converge to a point satisfying first order necessary optimality condition. In contrast, in this paper, we develop an efficient algorithm which finds the global optimal solution by converting the original problem into a sequence of very simple convex optimization problems parameterized by a single parameter α. The optimal solution corresponds to a particular value of α, which can be found by a simple one-dimensional search. The algorithm finds an ϵ-optimal solution $\mathbf{x}^*$ of (1), i.e.,

$$\frac{f_1(\mathbf{x}^*)}{f_2(\mathbf{x}^*)} \leq \min_{\mathbf{x}\in\mathcal{F}} \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} + \epsilon,$$

in a computational effort of order $O\left(n^3 \log\left(\frac{1}{\epsilon}\right)\right)$.

The paper is organized as follows. In the next section, we show how to recover the formulation of the RTLS problem as a quadratically constrained fractional quadratic problem. Section 3 describes a schematic algorithm designed to solve (1) for general quadratic functions f_1 and f_2 that provides the starting point of the analysis and the main results that are developed in section 4. In section 5 we return to the RTLS problem and give a detailed algorithm (RTLSC) for its solution. In order to illustrate the performance of algorithm RTLSC, two problems from the "Regularization Tools" [13] are employed: a problem that arises from the discretization of the inverse Laplace transform and an image deblurring problem. These numerical examples are reported in section 6, where we also compare the performance of our algorithm RTLSC with other well-known RTLS solvers. Some useful technical results used throughout the paper are collected in the appendix.

2. The RTLS problem. In this section we show how to recover a known formulation of the RTLS problem as a quadratically constrained fractional quadratic

programming. This result is well known [9, 15, 20]. However, we believe that the derivation we give below is simpler. The RTLS problem as stated in [20] is

$$\begin{array}{ll} \min_{\mathbf{E},\mathbf{r},\mathbf{x}} & \|\mathbf{E}\|^2 + \|\mathbf{r}\|^2 \\ \text{subject to} & (\mathbf{A}+\mathbf{E})\mathbf{x} = \mathbf{b}+\mathbf{r}, \\ & \mathbf{x} \in \mathcal{F}_2. \end{array} \tag{3}$$

To show that the RTLS problem (3) is a special case of problem (1), let us write (3) as

$$\min_{\mathbf{x}\in\mathcal{F}_2} \quad \min_{\mathbf{E},\mathbf{r}:(\mathbf{A}+\mathbf{E})\mathbf{x}=\mathbf{b}+\mathbf{r}} \|\mathbf{E}\|^2 + \|\mathbf{r}\|^2. \tag{4}$$

Next, fix $\mathbf{x} \in \mathcal{F}_2$ and consider the inner minimization problem in (4). Denote $\mathbf{w} = vec(\mathbf{E}, \mathbf{r})$, where, for a matrix $\mathbf{M}$, $vec(\mathbf{M})$ denotes the vector obtained by stacking the columns of $\mathbf{M}$. The linear constraint (in $\mathbf{E}$ and $\mathbf{r}$) $(\mathbf{A}+\mathbf{E})\mathbf{x} = \mathbf{b}+\mathbf{r}$ can be written as $\mathbf{Q}_\mathbf{x}\mathbf{w} = \mathbf{b} - \mathbf{A}\mathbf{x}$, where

$$\mathbf{Q}_\mathbf{x} = \begin{pmatrix} \widetilde{\mathbf{x}}^T & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{x}}^T & \dots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \widetilde{\mathbf{x}}^T \end{pmatrix}$$

and $\widetilde{\mathbf{x}} = (\mathbf{x}^T, -1)^T$. Thus, the inner minimization problem in (4) takes the form

$$\min_{\mathbf{Q}_\mathbf{x}\mathbf{w}=\mathbf{b}-\mathbf{A}\mathbf{x}} \|\mathbf{w}\|^2. \tag{5}$$

Using the KKT conditions, it is easy to see that the solution of (5) is attained at $\mathbf{w} = \mathbf{Q}_\mathbf{x}^T(\mathbf{Q}_\mathbf{x}\mathbf{Q}_\mathbf{x}^T)^{-1}(\mathbf{b}-\mathbf{A}\mathbf{x})$, and as a result the optimal value of problem (5) is equal to

$$(\mathbf{b}-\mathbf{A}\mathbf{x})^T(\mathbf{Q}_\mathbf{x}\mathbf{Q}_\mathbf{x}^T)^{-1}(\mathbf{b}-\mathbf{A}\mathbf{x}).$$

Since $\mathbf{Q}_\mathbf{x}\mathbf{Q}_\mathbf{x}^T = \|\widetilde{\mathbf{x}}\|^2\mathbf{I}$ we deduce that the value of the inner minimization problem (5) is equal to $\frac{\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2}{\|\widetilde{\mathbf{x}}\|^2} = \frac{\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2}{\|\mathbf{x}\|^2+1}$. Consequently, the value of the RTLS problem (3) reduces to

$$\min_{\mathbf{x}\in\mathcal{F}_2} \frac{\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2}{\|\mathbf{x}\|^2+1}, \tag{6}$$

which is indeed a special case of problem (1).

3. A schematic algorithm. We consider problem (1) and henceforth make the following assumption.

ASSUMPTION 1. f_2 is bounded below on $\mathcal{F}$ by a positive number N.

Let m and M be numbers such that

$$m \le \min_{\mathbf{x}\in\mathcal{F}} \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \le M. \tag{7}$$

Such bounds are easy to find; see section 4.3.

REMARK 3.1. For the RTLS problem (6), Assumption 1 is trivially satisfied for $N = 1$. The lower bound m can be chosen as 0 and M can be taken to be $f(0) = \|\mathbf{b}\|^2$.

Although both denominator and nominator in the RTLS problem (6) are [illegible], this property does not make the problem simpler since the quotient of convex functions is not necessarily convex.

A simple observation that goes back to Dinkelbach [4] and will enable us to solve (1) is the following.

[illegible]. The following two statements are equivalent:

1. $\min_{\mathbf{x}\in\mathcal{F}} \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \leq \alpha$.
2. $\min_{\mathbf{x}\in\mathcal{F}} \{f_1(\mathbf{x}) - \alpha f_2(\mathbf{x})\} \leq 0$.

Using the above observation, we can solve (1) by the following schematic bisection algorithm.

SCHEMATIC ALGORITHM

Initial Step: Set $lb_0 = m$ and $ub_0 = M$.

General Step: For every $k \geq 1$:

1. Define $\alpha_k = \frac{lb_{k-1}+ub_{k-1}}{2}$.
2. Calculate $\beta_k = \min_{\mathbf{x}\in\mathcal{F}} \{f_1(\mathbf{x}) - \alpha_k f_2(\mathbf{x})\}$.
   (a) If $\beta_k \leq 0$, then define $lb_k = lb_{k-1}$ and $ub_k = \alpha_k$.
   (b) If $\beta_k > 0$, then define $lb_k = \alpha_k$ and $ub_k = ub_{k-1}$.

Stopping Rule: Stop at the first iteration k^* that satisfies $ub_{k^*} - lb_{k^*} \leq \epsilon$.

Output:

$$\mathbf{x}^* \in \operatorname*{argmin}_{\mathbf{x}\in\mathcal{F}} \{f_1(\mathbf{x}) - ub_{k^*} f_2(\mathbf{x})\}. \tag{8}$$

PROPOSITION 3.1. [illegible] $\left\lceil \ln\left(\frac{M-m}{\epsilon}\right)/ln(2)\right\rceil$ [illegible] $\mathbf{x}^*$ [illegible] ϵ [illegible] (1). [illegible]

$$\mathbf{x}^* \in \mathcal{F}, \quad \alpha^* \leq \frac{f_1(\mathbf{x}^*)}{f_2(\mathbf{x}^*)} \leq \alpha^* + \epsilon,$$

[illegible] $\alpha^* = \min_{\mathbf{x}\in\mathcal{F}} \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}$

[illegible]. The length of the initial interval is $ub_0 - lb_0 = M - m$. By the definition of lb_k and ub_k, we have that for every $k \geq 1$, $ub_k - lb_k = \frac{1}{2}(ub_{k-1} - lb_{k-1})$, and therefore $ub_k - lb_k = (M-m)\left(\frac{1}{2}\right)^k$. From this it follows that k^*, the number of iterations of the schematic algorithm, is the smallest integer k satisfying

$$(M-m)\left(\frac{1}{2}\right)^k \leq \epsilon,$$

which is equivalent to $k \geq \left\lceil \ln\left(\frac{M-m}{\epsilon}\right)/ln(2)\right\rceil$. By (8), $\mathbf{x}^*$ is feasible, i.e., $\mathbf{x}^* \in \mathcal{F}$. Also, by the definition of the bisection process we have that $lb_k \leq \alpha^* \leq ub_k$. By (8) we have that $lb_k \leq \alpha^* \leq \frac{f_1(\mathbf{x}^*)}{f_2(\mathbf{x}^*)} \leq ub_k$ for every k and finally, since $ub_{k^*} \leq lb_{k^*} + \epsilon$, the result follows. □

[illegible] 3.2. By writing "min" and not "inf" in statements 1 and 2 of the observation and in the above scheme, we implicitly assumed that the minimum of the corresponding problems is attained (which is certainly the case when $\mathcal{F} = \mathcal{F}_1$).

Otherwise, the inequalities in the statements of the observation should be replaced by strict inequalities and the schematic algorithm revised accordingly. The schematic algorithm then terminates with a point $\mathbf{x}^*$, at which the objective value is at most ϵ away from the infimum. Thus henceforth we will assume that the minimum is attained.

To convert the schematic algorithm to a practical scheme we still need to address the following two questions:

1. How do we choose the lower and upper bound m and M?
2. How do we solve the subproblem

$$\min_{\mathbf{x}\in\mathcal{F}}\{f_1(\mathbf{x}) - \alpha f_2(\mathbf{x})\}? \tag{9}$$

The first question is rather easy (see section 4.3). The second one is seemingly more difficult since problem (9), like the original problem (1), is nonconvex. In the next section we give complete answers to these two questions.

4. Analysis and main results. In sections 4.1 and 4.2 we show how to efficiently solve the subproblem (9). We first transform the problem (9) into a convex optimization problem by using the methodology of Ben-Tal and Teboulle [2]. We then show that the solution of the derived convex optimization problem consists of one eigenvector decomposition and solutions of at most two one-dimensional secular equations [17]. Finally, in section 4.3 we show how to find the lower and upper bounds m and M.

4.1. Solving the subproblem in the case $\mathcal{F} = \mathcal{F}_1$. In this section we consider the case in which the feasible set is equal to $\{\mathbf{x} : L^2 \le \mathbf{x}^T\mathbf{T}\mathbf{x} \le U^2\}$, where $\mathbf{T}$ is a positive definite matrix. Notice that in this case, the feasible set is compact, and thus the minimum is always attained both in the original problem (1) and in the subproblem (9). First, we convert problem (9) to one with an Euclidean norm constraint by making the change of variables $\mathbf{s} = \mathbf{T}^{1/2}\mathbf{x}$. The result is the following optimization problem:

$$\min_{L^2\le\|\mathbf{s}\|^2\le U^2}\left\{f_1(\mathbf{T}^{-1/2}\mathbf{s}) - \alpha f_2(\mathbf{T}^{-1/2}\mathbf{s})\right\}. \tag{10}$$

Using the notation

$$\begin{aligned}\widetilde{\mathbf{A}} &= \mathbf{T}^{-1/2}(\mathbf{A}_1 - \alpha\mathbf{A}_2)\mathbf{T}^{-1/2},\\ \widetilde{\mathbf{b}} &= \mathbf{T}^{-1/2}(\mathbf{b}_1 - \alpha\mathbf{b}_2),\\ \widetilde{c} &= c_1 - \alpha c_2,\end{aligned}$$

we obtain that problem (10) is the same as

$$(\mathrm{P}): \quad \min_{L\le\|\mathbf{s}\|\le U}\{\mathbf{s}^T\widetilde{\mathbf{A}}\mathbf{s} - 2\widetilde{\mathbf{b}}^T\mathbf{s} + \widetilde{c}\}. \tag{11}$$

$\widetilde{\mathbf{A}}$ is symmetric and hence can be diagonalized by an orthogonal matrix $\mathbf{U}$, so that

$$\mathbf{U}^T\widetilde{\mathbf{A}}\mathbf{U} = \mathbf{D} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n), \tag{12}$$

where $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n$. Making the change of variables $\mathbf{s} = \mathbf{U}\mathbf{z}$ we obtain that (11) is equivalent to

$$\min_{L^2\le\|\mathbf{z}\|^2\le U^2}\left\{\sum_{j=1}^{n}(\lambda_j z_j^2 - 2f_j z_j) + \widetilde{c}\right\}, \tag{13}$$

where $\mathbf{f} = \mathbf{U}^T\mathbf{b}$. The following lemma will enable us to transform problem (13) into a convex optimization problem.

LEMMA 4.1. $(z_1^*, z_2^*, \ldots, z_n^*)$

$$\min_{L^2 \le \|\mathbf{z}\|^2 \le U^2} q(z),$$

(14) $$q(z) = \sum_{j=1}^{n} (\lambda_j z_j^2 - 2 f_j z_j).$$

$z_j^* f_j \ge 0$ $j = 1, 2, \ldots, n$ $f_j \ne 0$. Since $\mathbf{w} = (z_1^*, z_2^*, \ldots, z_n^*)$ is optimal it is in particular feasible, i.e., $L^2 \le \|\mathbf{w}\|^2 \le U^2$. An immediate result is that $(z_1^*, z_2^*, \ldots, z_{k-1}^*, -z_k^*, z_{k+1}^*, \ldots, z_n^*)$ is also feasible for every $k = 1, 2, \ldots, n$. Since $\mathbf{w}$ is optimal we have that for every $k = 1, 2, \ldots, n$,

(15) $$q(z_1^*, \ldots, z_n^*) \le q(z_1^*, \ldots, z_{k-1}^*, -z_k^*, z_{k+1}^*, \ldots, z_n^*).$$

Substituting (14) into (15) yields

$$\sum_{j=1}^{n} (\lambda_j (z_j^*)^2 - 2 f_j z_j^*) \le \sum_{j=1, j\ne k}^{n} (\lambda_j (z_j^*)^2 - 2 f_j z_j^*) + \lambda_k (-z_k^*)^2 + 2 f_k z_k^*.$$

Therefore, $f_k z_k^* \ge 0$, and the result follows. □

Note that if $f_j = 0$ for some j, then the objective function $q(z)$ is symmetric with respect to z_j and as a result we can arbitrarily restrict z_j to be nonnegative or nonpositive. In view of this and Lemma 4.1, we can make the change of variables

(16) $$z_j = \operatorname{sign}(f_j)\sqrt{v_j}, \quad j = 1, 2, \ldots, n,$$

where $v_j \ge 0$. Substituting (16) into (13), we conclude that problem (9) is equivalent to the optimization problem

(17) $$\min_{v_j \ge 0} \left\{ \sum_{j=1}^{n} \left(\lambda_j v_j - 2|f_j|\sqrt{v_j}\right) + \widetilde{c} : L^2 \le \sum_{j=1}^{n} v_j \le U^2 \right\}.$$

PROPOSITION 4.1. $\widetilde{\mathbf{A}} \in \mathbb{R}^{n\times n}$ $\widetilde{\mathbf{b}} \in \mathbb{R}^n, \widetilde{c} \in \mathbb{R}$ $\widetilde{\mathbf{A}}$ $\widetilde{\mathbf{A}} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ $\mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n$

$$\min_{L^2 \le \|\mathbf{s}\|^2 \le U^2} \left\{ \mathbf{s}^T \widetilde{\mathbf{A}} \mathbf{s} - 2\widetilde{\mathbf{b}}^T \mathbf{s} + \widetilde{c} \right\}$$

$\mathbf{s} = \mathbf{U}\mathbf{z}$

$$z_j = sign(f_j)\sqrt{v_j}, \quad j = 1, 2, \ldots, n,$$

$\mathbf{v}$ (17).

Proposition 4.1 shows that the main step in the schematic algorithm (step 2) consists of solving the linearly constrained convex optimization problem (17). This

will be done by solving the dual problem, since, as we are about to show, the latter requires the solution of at most two [illegible] convex problems.

To develop the dual problem of (17), we assign a nonnegative multiplier ξ to the linear inequality constraint $-\sum_{j=1}^{n} v_j + L^2 \le 0$ and a nonpositive multiplier η to the linear inequality constraint $-\sum_{j=1}^{n} v_j + U^2 \ge 0$ and form the Lagrangian of (17):

$$L(\mathbf{v},\eta,\xi) = \sum_{j=1}^{n} \left(\lambda_j v_j - 2|f_j|\sqrt{v_j}\right) - \eta\left(\sum_{j=1}^{n} v_j - U^2\right) + \xi\left(-\sum_{j=1}^{n} v_j + L^2\right) + \widetilde{c}$$

$$= \sum_{j=1}^{n} \left((\lambda_j - \eta - \xi)v_j - 2|f_j|\sqrt{v_j}\right) + \eta U^2 + \xi L^2 + \widetilde{c}. \tag{18}$$

Differentiating (18) with respect to v_j and equating to zero, we obtain

$$v_j = \frac{f_j^2}{(\lambda_j - \eta - \xi)^2}, \quad j = 1, 2, \ldots, n, \tag{19}$$

subject to the conditions $\eta + \xi \le \lambda_n, \eta \le 0$, and $\xi \ge 0$. Thus, the dual objective function is given by

$$\inf_{v_j \ge 0} L(\mathbf{v},\eta,\xi) = \begin{cases} h(\eta,\xi) & \text{if } \eta - \xi > -\lambda_n, \eta \le 0, \xi \ge 0, \\ -\infty & \text{otherwise,} \end{cases}$$

where

$$h(\eta,\xi) \stackrel{\triangle}{=} -\sum_{j=1}^{n} \frac{f_j^2}{\lambda_j - \eta - \xi} + \eta U^2 + \xi L^2 + \widetilde{c}$$

and the dual problem of (17) is

$$\text{(D)}: \qquad \max_{\eta,\xi} \{h(\eta,\xi) : \eta + \xi < \lambda_n, \eta \le 0, \xi \ge 0\}.$$

From duality theory for convex optimization problems we have that [19, 3]

$$val(\text{P}) = val(\text{D}),$$

where $val(\text{P})$ ($val(\text{D})$) denotes the optimal value of problem (P) (problem (D)). Now we note that the dual variables η and ξ cannot both be nonzero, since in that case we would have by the complementarity slackness condition that $\sum_{j=1}^{n} v_j$ is equal to both U^2 and L^2, which is clearly a contradiction. As a result, instead of considering the problem (D) in two variables, we can consider the following two single-variable convex optimization problems (maximization of concave functions subject to a simple convex bound constraint):

$$\text{(D1)}: \qquad \max_{\eta \le \min\{\lambda_n, 0\}} \underbrace{-\sum_{j=1}^{n} \frac{f_j^2}{\lambda_j - \eta} + \eta U^2 + \widetilde{c}}_{h(\eta,0)}$$

and

$$\text{(D2)}: \qquad \max_{0 \le \xi < \lambda_n} \underbrace{-\sum_{j=1}^{n} \frac{f_j^2}{\lambda_j - \xi} + \xi L^2 + \widetilde{c}}_{h(0,\xi)}.$$

We thus obtain that in order to solve (D), we need to follow the following three steps:

1. Find a solution η of (D1).
2. Find a solution ξ of (D2).
3. If $h(\eta,0) > h(0,\xi)$, then the solution of (D) is $(\eta,0)$. Otherwise, the solution is $(0,\xi)$.

Notice that both (D1) and (D2) are easy problems to solve since they consist of maximizing a concave function of a single variable. A very efficient algorithm for solving problems with an exact structure as (D1) and (D2) will be discussed at the end of this section.

We summarize our results on the solution of (11) in the following theorem.

THEOREM 4.1. [illegible] $\widetilde{\mathbf{A}} \in \mathbb{R}^{n\times n}$ [illegible] $\widetilde{\mathbf{b}} \in \mathbb{R}^n, \widetilde{c} \in \mathbb{R}$ [illegible] $\widetilde{\mathbf{A}}$ [illegible] $\widetilde{\mathbf{A}} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ [illegible] $\mathbf{D} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ [illegible] $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ [illegible]

$$\min_{L^2 \leq \|\mathbf{s}\|^2 \leq U^2} \left\{ \mathbf{s}^T \widetilde{\mathbf{A}} \mathbf{s} - 2\widetilde{\mathbf{b}}^T \mathbf{s} + \widetilde{c} \right\}$$

[illegible] $\mathbf{s} = \mathbf{U}\mathbf{z}$ [illegible] $\mathbf{z} \in \mathbb{R}^n$ [illegible]

$$z_j = \frac{f_j}{\lambda_j - \eta^* - \xi^*}, \quad j = 1, 2, \ldots, n,$$

[illegible] (η^*, ξ^*) [illegible]

$$(\eta^*, \xi^*) = \begin{cases} (\bar{\eta}, 0) & \text{[illegible] } [\lambda_n > 0 \text{ [illegible] } \quad h(\bar{\eta}, 0) > h(0, \bar{\xi})] \text{ [illegible] } \lambda_n \leq 0, \\ (0, \bar{\xi}) & \text{[illegible] } [\lambda_n > 0 \text{ [illegible] } \quad h(\bar{\eta}, 0) \leq h(0, \bar{\xi})], \end{cases}$$

[illegible] $\bar{\eta}$ [illegible] $\bar{\xi}$ [illegible] (D1) [illegible] (D2) [illegible]

As was already mentioned, solving problems (D1) and (D2) is an easy task; to demonstrate this fact, let us consider the solution of (D1) in the case $\lambda_n \leq 0$ (all other instances can be similarly treated). In this case, (D1) takes the following form:

$$\max_{\eta < \lambda_n} \left\{ -\sum_{j=1}^{n} \frac{f_j^2}{\lambda_j - \eta} + \eta U^2 + \widetilde{c} \right\}.$$

Since $h_1(\eta) = h(\eta, 0)$ is continuous and strictly concave for $\eta < \lambda_n$ and also satisfies

$$\lim_{\eta \to -\infty} h_1(\eta) = -\infty, \ \lim_{\eta \to \lambda_n^-} h_1(\eta) = -\infty,$$

we conclude that the maximum is obtained at a unique point $\eta < \lambda_n$ that satisfies $h_1'(\eta) = 0$. Therefore, in this case we need to find the unique root of the following so-called secular equation [17]:

$$\eta < \lambda_n, \quad \mathcal{G}(\eta) = U^2, \tag{20}$$

where

$$\mathcal{G}(\eta) \equiv \sum_{j=1}^{n} \frac{f_j^2}{(\eta - \lambda_j)^2}. \tag{21}$$

Finding the unique root, which lies to the left of λ_n, of the secular equation (20) is a well-studied problem (see, e.g., [17, 7]). Specifically, Melman [17] transforms the problem into the equivalent problem

$$\mathcal{G}^{-1/2}(\eta) = U^{-1} \tag{22}$$

for which Newton's method exhibits [illegible]. The algorithm is as follows.

ALGORITHM SEC.

Input: $(\mathbf{f}, \Lambda, U)$, where $\mathbf{f} \in \mathbb{R}^n$, $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and $U > 0$.

Output: $\eta^* < \lambda_n$ that satisfies $|\mathcal{G}(\eta^*) - U^2| < \epsilon_2$, where $\mathcal{G}$ is defined in (21).

Initial step: $\eta_0 = \lambda_n - \epsilon_1$.

General step: for every $k \geq 0$,

$$\eta_{k+1} = \eta_k + 2\frac{\mathcal{G}^{-1/2}(\eta_k) - U^{-1}}{\mathcal{G}^{-3/2}(\eta_k)\mathcal{G}'(\eta_k)}.$$

Stopping rule: Stop at the first iteration k^* that satisfies $|\mathcal{G}(\eta_{k^*}) - U^2| < \epsilon_2$. Set $\eta^* = \eta_{k^*}$.

In our implementation the tolerance parameters ϵ_1 and ϵ_2 take the values $\epsilon_1 = 10^{-4}, \epsilon_2 = 10^{-15}$. Melman's algorithm solves the secular equation very fast (typically 5 or 6 iterations suffice to achieve 15 digit accuracy independently of n).

[illegible]. To demonstrate the rate of convergence of algorithm SEC we consider problem (20) with $n = 100, \lambda_i = i, f_i = 1$ $(i = 1, 2, \ldots, 100)$, and $U = 1$. We compare algorithm SEC with a simple bisection algorithm with initial interval $[-100, \lambda_n]$ and an identical stopping criteria as the one of algorithm SEC.

TABLE 1
Quadratic rate of convergence of Melman's algorithm.

Iteration	$\mathcal{G}(\eta_k) - U^2$	
	Bisection	SEC
1	−0.9866	1.0e+8
2	−0.9676	0.6349
3	−0.9264	0.0365
4	−0.8378	0.0002
5	−0.6375	1.19e-008
6	−0.1358	−1.11e-016

From Table 1 it is clear that the algorithm exhibits quadratic rate of convergence right from the very first iteration. The bisection algorithm terminated in this example after 55 iterations.

The dominant computational effort when solving the subproblem in the case $\mathcal{F} = \mathcal{F}_1$ are (i) the calculation of the matrices $\mathbf{T}^{1/2}, \mathbf{T}^{-1/2}$ and (ii) the spectral decomposition of the matrix $\widetilde{\mathbf{A}}$. Each requires a computational effort of $O(n^3)$. By Proposition 3.1, the schematic algorithm requires solving $O(\log \epsilon^{-1})$ subproblems in order to generate a ϵ-global optimal solution. We thus conclude that the overall computational effort of the schematic algorithm is $O(n^3 \log \epsilon^{-1})$.

4.2. Solving the subproblem in the case $\mathcal{F} = \mathcal{F}_2$. Here we consider problem (9) in the case where the feasible set is $\mathcal{F}_2 = \{\mathbf{x} : \mathbf{x}^T\mathbf{B}\mathbf{x} \leq U^2\}$, where $\mathbf{B}$ is positive semidefinite but not positive definite. Thus, the subproblem in step 2 of the schematic algorithm under consideration here is

$$\beta^* = \min_{\mathbf{x}^T\mathbf{B}\mathbf{x} \leq U^2} \{\mathbf{x}^T\mathbf{A}\mathbf{x} - 2\mathbf{b}^T\mathbf{x} + c\}, \tag{23}$$

where

$$\mathbf{A} = \mathbf{A}_1 - \alpha\mathbf{A}_2, \quad \mathbf{b} = \mathbf{b}_1 - \alpha\mathbf{b}_2, \quad c = c_1 - \alpha c_2.$$

Notice that since $\mathbf{B}$ is singular the feasible set $\mathcal{F}_2$ is not compact, and therefore the solution of the subproblem (23) might be $-\infty$. This issue is addressed in the following.

LEMMA 4.2. [illegible] $\mathbf{A} \in \mathbb{R}^{n\times n}$ [illegible] $\mathbf{b} \in \mathbb{R}^n$ $c \in \mathbb{R}$ $U > 0$ [illegible] $\mathbf{B} \in \mathbb{R}^{n\times n}$ [illegible]

1. [illegible] $\lambda \geq 0$ [illegible] $\mathbf{A} + \lambda\mathbf{B} \succ \mathbf{0}$ [illegible] (23) [illegible] $\beta^* > -\infty$.
2. [illegible] $\lambda \geq 0$ [illegible] $\mathbf{A} + \lambda\mathbf{B} \succeq \mathbf{0}$ [illegible] (23) [illegible]

[illegible]. The optimal value β^* of problem (23) is finite if and only if the following statement is true:

$$\exists \mu \in \mathbb{R}, \quad \mathbf{x}^T\mathbf{B}\mathbf{x} \leq U^2 \Rightarrow \mathbf{x}^T\mathbf{A}\mathbf{x} + 2\mathbf{b}^T\mathbf{x} + c \geq \mu, \tag{24}$$

which by the S-lemma (see Lemma A.1 in the appendix) is equivalent to

$$\exists \mu \in \mathbb{R}, \lambda \in \mathbb{R}_+, \quad \begin{pmatrix} \mathbf{A} & -\mathbf{b} \\ -\mathbf{b}^T & c-\mu \end{pmatrix} \succeq \lambda \begin{pmatrix} -\mathbf{B} & \mathbf{0} \\ \mathbf{0} & U^2 \end{pmatrix}$$

and which can also be written as

$$\exists \mu \in \mathbb{R}, \lambda \in \mathbb{R}_+, \quad \begin{pmatrix} \mathbf{A}+\lambda\mathbf{B} & -\mathbf{b} \\ -\mathbf{b}^T & c-\mu-U^2 \end{pmatrix} \succeq \mathbf{0}. \tag{25}$$

Since a necessary condition for the validity of (25) is that there exists a $\lambda \geq 0$ such that $\mathbf{A} + \lambda\mathbf{B} \succeq \mathbf{0}$, we conclude that the second statement of the lemma is proven. Moreover, if there exists a $\lambda_0 \geq 0$ such that $\mathbf{A} + \lambda_0\mathbf{B} \succ \mathbf{0}$, then taking $\mu_0 < c - U^2 - \mathbf{b}^T(\mathbf{A}+\lambda_0\mathbf{B})^{-1}\mathbf{b}$ we have by Schur's complement (Lemma A.2) that the linear matrix inequality (LMI) (25) is satisfied for $\lambda = \lambda_0$ and $\mu = \mu_0$, and therefore $\beta^* > -\infty$ and the first statement of the lemma is proven. □

Notice that the only case not covered by Lemma 4.2 is the case where there is a $\lambda \geq 0$ such that $\mathbf{A} + \lambda\mathbf{B} \succeq \mathbf{0}$ but there does not exist a $\lambda \geq 0$ such that $\mathbf{A} + \lambda\mathbf{B} \succ \mathbf{0}$. Later, we will see that we can ignore this case.

In the next result we find equivalent conditions for the finiteness of the minimization problem (23) that can be easily checked and analyzed.

LEMMA 4.3. [illegible] $\mathbf{A} \in \mathbb{R}^{n\times n}$ [illegible] $\mathbf{B} \in \mathbb{R}^{n\times n}$ [illegible] r [illegible] $\mathbf{F}$ [illegible] $n \times (n-r)$ [illegible] $\mathbf{B}$ [illegible]

1. [illegible] $\lambda \geq 0$ [illegible] $\mathbf{A} + \lambda\mathbf{B} \succ \mathbf{0}$
2. $\mathbf{F}^T\mathbf{A}\mathbf{F} \succ \mathbf{0}$

[illegible]. First, since $\mathbf{B} \succeq \mathbf{0}$, statement 1 is equivalent to the same statement without the sign constraint on λ:

$$\exists \lambda \in \mathbb{R}, \quad \mathbf{A} + \lambda\mathbf{B} \succ \mathbf{0}.$$

By Finsler's theorem (see Theorem A.1 in the appendix), this condition is equivalent to the following statement:

$$\mathbf{x}^T\mathbf{A}\mathbf{x} > 0 \quad \text{for every } \mathbf{x} \neq \mathbf{0} \text{ such that } \mathbf{x}^T\mathbf{B}\mathbf{x} = 0. \tag{26}$$

Now, since $\mathbf{B} \succeq \mathbf{0}$, we have that $\mathbf{x}^T\mathbf{B}\mathbf{x} = 0$ is equivalent to $\mathbf{x} \in \text{Null}(\mathbf{B})$. Thus, (26) is equivalent to

$$\mathbf{x}^T\mathbf{A}\mathbf{x} > 0 \quad \text{for every } \mathbf{x} \neq 0 \text{ such that } \mathbf{x} \in \text{Null}(\mathbf{B}),$$

which is equivalent to saying that $\mathbf{F}^T\mathbf{A}\mathbf{F} \succ \mathbf{0}$. ∎

A direct consequence of Lemmas 4.3 and 4.2 is that if

$$\mathbf{F}^T\mathbf{A}\mathbf{F} \succ \mathbf{0}, \tag{27}$$

then $\beta^* > -\infty$ and if $\mathbf{F}^T\mathbf{A}\mathbf{F}$ is not positive semidefinite (i.e., has at least one negative eigenvector), then $\beta^* = -\infty$. In the case where condition (27) is satisfied we can simultaneously diagonalize $\mathbf{A}$ and $\mathbf{B}$ (see Appendix B), and therefore we can continue with the hidden convexity argument.

Let $\mathbf{C}$ be a nonsingular matrix that simultaneously diagonalizes $\mathbf{A}$ and $\mathbf{B}$:

$$\mathbf{C}^T\mathbf{B}\mathbf{C} = \operatorname{diag}(\underbrace{1,1,\ldots,1}_{r \text{ times}}, \underbrace{0,0,\ldots,0}_{n-r \text{ times}}),$$
$$\mathbf{C}^T\mathbf{A}\mathbf{C} = \operatorname{diag}(\lambda_1,\lambda_2,\ldots,\lambda_r, \underbrace{1,1,\ldots,1}_{n-r \text{ times}}),$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$ (see Appendix B for details). Making the change of variables $\mathbf{x} = \mathbf{C}\mathbf{z}$ we obtain that (23) is equivalent to

$$\min\left\{\sum_{j=1}^{r}\lambda_j z_j^2 + \sum_{j=r+1}^{n} z_j^2 - 2\sum_{j=1}^{n} f_j z_j + c : \sum_{j=1}^{r} z_j^2 \leq U^2\right\}, \tag{28}$$

where $\mathbf{f} = \mathbf{C}^T\mathbf{b}$. The same argument as in Lemma 4.1 shows that we can make the change of variables

$$z_j = \operatorname{sign}(f_j)\sqrt{v_j}, \quad j = 1,2,\ldots,n,$$

where $v_j \geq 0$. We obtain the following equivalent convex optimization problem:

$$\min_{v_j \geq 0}\left\{\sum_{j=1}^{r}\left(\lambda_j v_j - 2|f_j|\sqrt{v_j}\right) + \sum_{j=r+1}^{n}\left(v_j - 2|f_j|\sqrt{v_j}\right) + c : \sum_{j=1}^{r} v_j \leq U^2\right\}. \tag{29}$$

To develop the dual problem of (29), we assign a nonpositive multiplier λ to the linear inequality constraint $-\sum_{j=1}^{r} v_j + U^2 \geq 0$ and form the Lagrangian of (29) given by

$$L(\mathbf{v},\eta,\xi) = \sum_{j=1}^{r}\left(\lambda_j v_j - 2|f_j|\sqrt{v_j}\right) + \sum_{j=r+1}^{n}\left(v_j - 2|f_j|\sqrt{v_j}\right) - \lambda\left(\sum_{j=1}^{r} v_j - U^2\right) + c$$
$$= \sum_{j=1}^{r}\left((\lambda_j - \lambda)v_j - 2|f_j|\sqrt{v_j}\right) + \sum_{j=r+1}^{n}\left(v_j - 2|f_j|\sqrt{v_j}\right) + \lambda U^2 + c. \tag{30}$$

Differentiating (18) with respect to v_j and equating to zero, we obtain

$$v_j = \frac{f_j^2}{(\lambda_j - \lambda)^2}, \quad j = 1,2,\ldots,r,$$
$$v_j = f_j^2, \quad j = r+1,\ldots,n,$$

subject to the condition $\lambda \leq \min\{\lambda_n, 0\}$. Thus, the dual objective function is given by

$$h(\lambda) = \inf_{v_j \geq 0} L(\mathbf{v}, \eta, \xi) = \begin{cases} -\sum_{j=1}^{r} \frac{f_j^2}{\lambda_j - \lambda} + \lambda U^2 + d, & \lambda < \min\{\lambda_r, 0\}, \\ -\infty & \text{otherwise}, \end{cases}$$

where $d = c - \sum_{j=r+1}^{n} f_j^2$. The dual problem of (29) is therefore

$$(\mathrm{D}): \qquad \max_{\lambda \leq \min\{\lambda_r, 0\}} h(\lambda).$$

From duality theory for convex optimization problems we have that [19, 3]

$$val(\mathrm{P}) = val(\mathrm{D}).$$

The solution of (D) involves the solution of a single secular equation of the form (20). We summarize the above discussion in Theorem 4.2.

THEOREM 4.2. [illegible] $\mathbf{A} \in \mathbb{R}^{n \times n}$ [illegible] $\mathbf{B} \in \mathbb{R}^{n \times n}$ [illegible] r $\mathbf{b} \in \mathbb{R}^n$ [illegible] $c \in \mathbb{R}$ [illegible] $\mathbf{F}^T \mathbf{A} \mathbf{F} \succ \mathbf{0}$ [illegible] $\mathbf{F}$ [illegible] $n \times (n-r)$ [illegible] $\mathbf{B}$ [illegible] $\mathbf{C}$ [illegible]

$$\mathbf{C}^T \mathbf{B} \mathbf{C} = \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad \mathbf{C}^T \mathbf{A} \mathbf{C} = \begin{pmatrix} \Lambda & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n-r} \end{pmatrix},$$

[illegible] $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r)$ [illegible] $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$ [illegible]

$$\min_{\mathbf{x}^T \mathbf{B} \mathbf{x} \leq U^2} \left\{ \mathbf{x}^T \mathbf{A} \mathbf{x} - 2\mathbf{b}^T \mathbf{x} + c \right\}$$

[illegible] $\mathbf{x} = \mathbf{C}\mathbf{z}$ [illegible] $\mathbf{z} \in \mathbb{R}^n$ [illegible]

$$z_j = \begin{cases} \frac{f_j}{\lambda_j - \lambda}, & j = 1, 2, \ldots, r, \\ f_j, & j = r+1, \ldots, n \end{cases} \qquad (\mathbf{f} = \mathbf{C}^T \mathbf{b})$$

[illegible] λ [illegible]

$$\max_{\lambda \leq \min\{\lambda_r, 0\}} \left\{ -\sum_{j=1}^{r} \frac{f_j^2}{\lambda_j - \lambda} + \lambda U^2 \right\},$$

[illegible] (20).

We will impose an additional assumption on the quadratic function $f_2(\mathbf{x})$.

[illegible] 2. f_2 is a strongly convex function (i.e., $\mathbf{A}_2 \succ \mathbf{0}$).

Note that Assumption 2 is readily satisfied by the RTLS problem (6). Recall that in the schematic algorithm $\mathbf{A} = \mathbf{A}_1 - \alpha \mathbf{A}_2$, so (27) is equivalent to

$$\mathbf{F}^T \mathbf{A}_1 \mathbf{F} - \alpha \mathbf{F}^T \mathbf{A}_2 \mathbf{F} \succ \mathbf{0}. \tag{31}$$

$\mathbf{F}$ is full column rank and, by Assumption 2, we have that $\mathbf{A}_2$ is positive definite, and as a consequence $\mathbf{F}^T \mathbf{A}_2 \mathbf{F}$ is also positive definite. Multiplying (31) from the right and left by $\mathbf{Q} = (\mathbf{F}^T \mathbf{A}_2 \mathbf{F})^{-1/2}$, we obtain the following equivalent LMI:

$$\mathbf{Q}(\mathbf{F}^T \mathbf{A}_1 \mathbf{F})\mathbf{Q} - \alpha \mathbf{I} \succ \mathbf{0}.$$

The last LMI is equivalent to $\alpha < \lambda_{\min}(\mathbf{Q}(\mathbf{F}^T\mathbf{A}_1\mathbf{F})\mathbf{Q})$. We summarize this in the following proposition.

PROPOSITION 4.2. [illegible] $\bar{\alpha} = \lambda_{\min}(\mathbf{Q}(\mathbf{F}^T\mathbf{A}_1\mathbf{F})\mathbf{Q})$ [illegible] $\mathbf{Q} = (\mathbf{F}^T\mathbf{A}_2\mathbf{F})^{-1/2}$ [illegible] (23) [illegible] $\alpha < \bar{\alpha}$ [illegible] $-\infty$, $\alpha > \bar{\alpha}$

$\bar{\alpha}$ is of course an upper bound for the minimal value of the original problem (1), and thus, in the schematic algorithm, we will always take an upper bound M that is of most $\bar{\alpha}$. We therefore conclude that throughout the schematic algorithm, we need to consider only subproblems with a finite minimum that satisfies (27).

A similar argument to the one given in the case $\mathcal{F} = \mathcal{F}_1$ shows that in the case $\mathcal{F} = \mathcal{F}_2$ as well, the algorithm produces an ϵ-global optimal solution in a computational effort of $O(n^3 \log \epsilon^{-1})$.

4.3. Finding the bounds. In this section we present some suggestions for the lower and upper bounds m and M of the schematic algorithm. In the special case of the original RTLS problem, simpler bounds are derived in section 5.

4.3.1. The case $\mathcal{F} = \mathcal{F}_1$. In this case the constraint is given by $L^2 \le \mathbf{x}^T\mathbf{T}\mathbf{x} \le U^2$. From this it follows that $\|\mathbf{x}\|^2 \le U^2/\lambda_{\min}(\mathbf{T})$. We can therefore bound the objective function of problem (1) as follows:

$$\left|\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right| = \left|\frac{\mathbf{x}^T\mathbf{A}_1\mathbf{x}^T - 2\mathbf{b}_1^T\mathbf{x} + c_1}{\mathbf{x}^T\mathbf{A}_2\mathbf{x} - 2\mathbf{b}_2^T\mathbf{x} + c_2}\right| \le \frac{1}{N}\left|\mathbf{x}^T\mathbf{A}_1\mathbf{x}^T - 2\mathbf{b}_1^T\mathbf{x} + c_1\right|$$
$$\le \frac{1}{N}(|\mathbf{x}^T\mathbf{A}_1\mathbf{x}^T| + |2\mathbf{b}_1^T\mathbf{x}| + |c_1|) \le \frac{1}{N}\left(U^2\frac{\lambda_{\max}(\mathbf{A}_1)}{\lambda_{\min}(\mathbf{T})} + 2\frac{\|\mathbf{b}_1\|U}{\sqrt{\lambda_{\min}(\mathbf{T})}} + |c_1|\right).$$

Thus, we can choose m and M to be

$$M = \frac{1}{N}\left(U^2\frac{\lambda_{\max}(\mathbf{A}_1)}{\lambda_{\min}(\mathbf{T})} + 2\frac{\|\mathbf{b}_1\|U}{\sqrt{\lambda_{\min}(\mathbf{T})}} + |c_1|\right), \quad m = -M.$$

The only element in the definition of m and M which is not given explicitly is the positive number N, defined in Assumption 1. For the RTLS problem, where $f_2(\mathbf{x}) = \|\mathbf{x}\|^2 + 1$, we can take N to be equal to 1. Also, for other problems we can define N to be the optimal value of the minimization problem $\min_{L \le \|\mathbf{x}\|_{\mathbf{T}} \le U}\{\mathbf{x}^T\mathbf{A}_2\mathbf{x} - 2\mathbf{b}_2^T\mathbf{x} + c_2\}$.

4.3.2. The case $\mathcal{F} = \mathcal{F}_2$. In this case the constraint is given by $\mathbf{x}^T\mathbf{B}\mathbf{x} \le U^2$, where $\mathbf{B}$ is a positive semidefinite matrix. We consider the case where both Assumptions 1 and 2 hold true and that f_2 is bounded below in $\mathbb{R}^n$. The upper bound can be taken as $M = \bar{\alpha}$, where $\bar{\alpha}$ is given in Proposition 4.2. To find a lower bound m, we first make the change of variables $\mathbf{z} = \mathbf{x} - \mathbf{A}_2^{-1}\mathbf{b}_2$ resulting with the following form of the objective function:

$$\frac{\mathbf{z}^T\mathbf{A}_1\mathbf{z} - 2\mathbf{e}^T\mathbf{z} + f}{\mathbf{z}^T\mathbf{A}_2\mathbf{z} + d}, \tag{32}$$

where $d = c_2 - \mathbf{b}_2^T\mathbf{A}_2^{-1}\mathbf{b}_2 > 0$, $\mathbf{e} = \mathbf{b}_1 - \mathbf{A}_1\mathbf{A}_2^{-1}\mathbf{b}_2$, and $f = c_1 + \mathbf{b}_2^T\mathbf{A}_2^{-1}\mathbf{A}_1\mathbf{A}_2^{-1}\mathbf{b}_2 - 2\mathbf{b}_1^T\mathbf{A}_2^{-1}\mathbf{b}_2$.

The unconstrained minimum of the last expression (32) is a lower bound on the optimal value, and we can lower bound it using a relaxation technique.

$$\min_{\mathbf{z}} \frac{\mathbf{z}^T\mathbf{A}_1\mathbf{z} - 2\mathbf{e}^T\mathbf{z} + f}{\mathbf{z}^T\mathbf{A}_2\mathbf{z} + d} \overset{\mathbf{w}=\mathbf{A}_2^{1/2}\mathbf{z}}{=} \min_{\mathbf{w}} \frac{\mathbf{w}^T\mathbf{A}_2^{-1/2}\mathbf{A}_1\mathbf{A}_2^{-1/2}\mathbf{w} - 2\mathbf{e}^T\mathbf{A}_2^{-1/2}\mathbf{w} + f}{\|\mathbf{w}\|^2 + d}$$

$$
\begin{aligned}
&= \min_{\mathbf{w},t=\sqrt{d}} \frac{\mathbf{w}^T\mathbf{A}_2^{-1/2}\mathbf{A}_1\mathbf{A}_2^{-1/2}\mathbf{w} - \frac{2}{\sqrt{d}}\mathbf{e}^T\mathbf{A}_2^{-1/2}\mathbf{w}t + \frac{f}{d}t^2}{\|\mathbf{w}\|^2 + t^2} \\
&\geq \min_{\mathbf{w},t} \frac{\mathbf{w}^T\mathbf{A}_2^{-1/2}\mathbf{A}_1\mathbf{A}_2^{-1/2}\mathbf{w} - \frac{2}{\sqrt{d}}\mathbf{e}^T\mathbf{A}_2^{-1/2}\mathbf{w}t + \frac{f}{d}t^2}{\|\mathbf{w}\|^2 + t^2} \\
&= \lambda_{\min}\begin{pmatrix} \mathbf{A}_2^{-1/2}\mathbf{A}_1\mathbf{A}_2^{-1/2} & \frac{1}{\sqrt{d}}\mathbf{A}_2^{-1/2}\mathbf{e} \\ \frac{1}{\sqrt{d}}\mathbf{e}^T\mathbf{A}_2^{-1/2} & \frac{f}{d} \end{pmatrix}
\end{aligned}
$$

Thus, we can take

$$
m = \lambda_{\min}\begin{pmatrix} \mathbf{A}_2^{-1/2}\mathbf{A}_1\mathbf{A}_2^{-1/2} & \frac{1}{\sqrt{d}}\mathbf{A}_2^{-1/2}\mathbf{e} \\ \frac{1}{\sqrt{d}}\mathbf{e}^T\mathbf{A}_2^{-1/2} & \frac{f}{d} \end{pmatrix}.
$$

5. A detailed algorithm for the RTLS problem. In this section we use the results obtained so far to write in full details the schematic algorithm of section 3 as applied to the RTLS problem:

$$
\min_{\mathbf{x}} \left\{ f(\mathbf{x}) \equiv \frac{\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2}{\|\mathbf{x}\|^2+1} : \|\mathbf{L}\mathbf{x}\| \leq U \right\}. \tag{33}
$$

We call this algorithm RTLSC.

The RTLSC algorithm solves at each iteration a subproblem of the form

$$
\min\left\{\mathbf{x}^T\mathbf{Q}\mathbf{x} - 2\mathbf{d}^T\mathbf{x} : \|\mathbf{L}\mathbf{x}\| \leq U\right\}. \tag{34}
$$

The detailed algorithm SUBP for solving the latter problem is explicitly written below. It invokes three procedures:

- SDG—an algorithm for simultaneous diagonalization of two matrices, one of which is positive definite (see Appendix B).
- SDGP—an algorithm for simultaneous diagonalization of two matrices, one of which is positive semidefinite (see Appendix B).
- SEC—Melman's algorithm for solving secular equations given in section 4.

ALGORITHM SUBP.

Input: $(\mathbf{Q},\mathbf{d},\mathbf{L},U,\mathbf{F})$, where $\mathbf{Q} \in \mathbb{R}^{n\times n}$ is a symmetric matrix, $\mathbf{d} \in \mathbb{R}^n, \mathbf{L} \in \mathbb{R}^{r\times n}(r \leq n)$ is a full rank matrix, $U > 0$, and $\mathbf{F} \in \mathbb{R}^{n\times(n-r)}$ is a matrix whose columns are an orthogonal basis for the null space of $\mathbf{L}$.

Output: $(\mathbf{x}^*,\mu)$. $\mathbf{x}^*$ is an optimal solution to problem (34) and μ is the corresponding optimal value.

1. **If** $r < n$, then call algorithm SDG with input $(\mathbf{A},\mathbf{L}^T\mathbf{L},\mathbf{F})$ and obtain an output $(\mathbf{C},\Lambda)$. **Else** call algorithm SDGP with input $(\mathbf{A},\mathbf{L}^T\mathbf{L})$ and obtain an output $(\mathbf{C},\Lambda)$.
2. Set $\mathbf{f} = \mathbf{C}^T\mathbf{d}$.
3. **If** $\lambda_r > 0$ and $\sum_{j=1}^r \frac{f_j^2}{\lambda_j^2} < U^2$, then set $\lambda^* = 0$. **Else** call algorithm SEC with input $(\mathbf{f},\Lambda,U)$ and obtain an output λ^*.
4. Let $v_j = \frac{f_j}{\lambda_j-\lambda^*}, j = 1,\ldots,r$, and $v_j = f_j, j = r+1,\ldots,n$ $(\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_r))$.
5. Set $\mathbf{x}^* = \mathbf{C}\mathbf{v}$ and $\mu = (\mathbf{x}^*)^T\mathbf{Q}\mathbf{x}^* - 2\mathbf{f}^T\mathbf{x}^*$.

ALGORITHM RTLSC.

Input: $(\mathbf{A},\mathbf{b},\mathbf{L},U,ub,\epsilon)$, where $\mathbf{A} \in \mathbb{R}^{m\times n}(m \geq n), \mathbf{b} \in \mathbb{R}^m, \mathbf{L} \in \mathbb{R}^{r\times n}(r < n)$ has full row rank, $U > 0$, $ub > 0$ is an upper bound on the optimal function value, and $\epsilon > 0$ is a tolerance parameter.

Output: $\mathbf{x}^*$—an ϵ-optimal solution of problem (33).

1. Set $k = 0, lb_0 = 0, ub_0 = ub$.
2. Calculate a matrix $\mathbf{F} \in \mathbb{R}^{n\times(n-r)}$ whose columns are an orthogonal basis for the null space of $\mathbf{L}$.
3. **While** $ub_k - lb_k > \epsilon$, **do**
   (a) $\alpha_k = \frac{lb_k + ub_k}{2}$.
   (b) Call algorithm SUBP with input $(\mathbf{A}^T\mathbf{A} - \alpha_k\mathbf{I}, \mathbf{A}^T\mathbf{b}, \mathbf{L}, U)$ and obtain an output $(\mathbf{x}_k, \beta_k)$.
   (c) Calculate $f_k = f(\mathbf{x}_k)$.
   (d) **If** $\beta_k + \|\mathbf{b}\|^2 - \alpha_k > 0$, then

$$lb_{k+1} = \alpha_k, ub_{k+1} = \min\{ub_k, f_k\}, \tag{35}$$

   else

$$lb_{k+1} = lb_k, ub_{k+1} = \min\{\alpha_k, f_k\}. \tag{36}$$

   (e) Set $k \leftarrow k + 1$.
   End.
4. Define $\mathbf{x}^* = \mathbf{x}_m$, where m is chosen so that $f_m = \min\{f_0, f_1, \ldots, f_{k-1}\}$.

Choice of lower and upper bounds. In the case where $\mathbf{L}$ is square and nonsingular, the upper bound can be chosen as $ub = f(\tilde{\mathbf{x}})$, where $\tilde{\mathbf{x}}$ is any feasible point (such as $\mathbf{0}$). A tight upper bound can be obtained by choosing $\tilde{\mathbf{x}}$ as a solution of another method such as regularized least squares. In the rank deficient case, Proposition 4.2 implies that $\lambda_{\min}(\mathbf{F}^T\mathbf{A}\mathbf{F})$ is an upper bound on the optimal function value. Hence, an initial upper bound is given by $\min\{\lambda_{\min}(\mathbf{F}^T\mathbf{A}\mathbf{F}), f(\tilde{\mathbf{x}})\}$. This choice guarantees that all subproblems have a finite value.

5.1. Note that the update equations (35) and (36) for the upper bound ub_k are different from the naive implementation suggested in the schematic algorithm of section 3. The idea behind the revised update formulas is to incorporate the information gained at previous iterations in order to find better upper bounds. At each iteration we calculate a new feasible point $\mathbf{x}_k$, which induces a new upper bound $f_k \equiv f(\mathbf{x}_k)$ on the optimal function value. Thus, the update equation $ub_{k+1} = ub_k$ in the original schematic algorithm is converted to $ub_{k+1} = \min\{ub_k, f_k\}$. The following example demonstrates the advantage of using the new update equations.

. In this section we illustrate a single run of the RTLSC algorithm. We consider problem (33) with

$$n = 2,\ \mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix},\ b = \begin{pmatrix} 10 \\ 25 \end{pmatrix},\ \mathbf{L} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix},\ \rho = 10.$$

Table 2 describes the first six iterations of algorithm RTLSC. The initial upper bound ub_0 was chosen to be $f(\mathbf{0}) = \|\mathbf{b}\|^2 = 725$. Note that the decrease in the upper bound is very drastic at the first few iterations. The size of the interval $[lb_k, ub_k]$ decreases by a factor of 3000 between iteration 0 and iteration 1 (instead of a factor of 2 in the old update equations). The minimum value is equal to 0.047501 and is reached after only three iterations. This run is typical in the sense that usually the algorithm converges to a point after very few iterations.

6. Numerical examples. In order to test the performance of algorithm RTLSC, two problems from the "Regularization Tools" [13] are employed: a problem that arises from the discretization of the inverse Laplace transform and an image deblurring problem. The following algorithms are tested:

TABLE 2
Single run of algorithm RTLSC.

k (# of iterations)	lb_k	ub_k	α_k	f_k
0	0	725	362.5	0.240383
1	0	0.243083	0.121541	0.047513
2	0	0.047513	0.023756	0.085005
3	0.023756	0.047513	0.035634	0.047501
4	0.035634	0.047501	0.041567	0.047501
5	0.041567	0.047501	0.044534	0.047501
6	0.044534	0.047501	0.046017	0.047501

TABLE 3
Relative errors of various regularization solvers.

n	σ	RTLSC	QEP	RLS	GR	TTLS
20	1e-1	**5.9e-1**	**5.9e-1**	7.1e-1	4.6e+0	1.2e+0
	1e-2	**2.5e-1**	**2.5e-1**	**2.5e-1**	5.8e-1	8.1e-1
	1e-4	**7.4e-2**	**7.4e-2**	7.6e-2	9.7e-2	6.5e-1
100	1e-1	**2.2e-1**	**2.2e-1**	3.3e-1	2.3e+0	9.2e-1
	1e-2	**1.5e-1**	**1.5e-1**	1.7e-1	2.3e-1	7.0e-1
	1e-4	4.7e-2	4.7e-2	**2.9e-2**	3.1e-2	4.3e-1

- RLS—Regularized Least Squares. This is the solution to the problem

$$\min\{\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2 : \|\mathbf{L}\mathbf{x}\| \le \rho\},$$

  implemented in the function lsqi from [13].
- TTLS—Truncated Total Least Squares originating from [5] and implemented in the function ttls from [13].
- RTLSC—Our algorithm from section 5.
- QEP—Sima, Van Huffel, and Golub's solver for RTLS [20].
- GR—Guo and Renaut's eigenvalue method for RTLS [11] with the RLS solution as a starting vector.

6.1. Inverse Laplace transform. We consider the problem of estimating the function $f(t)$ from its given Laplace transform [21]:

$$\int_0^\infty e^{-st} f(t)dt = \frac{2}{(s+1/2)^3}.$$

By means of Gauss–Laguerre quadrature, the problem reduces to a linear system $\mathbf{A}\mathbf{x}=\mathbf{b}$. This system and its solution $\mathbf{x}_R$ are implemented in the function ilaplace(n,3) from [13]. The perturbed right-hand side is generated by

$$\widetilde{\mathbf{b}} = (\mathbf{A}+\sigma\mathbf{E})\mathbf{x}_R + \sigma\mathbf{e}, \tag{37}$$

where each component of $\mathbf{E}$ and $\mathbf{e}$ is generated from a standard normal distribution and σ runs thorough the values 1e-1, 1e-2, and 1e-4. The matrix $\mathbf{L}$ approximates the first-derivative operator implemented in the function get_l(n,1) from [13]. Two cases are tested: $m=n=20$ and $m=n=100$. Table 3 describes the relative error $\|\mathbf{x}-\mathbf{x}_R\|/\|\mathbf{x}_R\|$ averaged over 300 random realizations of $\mathbf{E}$ and $\mathbf{e}$.

The best results in each row are emphasized in boldface. The RTLSC and QEP methods give the best results in all but one case. The RLS also performed quite well. Note that the average relative error for the RTLSC and QEP solvers are equal. It

is interesting to note that not only the average was the same but in fact for all 1800 simulations of QEP and RTLSC, the results were the same. Incidentally, this provides an experimental evidence to the claim that QEP finds the global minimum, although such a theoretical claim was not proved in [20].

The CPU time in seconds of the three RTLS solvers averaged over 20 realizations of $\mathbf{E}$ and $\mathbf{e}$ is given in Table 4 (σ was fixed to be 1e-4). To make a fair comparison, we employed the same stopping rule for each of the methods: $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|/\|\mathbf{x}_k\| < 10^{-3}$.

TABLE 4
CPU time in seconds on a Pentium 4, 1.8Ghz.

n	RTLSC	QEP	GR
20	5.6e-2	5.3e-2	2e-1
50	1.9e-1	2.1e-1	1.2
100	2.27	2.4	23.2
200	3.2	3.1	112.3
1000	296	312	-

It is clear from Table 4 that RTLSC and QEP are significantly faster than GR. Moreover, RTLSC and QEP require more or less the same running time.

6.2. Image deblurring. We consider the problem of estimating a 32×32 two-dimensional image obtained from the sum of three harmonic oscillations:

$$x(z_1, z_2) = \sum_{l=1}^{3} a_i \cos(w_{l,1} z_1 + w_{l,2} z_2 + \phi_l), \quad \left(w_{l,i} = \frac{2\pi k_{l,i}}{n} \right), \ 1 \le z_1, z_2 \le 32,$$

where $k_{l,i} \in \mathbb{Z}^2$ (see Figure 1(A)). The specific values of the parameters are given in Table 5.

The image is blurred by atmospheric turbulence blur originating from [12] and implemented in the function blur(n,3,1) from [13].

The blurred image is generated by the relation (37) with $\sigma = 0.1$, which results in a highly noisy image (see Figure 1(B)).

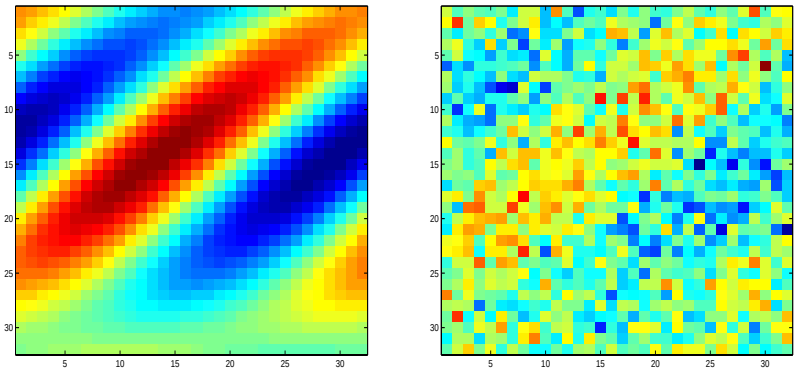
Choice of regularization matrix. We first ran algorithm RLS with standard regularization ($\mathbf{L} = \mathbf{I}$). The result is the poor image given in Figure 1(C). We then chose $\mathbf{L}$ as a discrete approximation of the Laplace operator [16] which is a two-dimensional convolution with the following mask:

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}.$$

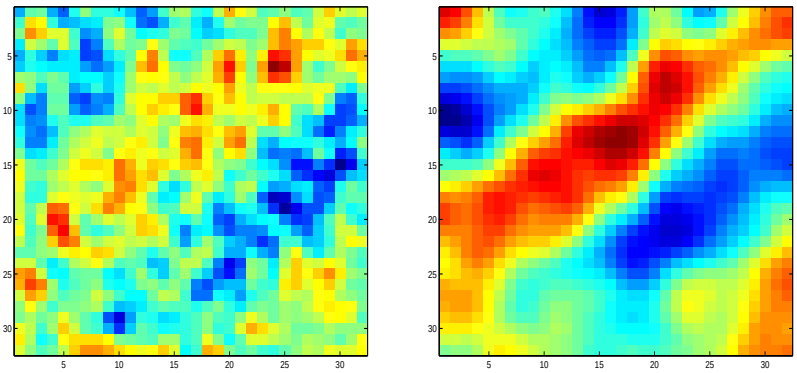
The above results demonstrate the importance of the choice of the regularization matrix. In the following experiments we use the nonstandard $\mathbf{L}$. The result for algorithm TTLS is given in Figure 1(E). A much improved result is obtained by our algorithm RTLSC (Figure 1(F)). Here again algorithm QEP gave the same result as algorithm RTLSC. Also, algorithm GR gave in this example the same image as RLS.

It is interesting to note that in this and many other examples algorithm RTLSC required only three iterations in order to produce quality reconstructions. As an illustration, Figure 2 shows the result of the first three iterations of algorithm RTLSC. The function values of the images generated in iterations 1, 2, and 3 are 2.0934, 1.5715, and 1.5566, respectively. The difference between the first and second iteration

(A) True Image (B) Observation



(C) RLS with $\mathbf{L} = \mathbf{I}$ (D) RLS with Laplace operator



(E) TTLS (F) RTLSC

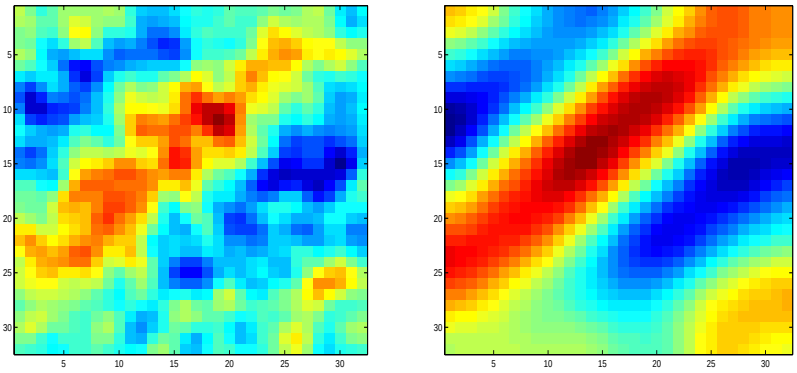


FIG. 1. *Results for different regularization solvers.*

is substantial. However, the image produced at the third iteration is almost identical to the image produced at the third iteration. Further iterations of RTLSC do not improve the image, although the function value reduces to the minimal value 1.5234.

TABLE 5
Image parameters.

l	a_l	$w_{l,1}$	$w_{l,2}$	ϕ_l
1	1.3936	0.1473	0.0982	5.8777
2	0.5579	0.0982	0.0982	5.7611
3	0.8529	0.0491	0.0982	2.5778

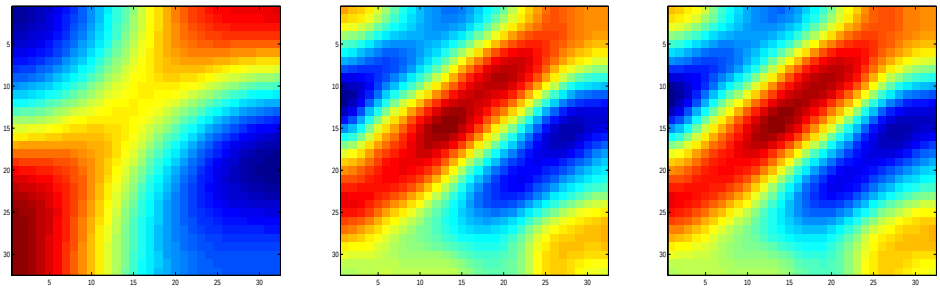


FIG. 2. *First three iterations of algorithm RTLSC.*

Appendix A. Known results.

LEMMA A.1 (S-lemma [1]). $\mathbf{A}$ $\mathbf{B}$ $n\times n$ $\mathbf{e},\mathbf{f}\in\mathbb{R}^n$ $g,h\in\mathbb{R}$

$$\mathbf{x}^T\mathbf{A}\mathbf{x}+2\mathbf{e}^T\mathbf{x}+g\geq 0 \tag{38}$$

$\bar{\mathbf{x}}$ $\bar{\mathbf{x}}^T\mathbf{A}\bar{\mathbf{x}}+2\mathbf{e}^T\bar{\mathbf{x}}+g>0$

$$\mathbf{x}^T\mathbf{B}\mathbf{x}+2\mathbf{f}^T\mathbf{x}+h\geq 0 \tag{39}$$

(38) λ

$$\begin{pmatrix}\mathbf{B} & \mathbf{f}\\ \mathbf{f}^T & h\end{pmatrix}\succeq\lambda\begin{pmatrix}\mathbf{A} & \mathbf{e}\\ \mathbf{e}^T & g\end{pmatrix}.$$

LEMMA A.2 (Schur's complement [1]).

$$\mathbf{M}=\begin{pmatrix}\mathbf{A} & \mathbf{B}^T\\ \mathbf{B} & \mathbf{C}\end{pmatrix}$$

$\mathbf{C}\succ 0$ $\mathbf{M}\succeq 0$ $\Delta_{\mathbf{C}}\geq 0$ $\Delta_{\mathbf{C}}$ $\mathbf{C}$ $\mathbf{M}$

$$\Delta_{\mathbf{C}}=\mathbf{A}-\mathbf{B}^T\mathbf{C}^{-1}\mathbf{B}.$$

THEOREM A.1 (Finsler's theorem [6]). $\mathbf{A}$ $\mathbf{B}$ $n\times n$

$$\mathbf{x}^T\mathbf{B}\mathbf{x}>0$$

$$\mathbf{x}^T\mathbf{A}\mathbf{x}=0$$

[illegible] $\alpha \in \mathbb{R}$ [illegible] $\mathbf{B} - \alpha\mathbf{A} \succ \mathbf{0}$

Appendix B. Algorithms for simultaneous diagonalization. In this section we recover an algorithm for the simultaneous diagonalization of an $n \times n$ symmetric matrix $\mathbf{A}$ and a positive semidefinite matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$ of rank $r(< n)$. We denote by $\mathbf{F}$ the $n \times (n-r)$ matrix whose columns are an orthogonal basis for the null space of $\mathbf{B}$ and assume that the condition

$$\mathbf{F}^T\mathbf{A}\mathbf{F} \succ \mathbf{0} \tag{40}$$

is satisfied, which implies that the matrices $\mathbf{A}$ and $\mathbf{B}$ are simultaneously diagonalizable by a nonsingular matrix. This fact follows directly from [18, Theorem 6.2.2]. Here we explicitly recover the algorithm that follows from [18] for the special case where (40) is satisfied.

Algorithm SDG.

Input: $(\mathbf{A}, \mathbf{B}, \mathbf{F})$, where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is a symmetric matrix, $\mathbf{B} \in \mathbb{R}^{n\times n}$ is a positive semidefinite of rank $r(r < n)$, and $\mathbf{F} \in \mathbb{R}^{n\times(n-r)}$ is a matrix whose columns are an orthogonal basis for the null space of $\mathbf{B}$.

Condition: $\mathbf{F}^T\mathbf{A}\mathbf{F} \succ \mathbf{0}$.

Output: $(\mathbf{C}, \Lambda)$. $\mathbf{C}$ is a nonsingular matrix and $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r)$ $(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r)$ is a diagonal matrix such that

$$\mathbf{C}^T\mathbf{B}\mathbf{C} = \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad \mathbf{C}^T\mathbf{A}\mathbf{C} = \begin{pmatrix} \Lambda & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n-r} \end{pmatrix}.$$

1. Find a full row rank $r \times n$ matrix $\mathbf{L}$ such that[1] $\mathbf{B} = \mathbf{L}^T\mathbf{L}$.
2. Define $\mathbf{M} = \mathbf{L}^T(\mathbf{L}\mathbf{L}^T)^{-1}$ ($\mathbf{M}$ is a right inverse of $\mathbf{L}$). We have $\mathbf{M}^T\mathbf{B}\mathbf{M} = \mathbf{I}_r$.
3. Define $\mathbf{S} = \begin{pmatrix} \mathbf{M} - \mathbf{F}(\mathbf{F}^T\mathbf{A}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{A}\mathbf{M}, & \mathbf{F} \end{pmatrix}$. We have

$$\mathbf{S}^T\mathbf{B}\mathbf{S} = \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \ \mathbf{S}^T\mathbf{A}\mathbf{S} = \begin{pmatrix} \mathbf{E} & \mathbf{0} \\ \mathbf{0} & \mathbf{F}^T\mathbf{A}\mathbf{F} \end{pmatrix},$$

   where $\mathbf{E}$ is an $r \times r$ symmetric matrix.
4. Find an $r \times r$ orthogonal matrix $\mathbf{Q}_1$ such that $\mathbf{Q}_1^T\mathbf{E}\mathbf{Q}_1 = \Lambda$, where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_r)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$.
5. Find an $(n-r) \times (n-r)$ matrix $\mathbf{Q}_2$ such that $\mathbf{Q}_2^T(\mathbf{F}^T\mathbf{A}\mathbf{F})\mathbf{Q}_2 = \mathbf{I}_{n-r}$ (this is possible since we assume that $\mathbf{F}^T\mathbf{A}\mathbf{F} \succ \mathbf{0}$).
6. Define

$$\mathbf{C} = \mathbf{S}\begin{pmatrix} \mathbf{Q}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2 \end{pmatrix}.$$

In the case where one of the matrices is positive definite, simultaneous diagonalization is always possible without any restrictions [14]. The procedure for simultaneous diagonalization in that case is much simpler and is given below.

Algorithm SDGP.

Input: $(\mathbf{A}, \mathbf{B})$, where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is a symmetric matrix and $\mathbf{B} \in \mathbb{R}^{n\times n}$ is a positive definite matrix.

Output: $(\mathbf{C}, \Lambda)$. $\mathbf{C}$ is a nonsingular matrix and $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ $(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n)$ is a diagonal matrix such that

$$\mathbf{C}^T\mathbf{B}\mathbf{C} = \mathbf{I}, \quad \mathbf{C}^T\mathbf{A}\mathbf{C} = \Lambda.$$

[1] This step can be done by, e.g., Cholesky's factorization. In some applications $\mathbf{B}$ is already given in that form.

1. Find a singular matrix $\mathbf{L}$ such that $\mathbf{B} = \mathbf{L}^T\mathbf{L}$.
2. Calculate the spectral decomposition of $(\mathbf{L}^T)^{-1}\mathbf{A}\mathbf{L}^{-1}$:

$$\mathbf{U}^T((\mathbf{L}^T)^{-1}\mathbf{A}\mathbf{L}^{-1})\mathbf{U} = \mathbf{D},$$

   where $\mathbf{U}$ is an orthogonal matrix, $\mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$.
3. Set $\mathbf{C} = \mathbf{L}^{-1}\mathbf{U}, \Lambda = \mathbf{D}$.

Acknowledgments. We thank the associate editor and two anonymous referees for their constructive comments.

REFERENCES

[1] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, MPS-SIAM Ser. Optim. 2, SIAM, Philadelphia, 2001.
[2] A. Ben-Tal and M. Teboulle, *Hidden convexity in some nonconvex quadratically constrained quadratic programming*, Math. Programming, 72 (1996), pp. 51–63.
[3] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, MA, 1999.
[4] W. Dinkelbach, *On nonlinear fractional programming*, Management Sci., 13 (1967), pp. 492–498.
[5] R. D. Fierro, G. H. Golub, P. C. Hansen, and D. P. O'Leary, *Regularization by truncated total least squares*, SIAM J. Sci. Comput., 18 (1997), pp. 1223–1241.
[6] P. Finsler, *Uber das Vorkommen definiter und semi-definiter Formen in scharen quadratische Formen*, Comment. Math. Helv., 9 (1937), pp. 188–192.
[7] W. Gander, G. H. Golub, and U. von Matt, *A constrained eigenvalue problem*, Linear Algebra Appl., 114/115 (1989), pp. 815–839.
[8] G. H. Golub, P. C. Hansen, and D. P. O'Leary, *Tikhonov regularization and total least squares*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 185–194.
[9] G. H. Golub and C. F. Van Loan, *An analysis of the total least squares problem*, SIAM J. Numer. Anal., 17 (1980), pp. 883–893.
[10] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.
[11] H. Guo and R. Renaut, *A regularized total least squares algorithm*, in Total Least Squares and Errors-in-Variables Modeling, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002, pp. 57–66.
[12] M. Hanke and P. C. Hansen, *Regularization methods for large-scale problems*, Surveys Math. Indust., 3 (1993), pp. 253–315.
[13] P. C. Hansen, *Regularization tools, a Matlab package for analysis of discrete regularization problems*, Numer. Algorithms, 6 (1994), pp. 1–35.
[14] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
[15] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, Frontiers Appl. Math. 9, SIAM, Philadelphia, 1991.
[16] A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice–Hall, Englewood Cliffs, NJ, 1989.
[17] A. Melman, *A unifying convergence analysis of second-order methods for secular equations*, Math. Comp., 66 (1997), pp. 333–344.
[18] C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and Its Applications*, John Wiley and Sons, New York, 1971.
[19] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[20] D. Sima, S. Van Huffel, and G. H. Golub, *Regularized total least squares based on quadratic eigenvalue problem solvers*, BIT, 44 (2004), pp. 793–812.
[21] J. M. Varah, *Pitfalls in the numerical solution of linear ill-posed problems*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 164–176.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 2, pp. 446–476

SMOOTHED ANALYSIS OF THE CONDITION NUMBERS AND GROWTH FACTORS OF MATRICES*

ARVIND SANKAR†, DANIEL A. SPIELMAN‡, AND SHANG-HUA TENG§

Abstract. Let $\bar{\boldsymbol{A}}$ be an arbitrary matrix and let $\boldsymbol{A}$ be a slight random perturbation of $\bar{\boldsymbol{A}}$. We prove that it is unlikely that $\boldsymbol{A}$ has a large condition number. Using this result, we prove that it is unlikely that $\boldsymbol{A}$ has large growth factor under Gaussian elimination without pivoting. By combining these results, we show that the smoothed precision necessary to solve $\boldsymbol{Ax} = \boldsymbol{b}$, for any $\boldsymbol{b}$, using Gaussian elimination without pivoting is logarithmic. Moreover, when $\bar{\boldsymbol{A}}$ is an all-zero square matrix, our results significantly improve the average-case analysis of Gaussian elimination without pivoting performed by Yeung and Chan (*SIAM J. Matrix Anal. Appl.*, 18 (1997), pp. 499–517).

Key words. smoothed analysis, condition number, Gaussian elimination, growth factor

AMS subject classifications. 15A18, 15A52, 65F05, 65F35

DOI. 10.1137/S0895479803436202

1. Introduction.

Spielman and Teng [ST04], introduced the smoothed analysis of algorithms to explain the success of algorithms and heuristics that could not be well understood through traditional worst-case and average-case analyses. Smoothed analysis is a hybrid of worst-case and average-case analyses in which one measures the maximum over inputs of the expected value of a measure of the performance of an algorithm on slight random perturbations of that input. For example, the smoothed complexity of an algorithm is the maximum over its inputs of the expected running time of the algorithm under slight perturbations of that input. If an algorithm has low smoothed complexity and its inputs are subject to noise, then it is unlikely that one will encounter an input on which the algorithm performs poorly. (See also the smoothed analysis homepage [Smo].)

Smoothed analysis is motivated by the existence of algorithms and heuristics that are known to work well in practice, but which are known to have poor worst-case performance. Average-case analysis was introduced in an attempt to explain the success of such heuristics. However, average-case analyses are often unsatisfying as the random inputs they consider may bear little resemblance to the inputs actually encountered in practice. Smoothed analysis attempts to overcome this objection by proving a bound that holds in every neighborhood of inputs.

In this paper, we prove that perturbations of arbitrary matrices are unlikely to have large condition numbers or large growth factors under Gaussian elimination without pivoting. As a consequence, we conclude that the smoothed precision necessary for Gaussian elimination is logarithmic. We obtain similar results for perturbations

*Received by the editors October 14, 2003; accepted for publication (in revised form) by D. P. O'Leary December 5, 2005; published electronically June 21, 2006.
http://www.siam.org/journals/simax/28-2/43620.html

†Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139 (nivedita@alum.mit.edu). Partially supported by NSF grant CCR-0112487.

‡Department of Computer Science, Yale University, New Haven, CT 06250 (spielman@cs.yale.edu). Partially supported by an Alfred P. Sloan Foundation Fellowship, and NSF grants CCR-0112487 and CCR-0324914.

§Department of Computer Science, Boston University, Boston, MA 02215 (steng@cs.bu.edu) and Akamai Technologies Inc., Cambridge, MA 02142. Partially supported by an Alfred P. Sloan Foundation Fellowship, NSF grants CCR-9972532 and CCR-0311430, and ITR CCR-0325630.

that affect only the nonzero and diagonal entries of symmetric matrices. We hope that these results will be a first step toward a smoothed analysis of Gaussian elimination with partial pivoting—an algorithm that is widely used in practice but known to have poor worst-case performance.

In the rest of this section, we recall the definitions of the condition numbers and growth factors of matrices, and review prior work on their average-case analysis. In section 3, we perform a smoothed analysis of the condition number of a matrix. In section 4, we use the results of section 3 to obtain a smoothed analysis of the growth factors of Gaussian elimination without pivoting. In section 5, we combine these results to obtain a smoothed bound on the precision needed by Gaussian elimination without pivoting. Definitions of zero-preserving perturbations and our results on perturbations that only affect the nonzero and diagonal entries of symmetric matrices appear in section 6. In the conclusion section, we explain how our results may be extended to larger families of perturbations, present some counter-examples, and suggest future directions for research. Other conjectures and open questions appear in the body of the paper.

The analysis in this paper requires many results from probability. Where reasonable, these have been deferred to the appendices.

1.1. Condition numbers and growth factors. We use the standard notation for the 1-, 2-, and ∞-norms of matrices and column vectors, and define

$$\|\boldsymbol{A}\|_{\max} = \max_{i,j} |\boldsymbol{A}_{i,j}| .$$

DEFINITION 1.1 (condition number). [illegible] $\boldsymbol{A}$ [illegible] $\boldsymbol{A}$[illegible]

$$\kappa(\boldsymbol{A}) = \|\boldsymbol{A}\|_2 \left\|\boldsymbol{A}^{-1}\right\|_2 .$$

The condition number measures how much the solution to a system $\boldsymbol{Ax} = \boldsymbol{b}$ changes as one makes slight changes to $\boldsymbol{A}$ and $\boldsymbol{b}$. A consequence is that if ones solves the linear system using fewer than $\log(\kappa(\boldsymbol{A}))$ bits of precision, one is likely to obtain a result far from a solution. For more information on the condition number of a matrix, we refer the reader to one of [GL83, TB97, Dem97].

The simplest and most often implemented method of solving linear systems is Gaussian elimination. Natural implementations of Gaussian elimination use $\mathcal{O}(n^3)$ arithmetic operations to solve a system of n linear equations in n variables. If the coefficients of these equations are specified using b bits, in the worst case it suffices to perform the elimination using $O(bn)$ bits of precision [GLS91]. This high precision may be necessary because the elimination may produce large intermediate entries [TB97]. However, in practice one usually obtains accurate answers using much less precision. In fact, it is rare to find an implementation of Gaussian elimination that uses anything more than double precision, and high-precision solvers are rarely used or needed in practice [TB97, TS90] (for example, LAPACK uses 64 bits [ABB+99]). One of the main results of this paper is that $\mathcal{O}(b + \log n)$ bits of precision usually suffice for Gaussian elimination in the smoothed analysis framework.

Since Wilkinson's seminal work [Wil61], it has been understood that it suffices to carry out Gaussian elimination with $b + \log_2(5n\kappa(\boldsymbol{A}) \|\boldsymbol{L}\|_\infty \|\boldsymbol{U}\|_\infty / \|\boldsymbol{A}\|_\infty + 3)$ bits of accuracy to obtain a solution that is accurate to b bits. In this formula, $\boldsymbol{L}$ and $\boldsymbol{U}$ are the LU-decomposition of $\boldsymbol{A}$; that is, $\boldsymbol{U}$ is the upper-triangular matrix and $\boldsymbol{L}$ is the lower-triangular matrix with 1s on the diagonal for which $\boldsymbol{A} = \boldsymbol{LU}$.

1.2. Prior work. The average-case behaviors of the condition numbers and growth factors of matrices have been studied both analytically and experimentally. In [Dem88], Demmel proved that it is unlikely that a Gaussian random matrix centered at the origin has large condition number. Demmel's bounds on the condition number were improved by Edelman [Ede88].

Average-case analysis of growth factors began with the experimental work of Trefethen and Schreiber [TS90], who found that Gaussian random matrices rarely have large growth factors under partial or full pivoting.

DEFINITION 1.2 (Gaussian matrix). $\boldsymbol{G}$ Gaussian random matrix σ^2 $\boldsymbol{G}$ 0 σ

Yeung and Chan [YC97] study the growth factors of Gaussian elimination without pivoting on Gaussian random matrices of variance 1. They define $\rho_{\boldsymbol{U}}$ and $\rho_{\boldsymbol{L}}$ by

$$\rho_{\boldsymbol{U}}(\boldsymbol{A}) = \|\boldsymbol{U}\|_\infty / \|\boldsymbol{A}\|_\infty, \quad \text{and} \quad \rho_{\boldsymbol{L}}(\boldsymbol{A}) = \|\boldsymbol{L}\|_\infty,$$

where $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$ is the LU-factorization of $\boldsymbol{A}$ obtained without pivoting. They prove the following theorem.

THEOREM 1.3 (Yeung-Chan). $c > 0$ $0 < b < 1$ $\boldsymbol{G}$ $n \times n$ 1 $\boldsymbol{G} = \boldsymbol{L}\boldsymbol{U}$ $\boldsymbol{G}$

$$\Pr[\rho_{\boldsymbol{L}}(\boldsymbol{G}) > x] \leq \frac{cn^3}{x},$$
$$\Pr[\rho_{\boldsymbol{U}}(\boldsymbol{G}) > x] \leq \min\left(\frac{cn^{7/2}}{x}, \frac{1}{n}\right) + \frac{cn^{5/2}}{x} + b^n.$$

As it is generally believed that partial pivoting is better than no pivoting, their result provides some intuition for the experimental results of Trefethen and Schreiber demonstrating that random matrices rarely have large growth factors under partial pivoting. However, we note that it is difficult to make this intuition rigorous as there are matrices $\boldsymbol{A}$ for which no pivoting has $\|\boldsymbol{L}\|_{\max}\|\boldsymbol{U}\|_{\max}/\|\boldsymbol{A}\|_{\max} = 2$ while partial pivoting has growth factor 2^{n-1}. (See also [Hig90].)

The running times of many numerical algorithms depend on the condition numbers of their inputs. For example, the number of iterations taken by the method of conjugate gradients can be bounded in terms of the square root of the condition number. Similarly, the running times of interior-point methods can be bounded in terms of condition numbers [Ren95]. Blum [Blu89] suggested that a complexity theory of numerical algorithms should be parameterized by the condition number of an input in addition to the input size. Smale [Sma97] proposed a complexity theory of numerical algorithms in which one:

1. proves a bound on the running time of an algorithm solving a problem in terms of its condition number, and then
2. proves that it is unlikely that a random problem instance has large condition number.

This program is analogous to the average-case complexity of theoretical computer science.

1.3. Our results. To better model the inputs that occur in practice, we propose replacing step 2 of Smale's program with

2′. proves that for every input instance it is unlikely that a slight random perturbation of that instance has large condition number.

That is, we propose to bound the smoothed value of the condition number. Our first result in this program is presented in section 3, where we improve upon Demmel's [Dem88] and Edelman's [Ede88] average-case results to show that a slight Gaussian perturbation of an arbitrary matrix is unlikely to have large condition number.

DEFINITION 1.4 (Gaussian perturbation). $\bar{\boldsymbol{A}}$ $n \times n$ $\boldsymbol{A}$ Gaussian perturbation $\bar{\boldsymbol{A}}$ σ^2 $\boldsymbol{A}$ $\boldsymbol{A} = \bar{\boldsymbol{A}} + \boldsymbol{G}$ $\boldsymbol{G}$ σ^2 $\boldsymbol{A}$ σ^2 $\bar{\boldsymbol{A}}$

In our smoothed analysis of the condition number, we consider an arbitrary $n \times n$ matrix $\bar{\boldsymbol{A}}$ of norm at most $\sqrt{n}$, and we bound the probability that $\kappa(\bar{\boldsymbol{A}} + \boldsymbol{G})$, the condition number of its Gaussian perturbation, is large, where $\boldsymbol{G}$ is a Gaussian random matrix of variance $\sigma^2 \leq 1$. We bound this probability in terms of σ and n. In contrast with the average-case analysis of Demmel and Edelman, our analysis can be interpreted as demonstrating that if there is a little bit of imprecision or noise in the entries of a matrix, then it is unlikely it is ill-conditioned. On the other hand, Edelman [Ede92] writes of random matrices:

> What is a mistake is to psychologically link a random matrix with the intuitive notion of a "typical" matrix or the vague concept of "any old matrix."

The reader might also be interested in recent work on the smoothed analysis of the condition numbers of linear programs [BD02, DST02, ST03].

In section 4, we use results from section 3 to perform a smoothed analysis of the growth factors of Gaussian elimination without pivoting. If one specializes our results to perturbations of an all-zero square matrix, then one obtains a bound on $\rho_{\boldsymbol{U}}$ that improves the bound obtained by Yeung and Chan by a factor of n and which agrees with their experimental observations. The result obtained for $\rho_{\boldsymbol{L}}$ also improves the bound of Yeung and Chan [YC97] by a factor of n. However, while Yeung and Chan compute the density functions of the distribution of the elements in $\boldsymbol{L}$ and $\boldsymbol{U}$, such precise estimates are not immediately available in our model. As a result, the techniques we develop are applicable to a wide variety of models of perturbations beyond the Gaussian. For example, one could use our techniques to obtain results of a similar nature if $\boldsymbol{G}$ were a matrix of random variables chosen uniformly in $[-1, 1]$. We comment further upon this in the conclusions section of the paper.

The less effect a perturbation has, the more meaningful the results of smoothed analysis are. As many matrices encountered in practice are sparse or have structure, it would be best to consider perturbations that respect their sparsity pattern or structure. Our first result in this direction appears in section 6, in which we consider the condition numbers and growth factors of perturbations of symmetric matrices that only alter their nonzero and diagonal elements. We prove results similar to those proved for dense perturbations of arbitrary matrices.

2. Notation and mathematical preliminaries. We use bold lower-case Roman letters such as $\boldsymbol{x}$, $\boldsymbol{a}$, $\boldsymbol{b}_j$ to denote vectors in $\mathbb{R}^?$. Whenever a vector, say $\boldsymbol{a} \in \mathbb{R}^n$ is present, its components will be denoted by lower-case Roman letters with subscripts, such as $a_1, \ldots, a_n$. Matrices are denoted by bold upper-case Roman letters such as $\boldsymbol{A}$ and scalars are denoted by lower-case roman letters. Indicator random variables and random event variables are denoted by upper-case Roman letters. Random variables

taking real values are denoted by upper-case Roman letters, except when they are components of a random vector or matrix.

The probability of an event A is written $\Pr[A]$, and the expectation of a variable X is written $\mathrm{E}[X]$. The indicator random variable for an event A is written $[A]$.

We write ln to denote the natural logarithm, base e, and explicitly write the base for all other logarithms.

For integers $a \leq b$, we let $a : b$ denote the set of integers $\{x : a \leq x \leq b\}$. For a matrix $\boldsymbol{A}$ we let $\boldsymbol{A}_{a:b,c:d}$ denote the submatrix of $\boldsymbol{A}$ indexed by rows in $a : b$ and columns in $c : d$.

We will bound many probabilities by applying the following proposition.

PROPOSITION 2.1 (minimum $\leq$ average $\leq$ maximum). $\mu(X,Y)$ X Y $\mu(X,Y)$ $A(X,Y)$ $F(X,Y)$

$$\min_X \Pr_Y [A(X,Y)] \leq \Pr_{X,Y} [A(X,Y)] \leq \max_X \Pr_Y [A(X,Y)], \quad \textit{and}$$
$$\min_X \mathrm{E}_Y [F(X,Y)] \leq \mathrm{E}_{X,Y} [F(X,Y)] \leq \max_X \mathrm{E}_Y [F(X,Y)],$$

Y $\mu(X,Y)$

We recall that a matrix $\boldsymbol{Q}$ is an orthonormal matrix if its inverse is equal to its transpose, that is, $\boldsymbol{Q}^T\boldsymbol{Q} = I$. In section 3 we will use the following proposition.

PROPOSITION 2.2 (orthonormal transformation of Gaussian). $\bar{\boldsymbol{A}}$ $\mathbb{R}^{n\times n}$ $\boldsymbol{Q}$ $\mathbb{R}^{n\times n}$ $\boldsymbol{A}$ $\bar{\boldsymbol{A}}$ σ^2 $\boldsymbol{Q}\boldsymbol{A}$ $\boldsymbol{Q}\bar{\boldsymbol{A}}$ σ^2

We will also use the following extension of Proposition 2.17 of [ST04].

PROPOSITION 2.3 (Gaussian measure of halfspaces). $\boldsymbol{t}$ $\mathbb{R}^n$ r $\bar{\boldsymbol{b}}$ $\mathbb{R}^n$ $\boldsymbol{b}$ $\bar{\boldsymbol{b}}$ σ^2

$$\Pr_{\boldsymbol{b}} [|\boldsymbol{t}^T\boldsymbol{b}| \leq r] \leq \frac{1}{\sqrt{2\pi}\sigma} \int_{t=-r}^{t=r} e^{-t^2/2\sigma^2} dt.$$

In this paper we will use the following properties of matrix norms and vector norms.

PROPOSITION 2.4 (product). $\boldsymbol{A}$ $\boldsymbol{B}$ $\boldsymbol{A}\boldsymbol{B}$ $1 \leq p \leq \infty$

$$\|\boldsymbol{A}\boldsymbol{B}\|_p \leq \|\boldsymbol{A}\|_p \|\boldsymbol{B}\|_p.$$

PROPOSITION 2.5 (vector norms). $\boldsymbol{a}$ $\mathbb{R}^n$ $\|\boldsymbol{a}\|_1/\sqrt{n} \leq \|\boldsymbol{a}\|_2 \leq \|\boldsymbol{a}\|_1$

PROPOSITION 2.6 (2-norm). $\boldsymbol{A}$

$$\|\boldsymbol{A}\|_2 = \left\|\boldsymbol{A}^T\right\|_2,$$

$\sqrt{\boldsymbol{A}^T\boldsymbol{A}}$

PROPOSITION 2.7 ($\|\boldsymbol{A}\|_\infty$: the maximum absolute row sum norm). $\boldsymbol{A}$

$$\|\boldsymbol{A}\|_\infty = \max_i \left\|\boldsymbol{a}_i^T\right\|_1,$$

$\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ $\boldsymbol{A}$ $\boldsymbol{D}$ $\boldsymbol{A}$

$$\|\boldsymbol{D}\|_\infty \leq \|\boldsymbol{A}\|_\infty.$$

PROPOSITION 2.8 ($\|\boldsymbol{A}\|_1$: the maximum absolute column sum norm). $\boldsymbol{A}$

$$\|\boldsymbol{A}\|_1 = \max_i \|\boldsymbol{a}_i\|_1,$$

$\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ $\boldsymbol{A}$

$$\|\boldsymbol{A}\|_1 = \left\|\boldsymbol{A}^T\right\|_\infty.$$

3. Smoothed analysis of the condition number of a matrix. In this section, we will prove the following theorem which shows that for every matrix it is unlikely that a slight perturbation of that matrix has large condition number.

THEOREM 3.1 (smoothed analysis of condition number). $\bar{\boldsymbol{A}}$ $n \times n$ $\left\|\bar{\boldsymbol{A}}\right\|_2 \leq \sqrt{n}$ $\boldsymbol{A}$ $\bar{\boldsymbol{A}}$ $\sigma^2 \leq 1$ $x \geq 1$

$$\Pr\left[\kappa(\boldsymbol{A}) \geq x\right] \leq \frac{14.1n\left(1 + \sqrt{2\ln(x)/9n}\right)}{x\sigma}.$$

As bounds on the norm of a random matrix are standard, we focus on the norm of the inverse. Recall that $1/\left\|\boldsymbol{A}^{-1}\right\|_2 = \min_{\boldsymbol{x}} \|\boldsymbol{A}\boldsymbol{x}\|_2 / \|\boldsymbol{x}\|_2$.

The first step in the proof is to bound the probability that $\left\|\boldsymbol{A}^{-1}\boldsymbol{v}\right\|_2$ is small for a fixed unit vector $\boldsymbol{v}$. This result is also used later (in section 4.1) in studying the growth factor. Using this result and an averaging argument, we then bound the probability that $\left\|\boldsymbol{A}^{-1}\right\|_2$ is large.

LEMMA 3.2 (projection of $\boldsymbol{A}^{-1}$). $\bar{\boldsymbol{A}}$ $\mathbb{R}^{n\times n}$ $\boldsymbol{A}$ $\bar{\boldsymbol{A}}$ σ^2 $\boldsymbol{v}$

$$\Pr\left[\left\|\boldsymbol{A}^{-1}\boldsymbol{v}\right\|_2 > x\right] < \sqrt{\frac{2}{\pi}}\frac{1}{x\sigma}.$$

. Let $\boldsymbol{Q}$ be an orthonormal matrix such that $\boldsymbol{Q}^T\boldsymbol{e}_1 = \boldsymbol{v}$. Let $\bar{\boldsymbol{B}} = \boldsymbol{Q}\bar{\boldsymbol{A}}$ and $\boldsymbol{B} = \boldsymbol{Q}\boldsymbol{A}$. By Proposition 2.2, $\boldsymbol{B}$ is a Gaussian perturbation of $\bar{\boldsymbol{B}}$ of variance σ^2. We have

$$\left\|\boldsymbol{A}^{-1}\boldsymbol{v}\right\|_2 = \left\|\boldsymbol{A}^{-1}\boldsymbol{Q}^T\boldsymbol{e}_1\right\|_2 = \left\|(\boldsymbol{Q}\boldsymbol{A})^{-1}\boldsymbol{e}_1\right\|_2 = \left\|\boldsymbol{B}^{-1}\boldsymbol{e}_1\right\|_2.$$

Thus, to prove the lemma it is sufficient to show

$$\Pr_{\boldsymbol{B}}\left[\left\|\boldsymbol{B}^{-1}\boldsymbol{e}_1\right\|_2 > x\right] < \sqrt{\frac{2}{\pi}}\frac{1}{x\sigma}.$$

We observe that

$$\left\|\boldsymbol{B}^{-1}\boldsymbol{e}_1\right\|_2 = \left\|(\boldsymbol{B}^{-1})_{:,1}\right\|_2,$$

the length of the first column of $\boldsymbol{B}^{-1}$. The first column of $\boldsymbol{B}^{-1}$, by the definition of the matrix inverse, is the vector that is orthogonal to every row of $\boldsymbol{B}$ but the first and

that has inner product 1 with the first row of $\boldsymbol{B}$. Hence its length is the reciprocal of the length of the projection of the first row of $\boldsymbol{B}$ onto the subspace orthogonal to the rest of the rows.

Let $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ be the rows of B and $\bar{\boldsymbol{b}}_1, \ldots, \bar{\boldsymbol{b}}_n$ be the rows of $\bar{\boldsymbol{B}}$. Note that $\boldsymbol{b}_i$ is a Gaussian perturbation of $\bar{\boldsymbol{b}}_i$ of variance σ^2. Let $\boldsymbol{t}$ be the unit vector that is orthogonal to the span of $\boldsymbol{b}_2, \ldots, \boldsymbol{b}_n$. Then

$$\left\|(\boldsymbol{B}^{-1})_{:,1}\right\|_2 = \left|\frac{1}{\boldsymbol{t}^T \boldsymbol{b}_1}\right|.$$

Thus,

$$\begin{aligned}
\Pr_{\boldsymbol{B}}\left[\left\|\boldsymbol{B}^{-1}\boldsymbol{v}\right\|_2 > x\right] &= \Pr_{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_n}\left[\left|\frac{1}{\boldsymbol{t}^T\boldsymbol{b}_1}\right| > x\right] \\
&\leq \max_{\boldsymbol{b}_2,\ldots,\boldsymbol{b}_n} \Pr_{\boldsymbol{b}_1}\left[\left|\boldsymbol{t}^T\boldsymbol{b}_1\right| < 1/x\right] \\
&< \sqrt{\frac{2}{\pi}}\frac{1}{x\sigma},
\end{aligned}$$

where the first inequality follows from Proposition 2.1 and the second inequality follows from Lemma A.2. □

THEOREM 3.3 (smallest singular value). Let $\bar{\boldsymbol{A}}$ be an arbitrary matrix in $\mathbb{R}^{n\times n}$, let $\boldsymbol{A}$ be a Gaussian perturbation of $\bar{\boldsymbol{A}}$ of variance σ^2, and let $x \geq \sqrt{n}$. Then

$$\Pr_{\boldsymbol{A}}\left[\left\|\boldsymbol{A}^{-1}\right\|_2 \geq x\right] \leq 2.35\frac{\sqrt{n}}{x\sigma}.$$

Proof. Let $\boldsymbol{v}$ be a uniformly distributed random unit vector in $\mathbb{R}^n$. It follows from Lemma 3.2 that

$$\Pr_{\boldsymbol{A},\boldsymbol{v}}\left[\left\|\boldsymbol{A}^{-1}\boldsymbol{v}\right\|_2 \geq x\right] \leq \sqrt{\frac{2}{\pi}}\frac{1}{x\sigma}. \tag{3.1}$$

Since $\boldsymbol{A}$ is a Gaussian perturbation of $\bar{\boldsymbol{A}}$, with probability 1 there is a unique pair $(\boldsymbol{u}, -\boldsymbol{u})$ of unit vectors such that $\left\|\boldsymbol{A}^{-1}\boldsymbol{u}\right\|_2 = \left\|\boldsymbol{A}^{-1}\right\|_2$. From the inequality

$$\left\|\boldsymbol{A}^{-1}\boldsymbol{v}\right\|_2 \geq \left\|\boldsymbol{A}^{-1}\right\|_2 \left|\boldsymbol{u}^T\boldsymbol{v}\right|,$$

we know that for every $c > 0$,

$$\begin{aligned}
\Pr_{\boldsymbol{A},\boldsymbol{v}}\left[\left\|\boldsymbol{A}^{-1}\boldsymbol{v}\right\|_2 \geq x\sqrt{c/n}\right] &\geq \Pr_{\boldsymbol{A},\boldsymbol{v}}\left[\left\|\boldsymbol{A}^{-1}\right\|_2 \geq x \text{ and } \left|\boldsymbol{u}^T\boldsymbol{v}\right| \geq \sqrt{c/n}\right] \\
&= \Pr_{\boldsymbol{A},\boldsymbol{v}}\left[\left\|\boldsymbol{A}^{-1}\right\|_2 \geq x\right] \Pr_{\boldsymbol{A},\boldsymbol{v}}\left[\left|\boldsymbol{u}^T\boldsymbol{v}\right| \geq \sqrt{c/n} \quad \middle| \left\|\boldsymbol{A}^{-1}\right\|_2 \geq x\right] \\
&= \Pr_{\boldsymbol{A}}\left[\left\|\boldsymbol{A}^{-1}\right\|_2 \geq x\right] \Pr_{\boldsymbol{A},\boldsymbol{v}}\left[\left|\boldsymbol{u}^T\boldsymbol{v}\right| \geq \sqrt{c/n} \quad \middle| \left\|\boldsymbol{A}^{-1}\right\|_2 \geq x\right] \\
&\geq \Pr_{\boldsymbol{A}}\left[\left\|\boldsymbol{A}^{-1}\right\|_2 \geq x\right] \min_{\boldsymbol{A}:\|\boldsymbol{A}^{-1}\|_2 \geq x} \Pr_{\boldsymbol{v}}\left[\left|\boldsymbol{u}^T\boldsymbol{v}\right| \geq \sqrt{c/n}\right]
\end{aligned}$$

(by Proposition 2.1)

$$\geq \Pr_{\boldsymbol{A}}\left[\left\|\boldsymbol{A}^{-1}\right\|_2 \geq x\right] \Pr_{G}\left[|G| \geq \sqrt{c}\right] \quad \text{(by Lemma B.1)},$$

where G is a Gaussian random variable with mean $\mathbf{0}$ and variance 1. To prove this last inequality, we first note that $\boldsymbol{v}$ is a random unit vector and independent from $\boldsymbol{u}$. Thus, in a basis of $\mathbb{R}^n$ in which $\boldsymbol{u}$ is the first vector, $\boldsymbol{v}$ is a uniformly distributed random unit vector with the first coordinate equal to $\boldsymbol{u}^T\boldsymbol{v}$, and so we may apply Lemma B.1 to bound $\Pr_{\boldsymbol{v}}\left[\,|\boldsymbol{u}^T\boldsymbol{v}| \geq \sqrt{c/n}\,\right]$ from below by $\Pr_G\left[|G| \geq \sqrt{c}\right]$. So,

$$\begin{aligned}\Pr_{\boldsymbol{A}}\left[\left\|\boldsymbol{A}^{-1}\right\|_2 \geq x\right] &\leq \frac{\Pr_{\boldsymbol{A},\boldsymbol{v}}\left[\left\|\boldsymbol{A}^{-1}\boldsymbol{v}\right\|_2 \geq x\sqrt{c/n}\right]}{\Pr_G\left[|G| \geq \sqrt{c}\right]} \\ &\leq \sqrt{\frac{2}{\pi}}\frac{\sqrt{n}}{x\sigma\sqrt{c}\Pr_G\left[|G| \geq \sqrt{c}\right]} \qquad \text{(by (3.1))}.\end{aligned}$$

Because this inequality is true for every c, we will choose a value for c that almost maximizes $\sqrt{c}\Pr_G\left[|G| \geq \sqrt{c}\right]$ and which in turn almost minimizes the right-hand side.

Choosing $c = 0.57$, and evaluating the error function numerically, we determine

$$\Pr_{\boldsymbol{A}}\left[\left\|\boldsymbol{A}^{-1}\right\|_2 \geq x\right] \leq 2.35\frac{\sqrt{n}}{x\sigma}. \qquad \square$$

Note that Theorem 3.3 gives a smoothed analogue of the following bound of Edelman [Ede88] on Gaussian random matrices.

THEOREM 3.4 (Edelman). Let $\boldsymbol{G} \in \mathbb{R}^{n\times n}$ [illegible] σ^2 [illegible]

$$\Pr_{\boldsymbol{G}}\left[\left\|\boldsymbol{G}^{-1}\right\|_2 \geq x\right] \leq \frac{\sqrt{n}}{x\sigma}.$$

As Gaussian random matrices can be viewed as Gaussian random perturbations of the $n \times n$ all-zero square matrix, Theorem 3.3 extends Edelman's theorem to Gaussian random perturbations of an arbitrary matrix. The constant 2.35 in Theorem 3.3 is bigger than Edelman's 1 for Gaussian random matrices. We conjecture that it is possible to reduce 2.35 in Theorem 3.3 to 1 as well.

CONJECTURE 1 (smallest singular value). Let $\bar{\boldsymbol{A}}$ [illegible] $\mathbb{R}^{n\times n}$ [illegible] $\boldsymbol{A}$ [illegible] $\bar{\boldsymbol{A}}$ [illegible] σ^2 [illegible]

$$\Pr_{\boldsymbol{A}}\left[\left\|\boldsymbol{A}^{-1}\right\|_2 \geq x\right] \leq \frac{\sqrt{n}}{x\sigma}.$$

We now apply Theorem 3.3 to prove Theorem 3.1.

[illegible] 3.1. As observed by Davidson and Szarek [DS01, Theorem II.7], one can apply inequality (1.4) of [LT91] to show that for all $k \geq 0$,

$$\Pr_{\boldsymbol{A}}\left[\left\|\bar{\boldsymbol{A}} - \boldsymbol{A}\right\|_2 \geq \sigma\left(2\sqrt{n} + k\right)\right] \leq e^{-k^2/2}.$$

Replacing σ by its upper bound of 1 and setting $\epsilon = e^{-k^2/2}$, we obtain

$$\Pr_{\boldsymbol{A}}\left[\,\left\|\bar{\boldsymbol{A}} - \boldsymbol{A}\right\|_2 \geq 2\sqrt{n} + \sqrt{2\ln(1/\epsilon)}\,\right] \leq \epsilon$$

for all $\epsilon \leq 1$. By assumption, $\left\|\bar{\boldsymbol{A}}\right\|_2 \leq \sqrt{n}$; so,

$$\Pr_{\boldsymbol{A}}\left[\,\|\boldsymbol{A}\|_2 \geq 3\sqrt{n} + \sqrt{2\ln(1/\epsilon)}\,\right] \leq \epsilon.$$

From the result of Theorem 3.3, we have

$$\Pr_{\boldsymbol{A}}\left[\left\|\boldsymbol{A}^{-1}\right\|_2 \geq \frac{2.35\sqrt{n}}{\epsilon\sigma}\right] \leq \epsilon.$$

Combining these two bounds, we find

$$\Pr_{\boldsymbol{A}}\left[\|\boldsymbol{A}\|_2\left\|\boldsymbol{A}^{-1}\right\|_2 \geq \frac{7.05n + 2.35\sqrt{2n\ln(1/\epsilon)}}{\epsilon\sigma}\right] \leq 2\epsilon.$$

So that we can express this probability in the form of $\Pr_{\boldsymbol{A}}\left[\|\boldsymbol{A}\|_2\left\|\boldsymbol{A}^{-1}\right\|_2 \geq x\right]$, for $x \geq 1$, we let

$$x = \frac{7.05n + 2.35\sqrt{2n\ln(1/\epsilon)}}{\epsilon\sigma}. \tag{3.2}$$

It follows from (3.2) and the assumption $\sigma \leq 1$ that $x\epsilon \geq 1$, implying $\ln(1/\epsilon) \leq \ln x$. From (3.2), we derive

$$\begin{aligned} 2\epsilon &= \frac{2\big(7.05n + 2.35\sqrt{2n\ln(1/\epsilon)}\,\big)}{x\sigma} \leq \frac{2\big(7.05n + 2.35\sqrt{2n\ln x}\,\big)}{x\sigma} \\ &\leq \frac{14.1n\big(1 + \sqrt{2\ln(x)/9n}\,\big)}{x\sigma}. \end{aligned}$$

Therefore, we conclude

$$\Pr\left[\|\boldsymbol{A}\|_2\left\|\boldsymbol{A}^{-1}\right\|_2 \geq x\right] \leq \frac{14.1n\big(1 + \sqrt{2\ln(x)/9n}\,\big)}{x\sigma}. \qquad \square$$

We conjecture that the $1+\sqrt{2\ln(x)/9n}$ term should be unnecessary because those matrices for which $\|\boldsymbol{A}\|_2$ is large are less likely to have $\left\|\boldsymbol{A}^{-1}\right\|_2$ large as well.

4. Growth factor of Gaussian elimination without pivoting. We now turn to proving a bound on the growth factor. We will consider a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ obtained from a Gaussian perturbation of variance σ^2 of an arbitrary matrix $\bar{\boldsymbol{A}}$ satisfying $\left\|\bar{\boldsymbol{A}}\right\|_2 \leq 1$. With probability 1, none of the diagonal entries that occur during elimination will be 0. So, in the spirit of Yeung and Chan [YC97], we analyze the growth factor of Gaussian elimination without pivoting. When we specialize our smoothed analyses to the case $\bar{\boldsymbol{A}} = 0$, we improve the bounds of Yeung and Chan (see Theorem 1.3) by a factor of n. Our improved bound on $\rho_{\boldsymbol{U}}$ agrees with their experimental analyses.

4.1. Growth in $\boldsymbol{U}$. We recall that

$$\rho_{\boldsymbol{U}}(\boldsymbol{A}) = \frac{\|\boldsymbol{U}\|_\infty}{\|\boldsymbol{A}\|_\infty}.$$

In this section, we give two bounds on $\rho_{\boldsymbol{U}}(\boldsymbol{A})$. The first will have a better dependence on σ, and the second will have a better dependence on n. It is the latter bound, Theorem 4.3, that agrees with the experiments of Yeung and Chan [YC97] when specialized to the average-case by setting $\bar{\boldsymbol{A}} = 0$ and $\sigma = 1$.

4.1.1. First bound.

THEOREM 4.1 (first bound on $\rho_{\boldsymbol{U}}(\boldsymbol{A})$). Let $\bar{\boldsymbol{A}}$ be any $n \times n$ matrix such that $\left\|\bar{\boldsymbol{A}}\right\|_2 \leq 1$, let $\boldsymbol{A}$ be a random perturbation of $\bar{\boldsymbol{A}}$ with variance $\sigma^2 \leq 1$, and let $x \geq \sigma\sqrt{2\pi}$. Then

$$\Pr\left[\rho_{\boldsymbol{U}}(\boldsymbol{A}) > 1 + x\right] < \frac{1}{\sqrt{2\pi}} \frac{n(n+1)}{x\sigma}.$$

PROOF. By Proposition 2.7.

$$\rho_{\boldsymbol{U}}(\boldsymbol{A}) = \frac{\|\boldsymbol{U}\|_\infty}{\|\boldsymbol{A}\|_\infty} = \max_i \frac{\left\|(\boldsymbol{U}_{i,:})^T\right\|_1}{\|\boldsymbol{A}\|_\infty}.$$

So, we need to bound the probability that the 1-norm of the vector defined by each row of $\boldsymbol{U}$ is large and then apply a union bound to bound the overall probability.

Fix for now a k between 2 and n. We denote the upper triangular segment of the kth row of $\boldsymbol{U}$ by $\boldsymbol{u}^T = \boldsymbol{U}_{k,k:n}$, and observe that $\boldsymbol{u}$ can be obtained from the formula

$$\boldsymbol{u}^T = \boldsymbol{a}^T - \boldsymbol{b}^T\boldsymbol{C}^{-1}\boldsymbol{D}, \tag{4.1}$$

where

$$\boldsymbol{a}^T = \boldsymbol{A}_{k,k:n}, \qquad \boldsymbol{b}^T = \boldsymbol{A}_{k,1:k-1}, \qquad \boldsymbol{C} = \boldsymbol{A}_{1:k-1,1:k-1}, \qquad \boldsymbol{D} = \boldsymbol{A}_{1:k-1,k:n}.$$

This expression for $\boldsymbol{u}$ follows immediately from

$$\boldsymbol{A}_{1:k,:} = \begin{pmatrix} \boldsymbol{C} & \boldsymbol{D} \\ \boldsymbol{b}^T & \boldsymbol{a}^T \end{pmatrix} = \begin{pmatrix} \boldsymbol{L}_{1:k-1,1:k-1} & 0 \\ \boldsymbol{L}_{k,1:k-1} & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{U}_{1:k-1,1:k-1} & \boldsymbol{U}_{1:k-1,k:n} \\ 0 & \boldsymbol{u}^T \end{pmatrix}.$$

From (4.1), we derive

$$\begin{aligned} \|\boldsymbol{u}\|_1 = \left\|\boldsymbol{a} - \left(\boldsymbol{b}^T\boldsymbol{C}^{-1}\boldsymbol{D}\right)^T\right\|_1 &\leq \|\boldsymbol{a}\|_1 + \|\left(\boldsymbol{b}^T\boldsymbol{C}^{-1}\boldsymbol{D}\right)^T\|_1 \\ &\leq \|\boldsymbol{a}^T\|_\infty + \|\left(\boldsymbol{C}^T\right)^{-1}\boldsymbol{b}\|_1 \ \|\boldsymbol{D}\|_\infty \end{aligned}$$

(by Propositions 2.4 and 2.8)

$$\leq \|\boldsymbol{A}\|_\infty (1 + \|(\boldsymbol{C}^T)^{-1}\boldsymbol{b}\|_1), \tag{4.2}$$

by Proposition 2.7.

We now bound the probability $\|(\boldsymbol{C}^T)^{-1}\boldsymbol{b}\|_1$ is large. By Proposition 2.5,

$$\|(\boldsymbol{C}^T)^{-1}\boldsymbol{b}\|_1 \leq \sqrt{k-1}\|(\boldsymbol{C}^T)^{-1}\boldsymbol{b}\|_2.$$

Note that $\boldsymbol{b}$ and $\boldsymbol{C}$ are independent of each other. Therefore,

$$\begin{aligned} \Pr_{\boldsymbol{b},\boldsymbol{C}}\left[\|\left(\boldsymbol{C}^T\right)^{-1}\boldsymbol{b}\|_1 > x\right] &\leq \Pr_{\boldsymbol{b},\boldsymbol{C}}\left[\|\left(\boldsymbol{C}^T\right)^{-1}\boldsymbol{b}\|_2 > x/\sqrt{k-1}\right] \\ &\leq \sqrt{\frac{2}{\pi}} \frac{\sqrt{k-1}\sqrt{(k-1)\sigma^2+1}}{x\sigma} < \sqrt{\frac{2}{\pi}} \frac{k}{x\sigma}, \end{aligned} \tag{4.3}$$

where the second inequality follows from Lemma 4.2 below and the last inequality follows from the assumption $\sigma^2 \leq 1$.

We now apply a union bound over the choices of k to obtain

$$\Pr\left[\rho_{\boldsymbol{U}}(\boldsymbol{A}) > 1 + x\right] < \sum_{k=2}^{n} \sqrt{\frac{2}{\pi}} \frac{k}{x\sigma} \leq \frac{1}{\sqrt{2\pi}} \frac{n(n+1)}{x\sigma}. \qquad \square$$

LEMMA 4.2. $\bar{\boldsymbol{C}}$ $\mathbb{R}^{d\times d}$ $\boldsymbol{C}$ $\bar{\boldsymbol{C}}$ σ^2 $\bar{\boldsymbol{b}}$ $\mathbb{R}^d$ $\left\|\bar{\boldsymbol{b}}\right\|_2 \leq 1$ $\boldsymbol{b}$ $\bar{\boldsymbol{b}}$ σ^2 $\boldsymbol{b}$ $\boldsymbol{C}$

$$\Pr_{\boldsymbol{b},\boldsymbol{C}}\left[\left\|\boldsymbol{C}^{-1}\boldsymbol{b}\right\|_2 \geq x\right] \leq \sqrt{\frac{2}{\pi}} \frac{\sqrt{\sigma^2 d + 1}}{x\sigma}.$$

. Let $\hat{\boldsymbol{b}}$ be the unit vector in the direction of $\boldsymbol{b}$. By applying Lemma 3.2, we obtain for all $\boldsymbol{b}$,

$$\Pr_{\boldsymbol{C}}\left[\left\|\boldsymbol{C}^{-1}\boldsymbol{b}\right\|_2 > x\right] = \Pr_{\boldsymbol{C}}\left[\|\boldsymbol{C}^{-1}\hat{\boldsymbol{b}}\|_2 > \frac{x}{\|\boldsymbol{b}\|_2}\right] \leq \sqrt{\frac{2}{\pi}} \frac{1}{x\sigma} \|\boldsymbol{b}\|_2 .$$

Let $\mu(\boldsymbol{b})$ denote the density according to which $\boldsymbol{b}$ is distributed. Then, we have

$$\begin{aligned}
\Pr_{\boldsymbol{b},\boldsymbol{C}}\left[\left\|\boldsymbol{C}^{-1}\boldsymbol{b}\right\|_2 > x\right] &= \int_{\boldsymbol{b}\in\mathbb{R}^d} \Pr_{\boldsymbol{C}}\left[\left\|\boldsymbol{C}^{-1}\boldsymbol{b}\right\|_2 > x\right] \mu(\boldsymbol{b}) d\boldsymbol{b} \\
&\leq \int_{\boldsymbol{b}\in\mathbb{R}^d} \left(\sqrt{\frac{2}{\pi}} \frac{1}{x\sigma} \|\boldsymbol{b}\|_2\right) \mu(\boldsymbol{b}) d\boldsymbol{b} \\
&= \sqrt{\frac{2}{\pi}} \frac{1}{x\sigma} \mathrm{E}_{\boldsymbol{b}}\left[\|\boldsymbol{b}\|_2\right].
\end{aligned}$$

It is known [KJ82, p. 277] that $\mathrm{E}_{\boldsymbol{b}}[\|\boldsymbol{b}\|_2^2] \leq \sigma^2 d + \|\bar{\boldsymbol{b}}\|_2^2$. As $\mathrm{E}\left[X\right] \leq \sqrt{\mathrm{E}\left[X^2\right]}$ for every positive random variable X, we have $\mathrm{E}_{\boldsymbol{b}}\left[\|\boldsymbol{b}\|_2\right] \leq \sqrt{\sigma^2 d + \|\bar{\boldsymbol{b}}\|_2^2} \leq \sqrt{\sigma^2 d + 1}$. $\square$

4.1.2. Second bound for $\rho_{\boldsymbol{U}}(\boldsymbol{A})$. In this section, we establish an upper bound on $\rho_{\boldsymbol{U}}(\boldsymbol{A})$ which dominates the bound in Theorem 4.1 for $\sigma \geq n^{-3/2}$.

If we specialize the parameters in this bound to $\bar{\boldsymbol{A}} = 0$ and $\sigma^2 = 1$, we improve the average-case bound proved by Yeung and Chan [YC97] (see Theorem 1.3) by a factor of n. Moreover, the resulting bound agrees with their experimental results.

THEOREM 4.3 (second bound on $\rho_{\boldsymbol{U}}(\boldsymbol{A})$). $\bar{\boldsymbol{A}}$ $n \times n$ $\left\|\bar{\boldsymbol{A}}\right\|_2 \leq 1$ $\boldsymbol{A}$ $\bar{\boldsymbol{A}}$ $\sigma^2 \leq 1$ $n \geq 2$

$$\Pr\left[\rho_{\boldsymbol{U}}(\boldsymbol{A}) > 1 + x\right] \leq \sqrt{\frac{2}{\pi}} \frac{1}{x} \left(\frac{2}{3} n^{3/2} + \frac{n}{\sigma} + \frac{4}{3} \frac{\sqrt{n}}{\sigma^2}\right).$$

As in the proof of Theorem 4.1, we will separately consider the kth row of U for each $2 \leq k \leq n$. For any such k, define $\boldsymbol{u}$, $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{C}$, and $\boldsymbol{D}$ as in the proof of Theorem 4.1.

In the case when $k = n$, we may apply (4.3) in the proof of Theorem 4.1, to show

$$\Pr\left[\frac{\|\boldsymbol{u}\|_1}{\|\boldsymbol{A}\|_\infty} > 1 + x\right] \leq \sqrt{\frac{2}{\pi}}\frac{n}{x\sigma}. \tag{4.4}$$

We now turn to the case $k \leq n-1$. By (4.1) and Proposition 2.5, we have

$$\begin{aligned}\|\boldsymbol{u}\|_1 \leq \|\boldsymbol{a}\|_1 + \|(\boldsymbol{b}^T\boldsymbol{C}^{-1}\boldsymbol{D})^T\|_1 &\leq \|\boldsymbol{a}\|_1 + \sqrt{k-1}\|\left(\boldsymbol{b}^T\boldsymbol{C}^{-1}\boldsymbol{D}\right)^T\|_2 \\ &= \|\boldsymbol{a}\|_1 + \sqrt{k-1}\left\|\boldsymbol{b}^T\boldsymbol{C}^{-1}\boldsymbol{D}\right\|_2.\end{aligned}$$

The last equation follows from Proposition 2.6. Therefore for all $k \leq n-1$,

$$\begin{aligned}\frac{\|\boldsymbol{u}\|_1}{\|\boldsymbol{A}\|_\infty} &\leq \frac{\|\boldsymbol{a}\|_1 + \sqrt{k-1}\left\|\boldsymbol{b}^T\boldsymbol{C}^{-1}\boldsymbol{D}\right\|_2}{\|\boldsymbol{A}\|_\infty} \\ &\leq 1 + \frac{\sqrt{k-1}\left\|\boldsymbol{b}^T\boldsymbol{C}^{-1}\boldsymbol{D}\right\|_2}{\|\boldsymbol{A}\|_\infty} \qquad \text{(by Proposition 2.7),} \\ &\leq 1 + \frac{\sqrt{k-1}\left\|\boldsymbol{b}^T\boldsymbol{C}^{-1}\boldsymbol{D}\right\|_2}{\|(\boldsymbol{A}_{n,:})^T\|_1} \qquad \text{(also by Proposition 2.7).}\end{aligned}$$

We now observe that for fixed $\boldsymbol{b}$ and $\boldsymbol{C}$, $(\boldsymbol{b}^T\boldsymbol{C}^{-1})\boldsymbol{D}$ is a Gaussian random row vector of variance $\|\boldsymbol{b}^T\boldsymbol{C}^{-1}\|_2^2\sigma^2$ centered at $(\boldsymbol{b}^T\boldsymbol{C}^{-1})\bar{\boldsymbol{D}}$, where $\bar{\boldsymbol{D}}$ is the center of $\boldsymbol{D}$. We have $\|\bar{\boldsymbol{D}}\|_2 \leq \|\bar{\boldsymbol{A}}\|_2 \leq 1$, by the assumptions of the theorem; so,

$$\left\|\boldsymbol{b}^T\boldsymbol{C}^{-1}\bar{\boldsymbol{D}}\right\|_2 \leq \left\|\boldsymbol{b}^T\boldsymbol{C}^{-1}\right\|_2\left\|\bar{\boldsymbol{D}}\right\|_2 \leq \left\|\boldsymbol{b}^T\boldsymbol{C}^{-1}\right\|_2.$$

Thus, if we let $\boldsymbol{t}^T = (\boldsymbol{b}^T\boldsymbol{C}^{-1}\boldsymbol{D})/\|\boldsymbol{b}^T\boldsymbol{C}^{-1}\|_2$, then for every fixed $\boldsymbol{b}$ and $\boldsymbol{C}$, $\boldsymbol{t}$ is a Gaussian random column vector in $\mathbb{R}^{n-k+1}$ of variance σ^2 centered at a vector of 2-norm at most 1. We also have

$$\Pr_{\boldsymbol{b},\boldsymbol{C},\boldsymbol{D}}\left[\left\|\boldsymbol{b}^T\boldsymbol{C}^{-1}\boldsymbol{D}\right\|_2 \geq x\right] = \Pr_{\boldsymbol{b},\boldsymbol{C},\boldsymbol{t}}\left[\left\|\boldsymbol{b}^T\boldsymbol{C}^{-1}\right\|_2\|\boldsymbol{t}\|_2 \geq x\right]. \tag{4.5}$$

It follows from Lemma 4.2 that

$$\Pr_{\boldsymbol{b},\boldsymbol{C}}\left[\left\|\boldsymbol{b}^T\boldsymbol{C}^{-1}\right\|_2 \geq x\right] \leq \sqrt{\frac{2}{\pi}}\frac{\sqrt{\sigma^2(k-1)+1}}{x\sigma}.$$

Hence, we may apply Corollary C.5 to show

$$\begin{aligned}\Pr_{\boldsymbol{b},\boldsymbol{C},\boldsymbol{t}}\left[\left\|\boldsymbol{b}^T\boldsymbol{C}^{-1}\right\|_2\|\boldsymbol{t}\|_2 \geq x\right] &\leq \sqrt{\frac{2}{\pi}}\frac{\sqrt{\sigma^2(k-1)+1}\sqrt{\sigma^2(n-k+1)+1}}{x\sigma} \\ &\leq \sqrt{\frac{2}{\pi}}\frac{\left(1+\frac{n\sigma^2}{2}\right)}{x\sigma}. \qquad (4.6)\end{aligned}$$

Note that $\boldsymbol{A}_{n,:}$ is a Gaussian perturbation of variance σ^2 of a row vector in $\mathbb{R}^n$. As $\boldsymbol{A}_{n,:}$ is independent of $\boldsymbol{b}$, $\boldsymbol{C}$, and $\boldsymbol{D}$, we can apply (4.5), (4.6), and Lemma C.4 to show

$$\begin{aligned}\Pr\left[\frac{\sqrt{k-1}\left\|\boldsymbol{b}^T\boldsymbol{C}^{-1}\boldsymbol{D}\right\|_2}{\|(\boldsymbol{A}_{n,:})^T\|_1} \geq x\right] &\leq \sqrt{\frac{2}{\pi}}\frac{\sqrt{k-1}\left(1+\frac{n\sigma^2}{2}\right)}{x\sigma}\,\mathrm{E}\left[\frac{1}{\|(\boldsymbol{A}_{n,:})^T\|_1}\right] \\ &\leq \sqrt{\frac{2}{\pi}}\frac{\sqrt{k-1}\left(1+\frac{n\sigma^2}{2}\right)}{x\sigma}\frac{2}{n\sigma}\end{aligned}$$

by Lemma A.4.

Applying a union bound over the choices for k, we obtain

$$\begin{aligned}\Pr\left[\rho_{\boldsymbol{U}}(\boldsymbol{A}) > 1+x\right] &\leq \left(\sum_{k=2}^{n-1}\sqrt{\frac{2}{\pi}}\frac{\sqrt{k-1}\left(1+\frac{n\sigma^2}{2}\right)}{x\sigma}\frac{2}{n\sigma}\right)+\sqrt{\frac{2}{\pi}}\frac{n}{x\sigma}\\ &\leq \sqrt{\frac{2}{\pi}}\frac{1}{x}\left(\frac{2}{3}\sqrt{n}\left(\frac{2}{\sigma^2}+n\right)+\frac{n}{\sigma}\right)\\ &= \sqrt{\frac{2}{\pi}}\frac{1}{x}\left(\frac{2}{3}n^{3/2}+\frac{n}{\sigma}+\frac{4}{3}\frac{\sqrt{n}}{\sigma^2}\right),\end{aligned}$$

where the second inequality follows from

$$\sum_{k=1}^{n-2}\sqrt{k} \leq \frac{2}{3}n^{3/2}. \qquad \square$$

4.2. Growth in $\boldsymbol{L}$. Let $\boldsymbol{L}$ be the lower-triangular part of the LU-factorization of $\boldsymbol{A}$. We have

$$\boldsymbol{L}_{(k+1):n,k} = \boldsymbol{A}^{(k-1)}_{(k+1):n,k} \Big/ \boldsymbol{A}^{(k-1)}_{k,k},$$

where we let $\boldsymbol{A}^{(k)}$ denote the matrix remaining after the first k columns have been eliminated. So, $\boldsymbol{A}^{(0)} = \boldsymbol{A}$.

Recall $\rho_{\boldsymbol{L}}(\boldsymbol{A}) = \|\boldsymbol{L}\|_\infty$, which is equal to the maximum absolute row sum of $\boldsymbol{L}$ (Proposition 2.7). We will show that it is unlikely that $\left\|\boldsymbol{L}_{(k+1):n,k}\right\|_\infty$ is large by proving that it is unlikely that $\left\|\boldsymbol{A}^{(k-1)}_{(k+1):n,k}\right\|_\infty$ is large while $\left|\boldsymbol{A}^{(k-1)}_{k,k}\right|$ is small.

THEOREM 4.4 ($\rho_{\boldsymbol{L}}(\boldsymbol{A})$). [illegible] $\bar{\boldsymbol{A}}$ [illegible] $n \times n$ [illegible] $\left\|\bar{\boldsymbol{A}}\right\|_2 \leq 1$ [illegible] $\boldsymbol{A}$ [illegible] $\bar{\boldsymbol{A}}$ [illegible] $\sigma^2 \leq 1$ [illegible] $n \geq 2$ [illegible]

$$\Pr\left[\rho_{\boldsymbol{L}}(\boldsymbol{A}) > x\right] \leq \sqrt{\frac{2}{\pi}}\frac{n^2}{x}\left(\frac{\sqrt{2}}{\sigma}+\sqrt{2\ln n}+\frac{1}{\sqrt{2\pi}\ln n}\right).$$

[illegible]. For each k between 1 and $n-1$, we have

$$\begin{aligned}\boldsymbol{L}_{(k+1):n,k} &= \frac{\boldsymbol{A}^{(k-1)}_{(k+1):n,k}}{\boldsymbol{A}^{(k-1)}_{k,k}}\\ &= \frac{\boldsymbol{A}_{(k+1):n,k} - \boldsymbol{A}_{(k+1):n,1:(k-1)}\boldsymbol{A}^{-1}_{1:(k-1),1:(k-1)}\boldsymbol{A}_{1:(k-1),k}}{\boldsymbol{A}_{k,k} - \boldsymbol{A}_{k,1:(k-1)}\boldsymbol{A}^{-1}_{1:(k-1),1:(k-1)}\boldsymbol{A}_{1:(k-1),k}}\\ &= \frac{\boldsymbol{A}_{(k+1):n,k} - \boldsymbol{A}_{(k+1):n,1:(k-1)}\boldsymbol{v}}{\boldsymbol{A}_{k,k} - \boldsymbol{A}_{k,1:(k-1)}\boldsymbol{v}},\end{aligned}$$

where we let $\boldsymbol{v} = \boldsymbol{A}^{-1}_{1:(k-1),1:(k-1)}\boldsymbol{A}_{1:(k-1),k}$. Since $\|\bar{\boldsymbol{A}}\|_2 \leq 1$, and all the terms $\boldsymbol{A}_{(k+1):n,k}$, $\boldsymbol{A}_{(k+1):n,1:(k-1)}$, $\boldsymbol{A}_{k,k}$, $\boldsymbol{A}_{k,1:(k-1)}$, and $\boldsymbol{v}$ are independent, we can apply

Lemma 4.5 to show that

$$\begin{aligned}
&\Pr\left[\left\|\boldsymbol{L}_{(k+1):n,k}\right\|_\infty > x\right] \\
&\quad\le \sqrt{\frac{2}{\pi}}\frac{1}{x}\left(\frac{\sqrt{2}}{\sigma} + \sqrt{2\ln(\max(n-k,2))} + \frac{1}{\sqrt{2\pi}\ln(\max(n-k,2))}\right) \\
&\quad\le \sqrt{\frac{2}{\pi}}\frac{1}{x}\left(\frac{\sqrt{2}}{\sigma} + \sqrt{2\ln n} + \frac{1}{\sqrt{2\pi}\ln n}\right),
\end{aligned}$$

where the last inequality follows the facts that $\sqrt{2z} + \frac{1}{\sqrt{2\pi z}}$ is an increasing function when $z \ge \pi^{-1/3}$, and $\ln 2 \ge \pi^{-1/3}$.

The theorem now follows by applying a union bound over the n choices for k and observing that $\|\boldsymbol{L}\|_\infty$ is at most n times the largest entry in $\boldsymbol{L}$. □

LEMMA 4.5 (vector ratio). [illegible] d [illegible] n [illegible] a $\boldsymbol{b}$ $\boldsymbol{x}$ [illegible] $\boldsymbol{Y}$ [illegible] $\bar{a} \in \mathbb{R}^1$ $\bar{\boldsymbol{b}} \in \mathbb{R}^d$ $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ [illegible] $\bar{\boldsymbol{Y}} \in \mathbb{R}^{n\times d}$ [illegible] σ^2 [illegible] $|\bar{a}| \le 1$ $\left\|\bar{\boldsymbol{b}}\right\|_2 \le 1$ $\|\bar{\boldsymbol{x}}\|_2 \le 1$ [illegible] $\left\|\bar{\boldsymbol{Y}}\right\|_2 \le 1$ [illegible] $\boldsymbol{v}$ [illegible] $\mathbb{R}^d$ [illegible] a $\boldsymbol{b}$ $\boldsymbol{x}$ [illegible] $\boldsymbol{Y}$ [illegible] $\sigma^2 \le 1$ [illegible]

$$\Pr\left[\frac{\|\boldsymbol{x}+\boldsymbol{Y}\boldsymbol{v}\|_\infty}{|a+\boldsymbol{b}^T\boldsymbol{v}|} > x\right] \le \sqrt{\frac{2}{\pi}}\frac{1}{x}\left(\frac{\sqrt{2}}{\sigma} + \sqrt{2\ln\max(n,2)} + \frac{1}{\sqrt{2\pi}\ln\max(n,2)}\right),$$

[illegible]. We begin by observing that $a+\boldsymbol{b}^T\boldsymbol{v}$ and each component of $\boldsymbol{x}+\boldsymbol{Y}\boldsymbol{v}$ is a Gaussian random variable of variance $\sigma^2(1+\|\boldsymbol{v}\|_2^2)$ whose mean has absolute value at most $1+\|\boldsymbol{v}\|_2$, and that all these variables are independent. By Lemma A.3,

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{x},\boldsymbol{Y}}\left[\|\boldsymbol{x}+\boldsymbol{Y}\boldsymbol{v}\|_\infty\right] &\le 1+\|\boldsymbol{v}\|_2 \\
&\quad + \left(\sigma\sqrt{(1+\|\boldsymbol{v}\|_2^2)}\right)\left(\sqrt{2\ln\max(n,2)} + \frac{1}{\sqrt{2\pi}\ln\max(n,2)}\right).
\end{aligned}$$

On the other hand, Lemma A.2 implies

$$\Pr_{a,\boldsymbol{b}}\left[\frac{1}{|a+\boldsymbol{b}^T\boldsymbol{v}|} > x\right] \le \sqrt{\frac{2}{\pi}}\frac{1}{x\sigma\sqrt{1+\|\boldsymbol{v}\|_2^2}}. \tag{4.7}$$

Thus, we can apply Corollary C.4 to show

$$\begin{aligned}
&\Pr\left[\frac{\|\boldsymbol{x}+\boldsymbol{Y}\boldsymbol{v}\|_\infty}{|a+\boldsymbol{b}^T\boldsymbol{v}|} > x\right] \\
&\quad\le \sqrt{\frac{2}{\pi}}\frac{1+\|\boldsymbol{v}\|_2 + \left(\sigma\sqrt{1+\|\boldsymbol{v}\|_2^2}\right)\left(\sqrt{2\ln\max(n,2)} + \frac{1}{\sqrt{2\pi}\ln\max(n,2)}\right)}{x\sigma\sqrt{1+\|\boldsymbol{v}\|_2^2}} \\
&\quad= \sqrt{\frac{2}{\pi}}\frac{1}{x}\left(\frac{1+\|\boldsymbol{v}\|_2}{\sigma\sqrt{1+\|\boldsymbol{v}\|_2^2}} + \frac{\left(\sigma\sqrt{1+\|\boldsymbol{v}\|_2^2}\right)\left(\sqrt{2\ln\max(n,2)} + \frac{1}{\sqrt{2\pi}\ln\max(n,2)}\right)}{\sigma\sqrt{1+\|\boldsymbol{v}\|_2^2}}\right) \\
&\quad\le \sqrt{\frac{2}{\pi}}\frac{1}{x}\left(\frac{\sqrt{2}}{\sigma} + \sqrt{2\ln\max(n,2)} + \frac{1}{\sqrt{2\pi}\ln\max(n,2)}\right),
\end{aligned}$$

where the last inequality follows from $(1+z)^2 \le 2(1+z^2)$ for all $z \ge 0$. □

5. Smoothed analysis of Gaussian elimination. We now combine the results from the previous sections to bound the smoothed precision needed in the application of Gaussian elimination without pivoting to obtain solutions to linear systems accurate to b bits.

THEOREM 5.1 (smoothed precision of Gaussian elimination). [illegible] $n > e^4$ [illegible] $\bar{\boldsymbol{A}}$ [illegible] $n \times n$ [illegible] $\|\bar{\boldsymbol{A}}\|_2 \leq 1$ [illegible] $\boldsymbol{A}$ [illegible] $\bar{\boldsymbol{A}}$ [illegible] $\sigma^2 \leq 1/4$ [illegible] $\boldsymbol{Ax} = \boldsymbol{b}$ [illegible] b [illegible]

$$b + \frac{11}{2}\log_2 n + 3\log_2\left(\frac{1}{\sigma}\right) + \log_2(1 + 2\sqrt{n}\sigma) + \frac{1}{2}\log_2\log_2 n + 6.83.$$

[illegible] By Wilkinson's theorem, we need the machine precision, ϵ_{mach}, to satisfy

$$5 \cdot 2^b n \rho_{\boldsymbol{L}}(\boldsymbol{A})\rho_{\boldsymbol{U}}(\boldsymbol{A})\kappa(\boldsymbol{A})\epsilon_{mach} \leq 1 \implies 2.33 + b + \log_2 n + \log_2(\rho_{\boldsymbol{L}}(\boldsymbol{A})) + \max(0, \log_2(\rho_{\boldsymbol{U}}(\boldsymbol{A}))) + \log_2(\kappa(\boldsymbol{A})) \leq \log_2(1/\epsilon_{mach}).$$

We will apply Lemma C.6 to bound these log terms. Theorem 4.1 tells us that

$$\Pr\left[\rho_U(\boldsymbol{A}) > 1 + x\right] \leq \frac{1}{\sqrt{2\pi}}\frac{n(n+1)}{x\sigma}.$$

To put this inequality into a form to which Lemma C.6 may be applied, we set

$$y = x\left(1 + \frac{\sqrt{2\pi}\sigma}{n(n+1)}\right),$$

to obtain

$$\Pr\left[\rho_U(\boldsymbol{A}) > y\right] \leq \left(\frac{1}{\sqrt{2\pi}}\frac{n(n+1)}{\sigma} + 1\right)\frac{1}{y}.$$

By Lemma C.6,

$$\begin{aligned}
\mathrm{E}\left[\max(0, \log_2 \rho_{\boldsymbol{U}}(\boldsymbol{A}))\right] &\leq \log_2\left(\frac{1}{\sqrt{2\pi}}\frac{n(n+1)}{\sigma} + 1\right) + \log_2 e \\
&\leq \log_2\left(n(n+1) + \sigma\sqrt{2\pi}\right) + \log_2\left(\frac{1}{\sigma}\right) + \log_2\left(\frac{e}{\sqrt{2\pi}}\right) \\
&\leq \log_2\left(1.02n^2\right) + \log_2\left(\frac{1}{\sigma}\right) + \log_2\left(\frac{e}{\sqrt{2\pi}}\right) \\
&\leq 2\log_2 n + \log_2\left(\frac{1}{\sigma}\right) + 0.15,
\end{aligned}$$

where in the second-to-last inequality, we used the assumptions $n \geq e^4$ and $\sigma \leq 1/2$. In the last inequality, we numerically computed $\log_2(1.02e/\sqrt{2\pi}) < 0.15$.

Theorem 4.4 and Lemma C.6 imply

$$\begin{aligned}
&\mathrm{E}\left[\log_2 \rho_{\boldsymbol{L}}(\boldsymbol{A})\right] \\
&\quad \le \log_2\left(\sqrt{\frac{2}{\pi}}n^2\left(\frac{\sqrt{2}}{\sigma}+\sqrt{2\ln n}+\frac{1}{\sqrt{2\pi}\ln n}\right)\right)+\log_2 e \\
&\quad \le 2\log_2 n+\log_2\left(\frac{1}{\sigma}+\sqrt{\ln n}\left(1+\frac{1}{2\sqrt{\pi}\ln n}\right)\right)+\log_2\left(\frac{2e}{\sqrt{\pi}}\right) \\
&\quad = 2\log_2 n+\log_2\left(\frac{1}{\sigma}\right)+\log_2\sqrt{\ln n} \\
&\qquad +\log_2\left(\frac{1}{\sqrt{\ln n}}+\sigma\left(1+\frac{1}{2\sqrt{\pi}\ln n}\right)\right)+\log_2\left(\frac{2e}{\sqrt{\pi}}\right)
\end{aligned}$$

using $\sigma \le \frac{1}{2}$ and $n > e^4$,

$$\begin{aligned}
&\le 2\log_2 n+\log_2\left(\frac{1}{\sigma}\right)+\frac{1}{2}\log_2\log_2 n+\log_2\left(1+\frac{1}{16\sqrt{\pi}}\right)+\log_2\left(\frac{2e}{\sqrt{\pi}}\right) \\
&\le 2\log_2 n+\log_2\left(\frac{1}{\sigma}\right)+\frac{1}{2}\log_2\log_2 n+1.67,
\end{aligned}$$

as $\log_2(1+1/16\sqrt{\pi})+\log_2(2e/\sqrt{\pi}) < 1.67$. Theorem 3.3 and Lemma C.6, along with the observation that $\log_2(2.35e) < 2.68$, imply

$$\mathrm{E}\left[\log_2\left\|\boldsymbol{A}^{-1}\right\|_2\right] \le \frac{1}{2}\log_2 n+\log_2\left(\frac{1}{\sigma}\right)+2.68.$$

Finally,

$$\mathrm{E}\left[\log_2(\|\boldsymbol{A}\|_2)\right] \le \log_2(1+2\sqrt{n}\sigma)$$

follows from the well-known facts that the expectation of $\left\|\boldsymbol{A}-\bar{\boldsymbol{A}}\right\|_2$ is at most $2\sqrt{n}\sigma$ (c.f., [Seg00]) and that $\mathrm{E}\left[\log_2(X)\right] \le \log_2 \mathrm{E}\left[X\right]$ for every positive random variable X. Thus, the expected number of digits of precision needed is at most

$$b+\frac{11}{2}\log_2 n+3\log_2\left(\frac{1}{\sigma}\right)+\log_2(1+2\sqrt{n}\sigma)+\frac{1}{2}\log_2\log_2 n+6.83. \qquad \square$$

The following conjecture would further improve the coefficient of $\log(1/\sigma)$.

CONJECTURE 2. $\bar{\boldsymbol{A}}$ $n \times n$ $\left\|\bar{\boldsymbol{A}}\right\|_2 \le 1$ $\boldsymbol{A}$ $\bar{\boldsymbol{A}}$ $\sigma^2 \le 1$

$$\Pr\left[\rho_{\boldsymbol{L}}(\boldsymbol{A})\rho_{\boldsymbol{U}}(\boldsymbol{A})\kappa(\boldsymbol{A}) > x\right] \le \frac{n^{c_1}\log^{c_2}(x)}{x\sigma},$$

c_1 c_2

6. Zero-preserving perturbations of symmetric matrices with diagonals. Many matrices that occur in practice are symmetric and sparse. Moreover, many matrix algorithms take advantage of this structure. Thus, it is natural to study the smoothed analysis of algorithms under perturbations that respect symmetry and nonzero structure. In this section, we study the condition numbers and growth factors

of Gaussian elimination without pivoting of symmetric matrices under perturbations that only alter their diagonal and nonzero entries.

DEFINITION 6.1 (Zero-preserving perturbations). $\bar{\boldsymbol{T}}$ zero-preserving perturbation of $\bar{\boldsymbol{T}}$ of variance σ^2 $\boldsymbol{T}$ 0 σ^2 $\bar{\boldsymbol{T}}$

Throughout this section, when we express a symmetric matrix $\boldsymbol{A}$ as $\boldsymbol{T}+\boldsymbol{D}+\boldsymbol{T}^T$, we mean that $\boldsymbol{T}$ is lower-triangular with zeros on the diagonal and $\boldsymbol{D}$ is a diagonal matrix. By making a zero-preserving perturbation to $\bar{\boldsymbol{T}}$, we preserve the symmetry of the matrix. The main results of this section are that the smoothed condition number and growth factors of symmetric matrices under zero-preserving perturbations to $\boldsymbol{T}$ and diagonal perturbations to $\boldsymbol{D}$ have distributions similar to those proved in sections 3 and 4 for dense matrices under dense perturbations.

6.1. Bounding the condition number. We begin by recalling that the singular values and vectors of symmetric matrices are the eigenvalues and eigenvectors.

LEMMA 6.2. $\bar{\boldsymbol{A}} = \bar{\boldsymbol{T}} + \bar{\boldsymbol{D}} + \bar{\boldsymbol{T}}^T$ $n \times n$ $\boldsymbol{T}$ $\bar{\boldsymbol{T}}$ σ^2 $\boldsymbol{G}_D$ σ^2 0 $\boldsymbol{T}$ $\boldsymbol{D} = \bar{\boldsymbol{D}} + \boldsymbol{G}_D$ $\boldsymbol{A} = \boldsymbol{T} + \boldsymbol{D} + \boldsymbol{T}^T$

$$\Pr\left[\left\|\boldsymbol{A}^{-1}\right\|_2 \geq x\right] \leq \sqrt{\frac{2}{\pi}}\frac{n^{3/2}}{x\sigma}.$$

. By Proposition 2.1,

$$\Pr_{\boldsymbol{T},\boldsymbol{G}_D}\left[\left\|(\boldsymbol{T}+\boldsymbol{D}+\boldsymbol{T}^T)^{-1}\right\|_2 \geq x\right] \leq \max_{\boldsymbol{T}} \Pr_{\boldsymbol{G}_D}\left[\left\|((\boldsymbol{T}+\bar{\boldsymbol{D}}+\boldsymbol{T}^T)+\boldsymbol{G}_D)^{-1}\right\|_2 \geq x\right].$$

The proof now follows from Lemma 6.3, taking $\boldsymbol{T}+\bar{\boldsymbol{D}}+\boldsymbol{T}^T$ as the base matrix. ∎

LEMMA 6.3. $\bar{\boldsymbol{A}}$ $n \times n$ $\boldsymbol{G}_D$ σ^2 0 $\boldsymbol{A} = \bar{\boldsymbol{A}} + \boldsymbol{G}_D$

$$\Pr\left[\left\|\boldsymbol{A}^{-1}\right\|_2 \geq x\right] \leq \sqrt{\frac{2}{\pi}}\frac{n^{3/2}}{x\sigma}.$$

. Let $x_1, \ldots, x_n$ be the diagonal entries of $\boldsymbol{G}_D$, and let

$$g = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \text{and } y_i = x_i - g.$$

Then,

$$\begin{aligned}\Pr_{y_1,\ldots,y_n,g}\left[\left\|(\bar{\boldsymbol{A}}+\boldsymbol{G}_D)^{-1}\right\|_2 \geq x\right] &= \Pr_{y_1,\ldots,y_n,g}\left[\left\|(\bar{\boldsymbol{A}}+\operatorname{diag}(y_1,\ldots,y_n)+g\boldsymbol{I})^{-1}\right\|_2 \geq x\right] \\ &\leq \max_{y_1,\ldots,y_n} \Pr_{g}\left[\left\|(\bar{\boldsymbol{A}}+\operatorname{diag}(y_1,\ldots,y_n)+g\boldsymbol{I})^{-1}\right\|_2 \geq x\right],\end{aligned}$$

where the last inequality follows from Proposition 2.1. The proof now follows from Proposition 6.4 and Lemma 6.5. ∎

PROPOSITION 6.4. $X_1, \ldots, X_n$ σ^2 $a_1, \ldots, a_n$

$$G = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad Y_i = X_i - G.$$

G σ^2/n $(1/n)\sum a_i$ $Y_1, \ldots, Y_n$

LEMMA 6.5. $\bar{\boldsymbol{A}}$ $n \times n$ G 0 σ^2/n $\boldsymbol{A} = \bar{\boldsymbol{A}} + G\boldsymbol{I}$

$$\Pr_{\boldsymbol{A}}\left[\left\|\boldsymbol{A}^{-1}\right\|_2 \geq x\right] \leq \sqrt{\frac{2}{\pi}}\frac{n^{3/2}}{x\sigma}.$$

. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of $\bar{\boldsymbol{A}}$. Then,

$$\left\|(\bar{\boldsymbol{A}} + G\boldsymbol{I})^{-1}\right\|_2^{-1} = \min_i |\lambda_i + G|.$$

Thus,

$$\Pr_{\boldsymbol{A}}\left[\left\|\boldsymbol{A}^{-1}\right\|_2 \geq x\right] = \Pr_{G}\left[\min_i |\lambda_i - G| < \frac{1}{x}\right] \leq \sum_i \Pr_{G}\left[|\lambda_i - G| < \frac{1}{x}\right]$$
$$\leq \sum_i \sqrt{\frac{2}{\pi}}\frac{\sqrt{n}}{x\sigma} \leq \sqrt{\frac{2}{\pi}}\frac{n^{3/2}}{x\sigma},$$

where the second-to-last inequality follows from Lemma A.2 for $\mathbb{R}^1$. □

As in section 3, we can now prove the following theorem.

THEOREM 6.6 (condition number of symmetric matrices). $\bar{\boldsymbol{A}} = \bar{\boldsymbol{T}} + \bar{\boldsymbol{D}} + \bar{\boldsymbol{T}}^T$ $n \times n$ $\left\|\bar{\boldsymbol{A}}\right\|_2 \leq \sqrt{n}$ $\sigma^2 \leq 1$ $\boldsymbol{T}$ $\bar{\boldsymbol{T}}$ σ^2 $\boldsymbol{G}_D$ σ^2 0 $\boldsymbol{T}$ $\boldsymbol{D} = \bar{\boldsymbol{D}} + \boldsymbol{G}_D$ $\boldsymbol{A} = \boldsymbol{T} + \boldsymbol{D} + \boldsymbol{T}^T$

$$\Pr\left[\kappa(\boldsymbol{A}) \geq x\right] \leq 6\sqrt{\frac{2}{\pi}}\frac{n^{7/2}}{x\sigma}(1 + \sqrt{2\ln(x)/9n}).$$

. As in the proof of Theorem 3.1, we can apply the techniques used in the proof of [DS01, Theorem II.7], to show

$$\Pr\left[\left\|\bar{\boldsymbol{A}} - \boldsymbol{A}\right\|_2 \geq 2\sqrt{n} + k\right] < e^{-k^2/2}.$$

The rest of the proof follows the outline of the proof of Theorem 3.1, using Lemma 6.2 instead of Theorem 3.3. □

6.2. Bounding entries in $\boldsymbol{U}$. In this section, we will prove the following theorem.

THEOREM 6.7 ($\rho_{\boldsymbol{U}}(\boldsymbol{A})$ of symmetric matrices). $\bar{\boldsymbol{A}} = \bar{\boldsymbol{T}} + \bar{\boldsymbol{D}} + \bar{\boldsymbol{T}}^T$ $n \times n$ $\left\|\bar{\boldsymbol{A}}\right\|_2 \leq 1$ $\sigma^2 \leq 1$ $\boldsymbol{T}$

$\bar{\boldsymbol{T}}$ σ^2 $\boldsymbol{G}_D$
σ^2 0
$\boldsymbol{T}$ $\boldsymbol{D} = \bar{\boldsymbol{D}} + \boldsymbol{G}_D$ $\boldsymbol{A} = \boldsymbol{T} + \boldsymbol{D} + \boldsymbol{T}^T$

$$\Pr\left[\rho_{\boldsymbol{U}}(\boldsymbol{A}) > 1 + x\right] \leq \frac{2}{7}\sqrt{\frac{2}{\pi}}\frac{n^3}{x\sigma}.$$

We proceed as in the proof of Theorem 4.1. For k between 2 and n, we define $\boldsymbol{u}$, $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{C}$ as in the proof of Theorem 4.1. By (4.2)

$$\frac{\|\boldsymbol{u}\|_1}{\|\boldsymbol{A}\|_\infty} \leq 1 + \left\|(\boldsymbol{C}^T)^{-1}\boldsymbol{b}\right\|_1 \leq 1 + \sqrt{k-1}\left\|\boldsymbol{b}^T\boldsymbol{C}^{-1}\right\|_2 \leq 1 + \sqrt{k-1}\,\|\boldsymbol{b}\|_2\left\|\boldsymbol{C}^{-1}\right\|_2.$$

Hence

$$\begin{aligned}
\Pr\left[\frac{\|\boldsymbol{u}\|_1}{\|\boldsymbol{A}\|_\infty} > 1 + x\right] &\leq \Pr\left[\|\boldsymbol{b}\|_2\left\|\boldsymbol{C}^{-1}\right\|_2 > \frac{x}{\sqrt{k-1}}\right] \\
&\leq \mathrm{E}\left[\|\boldsymbol{b}\|_2\right]\sqrt{\frac{2}{\pi}}\frac{(k-1)^2}{x\sigma}, \quad \text{by Lemmas 6.2 and C.4,} \\
&\leq \sqrt{1 + j\sigma^2}\sqrt{\frac{2}{\pi}}\frac{(k-1)^2}{x\sigma},
\end{aligned}$$

where j is the number of nonzeros in $\boldsymbol{b}$,

$$\leq \sqrt{\frac{2}{\pi}}\frac{\sqrt{k}(k-1)^2}{x\sigma}.$$

Applying a union bound over k,

$$\Pr\left[\rho_U(\boldsymbol{A}) > x\right] \leq \sqrt{\frac{2}{\pi}}\frac{1}{x\sigma}\sum_{k=2}^{n}\sqrt{k}(k-1)^2 \leq \frac{2}{7}\sqrt{\frac{2}{\pi}}\frac{n^{7/2}}{x\sigma}. \qquad \square$$

6.3. Bounding entries in $\boldsymbol{L}$. As in section 4.2, we derive a bound on the growth factor of $\boldsymbol{L}$. As before, we will show that it is unlikely that $\boldsymbol{A}_{j,k}^{(k-1)}$ is large while $\boldsymbol{A}_{k,k}^{(k-1)}$ is small. However, our techniques must differ from those used in section 4.2, as the proof in that section made critical use of the independence of $\boldsymbol{A}_{k,1:(k-1)}$ and $\boldsymbol{A}_{1:(k-1),k}$.

THEOREM 6.8 ($\rho_{\boldsymbol{L}}(\boldsymbol{A})$ of symmetric matrices). $\sigma^2 \leq 1$ $n \geq 2$
$\bar{\boldsymbol{A}} = \bar{\boldsymbol{T}} + \bar{\boldsymbol{D}} + \bar{\boldsymbol{T}}^T$ $n \times n$ $\left\|\bar{\boldsymbol{A}}\right\|_2 \leq 1$
$\boldsymbol{T}$ $\bar{\boldsymbol{T}}$ σ^2 $\boldsymbol{G}_D$
$\sigma^2 \leq 1$ 0
$\boldsymbol{T}$ $\boldsymbol{D} = \bar{\boldsymbol{D}} + \boldsymbol{G}_D$ $\boldsymbol{A} = \boldsymbol{T} + \boldsymbol{D} + \boldsymbol{T}^T$

$$\forall\, x \geq \sqrt{\frac{2}{\pi}}\frac{1}{\sigma^2}, \quad \Pr\left[\rho_{\boldsymbol{L}}(\boldsymbol{A}) > x\right] \leq \frac{3.2n^4}{x\sigma^2}\ln^{3/2}\left(e\sqrt{\frac{\pi}{2}}x\sigma^2\right).$$

Using Lemma 6.9, we obtain for all k

$$\Pr\left[\exists j > k : |\boldsymbol{L}_{j,k}| > x\right] \leq \Pr\left[\left\|\boldsymbol{L}_{(k+1):n,k}\right\|_2 > x\right] \leq \frac{3.2n^2}{x\sigma^2}\ln^{3/2}\left(e\sqrt{\frac{\pi}{2}}x\sigma^2\right).$$

Applying a union bound over the choices for k, we then have

$$\Pr\left[\exists j,k : |\boldsymbol{L}_{j,k}| > x\right] \leq \frac{3.2n^3}{x\sigma^2} \ln^{3/2}\left(e\sqrt{\frac{\pi}{2}}x\sigma^2\right).$$

The result now follows from the fact that $\|\boldsymbol{L}\|_\infty$ is at most n times the largest entry in $\boldsymbol{L}$. □

LEMMA 6.9. [illegible] 6.8

$$\forall\, x \geq \sqrt{\frac{2}{\pi}}\frac{1}{\sigma^2}, \quad \Pr\left[\left\|\boldsymbol{L}_{(k+1):n,k}\right\|_2 > x\right] \leq \frac{3.2n^2}{x\sigma^2} \ln^{3/2}\left(e\sqrt{\frac{\pi}{2}}x\sigma^2\right).$$

[illegible]. We recall that

$$\boldsymbol{L}_{k+1:n,k} = \frac{\boldsymbol{A}_{k+1:n,k} - \boldsymbol{A}_{k+1:n,1:k-1}\boldsymbol{A}^{-1}_{1:k-1,1:k-1}\boldsymbol{A}_{1:k-1,k}}{\boldsymbol{A}_{k,k} - \boldsymbol{A}_{k,1:k-1}\boldsymbol{A}^{-1}_{1:k-1,1:k-1}\boldsymbol{A}_{1:k-1,k}}.$$

Because of the symmetry of $\boldsymbol{A}$, $\boldsymbol{A}_{k,1:k-1}$ is the same as $\boldsymbol{A}_{1:k-1,k}$, so we can no longer use the proof technique that worked in section 4.2. Instead, we will bound the tails of the numerator and denominator separately, exploiting the fact that only the denominator depends upon $\boldsymbol{A}_{k,k}$.

Consider the numerator first. Setting $\boldsymbol{v} = \boldsymbol{A}^{-1}_{1:k-1,1:k-1}\boldsymbol{A}_{1:k-1,k}$, the numerator can be written $\boldsymbol{A}_{k+1:n,1:k}\binom{-\boldsymbol{v}}{1}$. We will now prove that for all $x \geq 1/\sigma$,

$$\Pr_{\substack{\boldsymbol{A}_{k+1:n,1:k}\\ A_{1:k-1,1:k}}}\left[\left\|\boldsymbol{A}_{k+1:n,1:k}\begin{pmatrix}-\boldsymbol{v}\\1\end{pmatrix}\right\|_\infty > x\right] \leq \sqrt{\frac{2}{\pi}}\left(\frac{2n^2(1+\sigma\sqrt{2\ln(x\sigma)})+n}{x\sigma}\right). \tag{6.1}$$

Let

$$c = \frac{1}{1+\sigma\sqrt{2\ln(x\sigma)}}, \tag{6.2}$$

which implies $\frac{1-c}{c\sigma} = \sqrt{2\ln(x\sigma)}$. It suffices to prove (6.1) for all x for which the right-hand side is less than 1. Given that $x \geq 1/\sigma$, it suffices to consider x for which $cx \geq 2$ and $x\sigma \geq 2$.

We use the parameter c to divide the probability as follows:

$$\Pr_{\substack{\boldsymbol{A}_{k+1:n,1:k}\\ A_{1:k-1,1:k}}}\left[\left\|\boldsymbol{A}_{k+1:n,1:k}\begin{pmatrix}-\boldsymbol{v}\\1\end{pmatrix}\right\|_\infty > x\right] \leq \Pr_{\boldsymbol{A}_{1:(k-1),1:k}}\left[\left\|\begin{pmatrix}-\boldsymbol{v}\\1\end{pmatrix}\right\|_2 > cx\right] \tag{6.3}$$

$$+ \Pr_{\boldsymbol{A}_{k+1:n,1:k}}\left[\left\|\boldsymbol{A}_{k+1:n,1:k}\begin{pmatrix}-\boldsymbol{v}\\1\end{pmatrix}\right\|_\infty > \frac{1}{c}\left\|\begin{pmatrix}-\boldsymbol{v}\\1\end{pmatrix}\right\|_2 \,\middle|\, \left\|\begin{pmatrix}-\boldsymbol{v}\\1\end{pmatrix}\right\|_2 \leq cx\right]. \tag{6.4}$$

To evaluate (6.4), we note that once $\boldsymbol{v}$ is fixed, each component of $\boldsymbol{A}_{k+1:n,1:k}\binom{-\boldsymbol{v}}{1}$ is a Gaussian random variable of variance $\left\|\binom{-\boldsymbol{v}}{1}\right\|_2^2\sigma^2$ and mean at most $\left\|\bar{\boldsymbol{A}}_{k+1:n,1:k}\binom{-\boldsymbol{v}}{1}\right\|_2 \leq \left\|\binom{-\boldsymbol{v}}{1}\right\|_2$. So,

$$\left\|\boldsymbol{A}_{k+1:n,1:k}\begin{pmatrix}-\boldsymbol{v}\\1\end{pmatrix}\right\|_\infty > \frac{1}{c}\left\|\begin{pmatrix}-\boldsymbol{v}\\1\end{pmatrix}\right\|_2$$

implies one of the Gaussian random variables differs from its mean by more than $(1/c-1)/\sigma$ times it standard deviation, and we can therefore apply Lemma A.1 and a union bound to derive

$$(6.4) \leq \sqrt{\frac{2}{\pi}} \frac{n e^{-\frac{1}{2}\left(\frac{1-c}{c\sigma}\right)^2}}{\frac{1-c}{c\sigma}} = \sqrt{\frac{2}{\pi}} \frac{n}{x\sigma\sqrt{2\ln(x\sigma)}}.$$

To bound (6.3), we note that Lemma 6.2 and Corollary C.5 imply

$$\Pr_{\boldsymbol{A}_{1:(k-1),1:k}} \left[\left\| \boldsymbol{A}^{-1}_{1:k-1,1:k-1} \boldsymbol{A}_{1:k-1,k} \right\|_2 > y \right] \leq \sqrt{\frac{2}{\pi}} \frac{n^2}{y\sigma},$$

and so

$$\begin{aligned}
\Pr_{\boldsymbol{A}_{1:(k-1),1:k}} \left[\left\| \begin{pmatrix} -\boldsymbol{v} \\ 1 \end{pmatrix} \right\|_2 > cx \right] &\leq \Pr_{\boldsymbol{A}_{1:(k-1),1:k}} \left[\left\| \boldsymbol{A}^{-1}_{1:k-1,1:k-1} \boldsymbol{A}_{1:k-1,k} \right\|_2 > cx - 1 \right] \\
&\leq \sqrt{\frac{2}{\pi}} \frac{n^2}{(cx-1)\sigma} \\
&= \sqrt{\frac{2}{\pi}} \frac{n^2}{(cx\sigma(1-1/cx))} \\
&= \sqrt{\frac{2}{\pi}} \frac{n^2(1+\sigma\sqrt{2\ln(x\sigma)})}{x\sigma\,(1-1/cx)} \\
&\leq \sqrt{\frac{2}{\pi}} \frac{2n^2(1+\sigma\sqrt{2\ln(x\sigma)})}{x\sigma}, \text{ by } cx \geq 2.
\end{aligned}$$

So,

$$\begin{aligned}
(6.5) \qquad & \Pr_{\substack{\boldsymbol{A}_{k+1:n,1:k} \\ A_{1:k-1,1:k}}} \left[\left\| \boldsymbol{A}_{k+1:n,1:k} \begin{pmatrix} -\boldsymbol{v} \\ 1 \end{pmatrix} \right\|_\infty > x \right] \\
&\leq \sqrt{\frac{2}{\pi}} \left(\frac{n}{x\sigma\sqrt{2\ln(x\sigma)}} + \frac{2n^2(1+\sigma\sqrt{2\ln(x\sigma)})}{x\sigma} \right) \\
&\leq \sqrt{\frac{2}{\pi}} \left(\frac{2n^2(1+\sigma\sqrt{2\ln(x\sigma)})+n}{x\sigma} \right),
\end{aligned}$$

by the assumption $x\sigma \geq 2$, which proves (6.1).

As for the denominator, we note that $\boldsymbol{A}_{k,k}$ is independent of all other terms, and hence

$$(6.6) \qquad \Pr \left[\left| \boldsymbol{A}_{k,k} - \boldsymbol{A}_{k,1:k-1} \boldsymbol{A}^{-1}_{1:k-1,1:k-1} \boldsymbol{A}_{1:k-1,k} \right| < 1/x \right] \leq \sqrt{\frac{2}{\pi}} \frac{1}{x\sigma},$$

by Lemma A.2. Applying Corollary C.3 with

$$\alpha = \sqrt{\frac{2}{\pi}}(2n^2+n), \qquad \beta = \frac{4n^2\sigma}{\sqrt{\pi}}, \qquad \gamma = \sqrt{\frac{2}{\pi}}$$

to combine (6.5) with (6.6), we derive the bound

$$\begin{aligned}
&\frac{2}{\pi x\sigma^2}(2n^2+n+((2+4\sqrt{2}\sigma/3)n^2+n)\ln^{3/2}(\sqrt{\pi/2}x\sigma^2)) \\
&\quad \le \frac{2n^2}{\pi x\sigma^2}(3+4\sqrt{2}\sigma/3)(\ln^{3/2}(\sqrt{\pi/2}x\sigma^2)+1) \\
&\quad \le \frac{3.2n^2}{x\sigma^2}\ln^{3/2}(e\sqrt{\pi/2}x\sigma^2),
\end{aligned}$$

as $\sigma \le 1$. ∎

7. Conclusions and open problems.

7.1. Generality of results. In this paper, we have presented bounds on the smoothed values of the condition number and growth factors assuming the input matrix is subjected to a slight Gaussian perturbation. We would like to point out here that our results can be extended to some other families of perturbations.

With the exception of the proof of Theorem 3.3, the only properties of Gaussian random vectors that we used in sections 3 and 4 are:

1. There is a constant c for which the probability that a Gaussian random vector has distance less than ϵ to a hyperplane is at most $c\epsilon$, and
2. It is exponentially unlikely that a Gaussian random vector lies far from its mean.

Moreover, a result similar to Theorem 3.3 but with an extra factor of d could be proved using just fact 1.

In fact, results of a character similar to ours would still hold if the second condition were reduced to a polynomial probability. Many other families of perturbations share these properties. For example, similar results would hold if we let $\boldsymbol{A} = \bar{\boldsymbol{A}} + \boldsymbol{U}$, where $\boldsymbol{U}$ is a matrix of variables independently uniformly chosen in $[-\sigma, \sigma]$, or if $\boldsymbol{A} = \bar{\boldsymbol{A}} + \boldsymbol{S}$, where the columns of $\boldsymbol{S}$ are chosen uniformly among those vectors of norm at most σ.

7.2. Counter-examples. The results of sections 3 and 4 do not extend to zero-preserving perturbations for nonsymmetric matrices. For example, the following matrix remains ill-conditioned under zero-preserving perturbations:

$$\begin{pmatrix}
1 & -2 & 0 & 0 & 0 \\
0 & 1 & -2 & 0 & 0 \\
0 & 0 & 1 & -2 & 0 \\
0 & 0 & 0 & 1 & -2 \\
0 & 0 & 0 & 0 & 1
\end{pmatrix}.$$

A symmetric matrix that remains ill-conditioned under zero-preserving perturbations that do not alter the diagonal can be obtained by locating the above matrix in the

upper-right quadrant, and its transpose in the lower-left quadrant:

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

The following matrix maintains large growth factor under zero-preserving perturbations, regardless of whether partial pivoting or no pivoting is used:

$$\begin{bmatrix}
1.1 & 0 & 0 & 0 & 0 & 1 \\
-1 & 1.1 & 0 & 0 & 0 & 1 \\
-1 & -1 & 1.1 & 0 & 0 & 1 \\
-1 & -1 & -1 & 1.1 & 0 & 1 \\
-1 & -1 & -1 & -1 & 1.1 & 1 \\
-1 & -1 & -1 & -1 & -1 & 1
\end{bmatrix}.$$

These examples can be easily normalized to so that their 2-norms are equal to 1.

7.3. Open problems. Questions that naturally follow from this work are:

- What is the probability that the perturbation of an arbitrary matrix has large growth factors under Gaussian elimination with partial pivoting?
- What is the probability that the perturbation of an arbitrary matrix has large growth factors under Gaussian elimination with complete pivoting?
- Can zero-preserving perturbations of symmetric matrices have large growth factors under partial pivoting or under complete pivoting?
- Can zero-preserving perturbations of arbitrary matrices have large growth factors under complete pivoting?

For the first question, we point out that experimental data of Trefethen and Bau [TB97, p. 168] suggest that the probability that the perturbation of an arbitrary matrix has large growth factor under partial pivoting may be exponentially smaller than without pivoting. This leads us to the following conjecture.

CONJECTURE 3. $\bar{\boldsymbol{A}}$ $n \times n$ $\left\|\bar{\boldsymbol{A}}\right\|_2 \leq 1$ $\boldsymbol{A}$ $\bar{\boldsymbol{A}}$ $\sigma^2 \leq 1$ $\boldsymbol{U}$ $\boldsymbol{A}$ k_1 k_2 α

$$\Pr\left[\|\boldsymbol{U}\|_{\max}/\|\boldsymbol{A}\|_{\max} > x+1\right] \leq n^{k_1} e^{-\alpha x^{k_2}\sigma}.$$

Finally, we ask whether similar analyses can be performed for other algorithms of numerical analysis. One might start by extending Smale's program by analyzing the smoothed values of other condition numbers.

7.4. Recent progress. Since the announcement of our result, Wschebor [Wsc04] improved the smoothed bound on the condition number.

THEOREM 7.1 (Wschebor). $\bar{\boldsymbol{A}}$ $n \times n$ $\boldsymbol{A}$ $\bar{\boldsymbol{A}}$ $\sigma^2 \leq 1$

$$\Pr\left[\kappa(\boldsymbol{A}) \geq x\right] \leq \frac{n}{x}\left(\frac{1}{4\sqrt{2\pi n}} + 7\left(5 + \frac{4\left\|\bar{\boldsymbol{A}}\right\|_2^2(1+\log n)}{\sigma^2 n}\right)^{1/2}\right).$$

When $\left\|\bar{\boldsymbol{A}}\right\|_2 \leq \sqrt{n}$, his result implies

$$\Pr\left[\kappa(\boldsymbol{A}) \geq x\right] \leq O\left(\frac{n\log n}{x\sigma}\right),$$

we conjecture the following stronger statement is true.

CONJECTURE 4. $\bar{\boldsymbol{A}}$ $n \times n$ $\left\|\bar{\boldsymbol{A}}\right\|_2 \leq \sqrt{n}$ $\boldsymbol{A}$ $\bar{\boldsymbol{A}}$ $\sigma^2 \leq 1$

$$\Pr\left[\kappa(\boldsymbol{A}) \geq x\right] \leq O\left(\frac{n}{x\sigma}\right).$$

Appendix A. Gaussian random variables.

LEMMA A.1. X 0 1 $k \geq 1$

$$\Pr\left[X \geq k\right] \leq \frac{1}{\sqrt{2\pi}}\frac{e^{-\frac{1}{2}k^2}}{k}.$$

We have

$$\Pr\left[X \geq k\right] = \frac{1}{\sqrt{2\pi}}\int_k^\infty e^{-\frac{1}{2}x^2}\,dx$$

putting $t = \frac{1}{2}x^2$,

$$\begin{aligned}
&= \frac{1}{\sqrt{2\pi}}\int_{\frac{1}{2}k^2}^\infty \frac{e^{-t}}{\sqrt{2t}}\,dt \\
&\leq \frac{1}{\sqrt{2\pi}}\int_{\frac{1}{2}k^2}^\infty \frac{e^{-t}}{k}\,dt \\
&= \frac{1}{\sqrt{2\pi}}\frac{e^{-\frac{1}{2}k^2}}{k}. \qquad \square
\end{aligned}$$

LEMMA A.2. $\boldsymbol{x}$ d σ^2 $\boldsymbol{t}$ λ

$$\Pr\left[\left|\boldsymbol{t}^T\boldsymbol{x} - \lambda\right| \leq \epsilon\right] \leq \sqrt{\frac{2}{\pi}}\frac{\epsilon}{\sigma}.$$

LEMMA A.3. Let $g_1, \ldots, g_n$ be [illegible] 0 [illegible] 1 [illegible]

$$\mathrm{E}[\max_i |g_i|] \leq \sqrt{2\ln(\max(n,2))} + \frac{1}{\sqrt{2\pi}\ln(\max(n,2))}.$$

PROOF. For any $a \geq 1$,

$$\begin{aligned}
\mathrm{E}[\max_i |g_i|] &= \int_{t=0}^{\infty} \Pr\left[\max_i |g_i| \geq t\right] dt \\
&\leq \int_{t=0}^{a} 1\, dt + \int_{a}^{\infty} n \Pr\left[|g_1| \geq t\right] dt \\
&\leq a + \int_{a}^{\infty} n \frac{2}{\sqrt{2\pi}} \frac{e^{-\frac{1}{2}t^2}}{t}\, dt \qquad \text{(applying Lemma A.1)} \\
&= a + \frac{2n}{\sqrt{2\pi}} \int_{a}^{\infty} \frac{e^{-\frac{1}{2}t^2}}{t^2}\, d\left(\frac{1}{2}t^2\right) \\
&\leq a + \frac{2n}{\sqrt{2\pi}} \frac{1}{a^2} \int_{a}^{\infty} e^{-\frac{1}{2}t^2}\, d\left(\frac{1}{2}t^2\right) \\
&= a + \frac{2n}{\sqrt{2\pi}} \frac{1}{a^2} e^{-\frac{1}{2}a^2}.
\end{aligned}$$

Setting $a = \sqrt{2\ln(\max(n,2))}$, which is greater than 1 for all $n \geq 1$, we obtain the following upper bound on the expectation:

$$\begin{aligned}
&\sqrt{2\ln(\max(n,2))} + \frac{2n}{\sqrt{2\pi}} \frac{1}{2\ln(\max(n,2))} \frac{1}{\max(n,2)} \leq \sqrt{2\ln(\max(n,2))} \\
&\quad + \frac{1}{\sqrt{2\pi}\ln(\max(n,2))}. \qquad \square
\end{aligned}$$

LEMMA A.4 (expectation of reciprocal of the 1-norm of a Gaussian vector). Let $\bar{\boldsymbol{a}}$ be [illegible] $\mathbb{R}^n$ [illegible] $n \geq 2$ [illegible] $\boldsymbol{a}$ [illegible] $\bar{\boldsymbol{a}}$ [illegible] σ^2 [illegible]

$$\mathrm{E}\left[\frac{1}{\|\boldsymbol{a}\|_1}\right] \leq \frac{2}{n\sigma}.$$

PROOF. Let $\boldsymbol{a} = (a_1, \ldots, a_n)$. It is clear that the expectation of $1/\|\boldsymbol{a}\|_1$ is maximized if $\bar{\boldsymbol{a}} = \boldsymbol{0}$, so we will make this assumption. Without loss of generality, we also assume $\sigma^2 = 1$. For general σ, we can simply scale the bound by the factor $1/\sigma$.

Recall that the Laplace transform of a positive random variable X is defined by

$$\mathcal{L}[X](t) = \mathrm{E}_X[e^{-tX}]$$

and the expectation of the reciprocal of a random variable is simply the integral of its Laplace transform.

Let X be the absolute value of a standard normal random variable. The Laplace transform of X is given by

$$\begin{aligned}\mathcal{L}[X](t) &= \sqrt{\frac{2}{\pi}} \int_0^\infty e^{-tx} e^{-\frac{1}{2}x^2}\, dx \\ &= \sqrt{\frac{2}{\pi}} e^{\frac{1}{2}t^2} \int_0^\infty e^{-\frac{1}{2}(x+t)^2}\, dx \\ &= \sqrt{\frac{2}{\pi}} e^{\frac{1}{2}t^2} \int_t^\infty e^{-\frac{1}{2}x^2}\, dx \\ &= e^{\frac{1}{2}t^2} \operatorname{erfc}\left(\frac{t}{\sqrt{2}}\right).\end{aligned}$$

Taking second derivatives and applying the inequality (c.f., [AS64, 26.2.13])

$$\frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-\frac{1}{2}x^2}\, dx \geq \frac{e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}} \frac{1}{x + 1/x},$$

we find that $e^{\frac{1}{2}t^2} \operatorname{erfc}(\frac{t}{\sqrt{2}})$ is convex.

We now set a constant $c = 2.4$ and set α to satisfy

$$1 - \frac{\sqrt{c/\pi}}{\alpha} = e^{\frac{1}{2}(c/\pi)} \operatorname{erfc}\left(\frac{\sqrt{c/\pi}}{\sqrt{2}}\right).$$

Numerically, we find that $\alpha \approx 1.9857 < 2$.

As $e^{\frac{1}{2}t^2} \operatorname{erfc}(\frac{t}{\sqrt{2}})$ is convex, we have the upper bound

$$e^{\frac{1}{2}t^2} \operatorname{erfc}\left(\frac{t}{\sqrt{2}}\right) \leq 1 - \frac{t}{\alpha}, \quad \text{for} \quad 0 \leq t \leq \sqrt{c/\pi}.$$

For $t > \sqrt{c/\pi}$, we apply the upper bound

$$e^{\frac{1}{2}t^2} \operatorname{erfc}\left(\frac{t}{\sqrt{2}}\right) \leq \sqrt{\frac{2}{\pi}} \frac{1}{t},$$

which follows from Lemma A.1.

We now have

$$\begin{aligned}\mathrm{E}\left[\frac{1}{\|\boldsymbol{a}\|_1}\right] &= \int_0^\infty \left(e^{\frac{1}{2}t^2} \operatorname{erfc}(t/\sqrt{2})\right)^n\, dt \\ &\leq \int_0^{\sqrt{c/\pi}} \left(1 - \frac{t}{\alpha}\right)^n\, dt + \int_{\sqrt{c/\pi}}^\infty \left(\sqrt{\frac{2}{\pi}} \frac{1}{t}\right)^n\, dt \\ &\leq \frac{\alpha}{n+1} + \sqrt{\frac{2}{\pi}} \frac{(2/c)^{(n-1)/2}}{n-1} \\ &< \frac{2}{n+1} + \sqrt{\frac{2}{\pi}} \frac{(2/c)^{(n-1)/2}}{n-1} \\ &\leq \frac{2}{n-1},\end{aligned}$$

for $n \geq 2$. To verify this last equality, one can multiply through by $(n+1)(n-1)$ to obtain

$$\sqrt{\frac{2}{\pi}}(n+1)(2/c)^{(n-1)/2} \leq 4,$$

which one can verify by taking the derivative of the left-hand side to find the point where it is maximized, $n = (2 + \ln(5/6))/\ln(6/5)$. ∎

Appendix B. Random point on a sphere.

LEMMA B.1. [illegible] $d \geq 2$ [illegible] $(u_1, \ldots, u_d)$ [illegible] $\mathbb{R}^d$ [illegible] $c \leq 1$

$$\Pr\left[|u_1| \geq \sqrt{\frac{c}{d}}\right] \geq \Pr\left[|G| \geq \sqrt{c}\right],$$

[illegible] G [illegible] 1 [illegible] 0 [illegible]. We may obtain a random unit vector by choosing d independent Gaussian random variables of variance 1 and mean 0, $x_1, \ldots, x_d$, and setting

$$u_i = \frac{x_i}{\sqrt{x_1^2 + \cdots + x_d^2}}.$$

We have

$$\begin{aligned}
\Pr\left[u_1^2 \geq \frac{c}{d}\right] &= \Pr\left[\frac{x_1^2}{x_1^2 + \cdots + x_d^2} \geq \frac{c}{d}\right] \\
&= \Pr\left[\frac{(d-1)x_1^2}{x_2^2 + \cdots + x_d^2} \geq \frac{(d-1)c}{d-c}\right] \\
&\geq \Pr\left[\frac{(d-1)x_1^2}{x_2^2 + \cdots + x_d^2} \geq c\right], \quad \text{since } c \leq 1.
\end{aligned}$$

We now note that

$$t_d \stackrel{\text{def}}{=} \frac{\sqrt{(d-1)}x_1}{\sqrt{x_2^2 + \cdots + x_d^2}}$$

is a random variable distributed according to the t-distribution with $d-1$ degrees of freedom. The lemma now follows from the fact (cf., [JKB95, Chapter 28, section 2] or [AS64, 26.7.5]) that, for $c > 0$,

$$\Pr\left[t_d > \sqrt{c}\right] \geq \Pr\left[G > \sqrt{c}\right],$$

and that the distributions of t_d and G are symmetric about the origin. ∎

Appendix C. Combination lemmas.

LEMMA C.1. [illegible] A [illegible] B [illegible]

1. $\Pr[A \geq x] \leq f(x)$ [illegible]
2. $\Pr[B \geq x | A] \leq g(x)$

[illegible] g [illegible] $\lim_{x\to\infty} g(x) = 0$ [illegible]

$$\Pr[AB \geq x] \leq \int_0^\infty f\left(\frac{x}{t}\right)(-g'(t))\,dt.$$

Let μ_A denote the probability measure associated with A. We have

$$\begin{aligned}\Pr[AB \geq x] &= \int_0^\infty \Pr_B[B \geq x/s | A]\, d\mu_A(s) \\ &\leq \int_0^\infty g\left(\frac{x}{s}\right) d\mu_A(s),\end{aligned}$$

integrating by parts,

$$\begin{aligned}&= \int_0^\infty \Pr[A \geq s]\, \frac{d}{ds} g\left(\frac{x}{s}\right) ds \\ &\leq \int_0^\infty f(s) \frac{d}{ds} g\left(\frac{x}{s}\right) ds,\end{aligned}$$

setting $t = x/s$

$$= \int_0^\infty f\left(\frac{x}{t}\right)(-g'(t))\, dt. \qquad \square$$

COROLLARY C.2 (linear-linear). A B

1. $\Pr[A \geq x] \leq \frac{\alpha}{x}$
2. $\Pr[B \geq x | A] \leq \frac{\beta}{x}$

$\alpha, \beta > 0$

$$\Pr[AB \geq x] \leq \frac{\alpha\beta}{x}\left(1 + \max\left(0, \ln\left(\frac{x}{\alpha\beta}\right)\right)\right).$$

As the probability of an event can be at most 1,

$$\begin{aligned}\Pr[A \geq x] &\leq \min\left(\frac{\alpha}{x}, 1\right) \stackrel{\text{def}}{=} f(x), \text{ and} \\ \Pr[B \geq x] &\leq \min\left(\frac{\beta}{x}, 1\right) \stackrel{\text{def}}{=} g(x).\end{aligned}$$

Applying Lemma C.1 while observing

- $g'(t) = 0$ for $t \in [0, \beta]$ and
- $f(x/t) = 1$ for $t \geq x/\alpha$,

we obtain

$$\begin{aligned}\Pr[AB \geq x] &\leq \int_0^\beta \frac{\alpha t}{x} \cdot 0\, dt + \max\left(0, \int_\beta^{x/\alpha} \frac{\alpha t}{x}\frac{\beta}{t^2}\, dt\right) + \int_{x/\alpha}^\infty \frac{\beta}{t^2}\, dt \\ &= \max\left(0, \frac{\alpha\beta}{x}\int_\beta^{x/\alpha} \frac{dt}{t}\right) + \frac{\alpha\beta}{x} \\ &= \frac{\alpha\beta}{x}\left(1 + \max\left(0, \ln\left(\frac{x}{\alpha\beta}\right)\right)\right),\end{aligned}$$

where the max appears in case $x/\alpha < \beta$. $\square$

COROLLARY C.3. A B

1. $\forall x \geq 1/\sigma \;\; \Pr[A \geq x] \leq \min\left(1, \frac{\alpha+\beta\sqrt{\ln x\sigma}}{\sigma x}\right)$
2. $\Pr[B \geq x|A] \leq \frac{\gamma}{x\sigma}$

$\alpha \geq 1$ $\beta, \gamma, \sigma > 0$

$$\forall x \geq \gamma/\sigma^2, \Pr[AB \geq x] \leq \frac{\alpha\gamma}{x\sigma^2}\left(1 + \left(\frac{2\beta}{3\alpha} + 1\right) ln^{3/2}\left(\frac{x\sigma^2}{\gamma}\right)\right).$$

. Define f and g by

$$f(x) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{for } x \leq \frac{\alpha}{\sigma} \\ \frac{\alpha+\beta\sqrt{\ln x\sigma}}{x\sigma} & \text{for } x > \frac{\alpha}{\sigma} \end{cases}$$

$$g(x) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{for } x \leq \frac{\gamma}{\sigma} \\ \frac{\gamma}{x\sigma} & \text{for } x > \frac{\gamma}{\sigma}. \end{cases}$$

Applying Lemma C.1 while observing

- $g'(t) = 0$ for $t \in \left[0, \frac{\gamma}{\sigma}\right]$, and
- $f(x/t) = 1$ for $t \geq x\sigma/\alpha$,

we obtain

$$\begin{aligned} \Pr[AB \geq x] &\leq \int_{\gamma/\sigma}^{x\sigma/\alpha} \frac{\alpha + \beta\sqrt{\ln(x\sigma/t)}}{x\sigma/t} \frac{\gamma}{t^2\sigma}\, dt + \int_{x\sigma/\alpha}^{\infty} \frac{\gamma}{\sigma t^2}\, dt \\ &= \int_{\gamma/\sigma}^{x\sigma/\alpha} \frac{\alpha + \beta\sqrt{\ln(x\sigma/t)}}{x\sigma^2} \frac{\gamma}{t}\, dt + \frac{\alpha\gamma}{x\sigma^2} \end{aligned}$$

(substituting $s = \sqrt{\ln(x\sigma/t)}, t = x\sigma e^{-s^2}$, which is defined as $x \geq \gamma/\sigma^2$,)

$$\begin{aligned} &= \int_{\sqrt{\ln(x\sigma^2/\gamma)}}^{\sqrt{\ln\alpha}} \frac{\alpha + \beta s}{x\sigma^2} \frac{\gamma}{x\sigma e^{-s^2}} x\sigma(-2se^{-s^2})\, ds + \frac{\alpha\gamma}{x\sigma^2} \\ &= \frac{\gamma}{x\sigma^2} \int_{\sqrt{\ln\alpha}}^{\sqrt{\ln(x\sigma^2/\gamma)}} 2s(\alpha + \beta s)\, ds + \frac{\alpha\gamma}{x\sigma^2} \\ &= \frac{\alpha\gamma}{x\sigma^2}\left(1 + \ln\left(\frac{x\sigma^2}{\alpha\gamma}\right) + \frac{2\beta}{3\alpha}\left(\ln^{3/2}\left(\frac{x\sigma^2}{\gamma}\right) - \ln^{3/2}\alpha\right)\right) \\ &\leq \frac{\alpha\gamma}{x\sigma^2}\left(1 + \left(\frac{2\beta}{3\alpha} + 1\right) ln^{3/2}\left(\frac{x\sigma^2}{\gamma}\right)\right), \end{aligned}$$

as $\alpha \geq 1$. □

LEMMA C.4 (linear-bounded expectation). A B C

$$\Pr[A \geq x] \leq \frac{\alpha}{x},$$

$\alpha > 0$

$$\forall A, \; \Pr[B \geq x|A] \leq \Pr[C \geq x].$$

$$\Pr[AB \geq x] \leq \frac{\alpha}{x} \mathrm{E}[C].$$

Let $g(x)$ be the distribution function of C. By Lemma C.1, we have

$$\begin{aligned}\Pr[AB \geq x] &\leq \int_0^\infty \left(\frac{\alpha t}{x}\right)(-(1-g)'(t))\,dt \\ &= \frac{\alpha}{x}\int_0^\infty t(g'(t))\,dt \\ &= \frac{\alpha}{x}\,\mathrm{E}[C]. \qquad \square\end{aligned}$$

COROLLARY C.5 (linear-chi). Let A [illegible] such that

$$\Pr[A \geq x] \leq \frac{\alpha}{x}.$$

[illegible] $\alpha > 0$ [illegible] $\boldsymbol{b}$ [illegible] d [illegible] σ^2 [illegible] t [illegible] $B = \|\boldsymbol{b}\|_2$ [illegible]

$$\Pr[AB \geq x] \leq \frac{\alpha\sqrt{\sigma^2 d + t^2}}{x}.$$

As $\mathrm{E}[B] \leq \sqrt{\mathrm{E}[B^2]}$, and it is known [KJ82, p. 277] that the expected value of B^2—the noncentral χ^2-distribution with noncentrality parameter $\|\bar{\boldsymbol{b}}\|_2^2$—is $\sigma^2 d + \|\bar{\boldsymbol{b}}\|_2^2$, the corollary follows from Lemma C.4. $\square$

LEMMA C.6 (linear to log). Let A [illegible] $A_0 \geq 1$ [illegible] $\alpha \geq 1$ [illegible] $x \geq A_0$

$$\Pr_A[A \geq x] \leq \frac{\alpha}{x}.$$

[illegible]

$$\mathrm{E}_A[\max(0, \ln A)] \leq \ln\max(A_0, \alpha) + 1.$$

[illegible]

$$\begin{aligned}\mathrm{E}_A[\max(0,\ln A)] &= \int_{x=0}^\infty \Pr_A[\max(0,\ln A) \geq x]\,dx \\ &\leq \int_{x=0}^{\ln\max(A_0,\alpha)} 1\,dx + \int_{x=\ln\max(A_0,\alpha)}^\infty \Pr_A[\ln A \geq x]\,dx \\ &\leq \int_{x=0}^{\ln\max(A_0,\alpha)} dx + \int_{x=\ln\max(A_0,\alpha)}^\infty \alpha e^{-x}dx \\ &\leq \ln\max(A_0,\alpha) + 1. \qquad \square\end{aligned}$$

Acknowledgments. We thank Alan Edelman for suggesting the name "smoothed analysis," for suggesting we examine growth factors, and for his continuing support of our efforts. We thank Juan Cuesta and Mario Wschebor for pointing out some mistakes in an early draft of this paper. We thank Felipe Cucker for bringing Wschebor's paper [Wsc04] to our attention. Finally, we thank the referees for their extraordinary efforts and many helpful suggestions.

REFERENCES

[ABB+99] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, LAPACK Users' Guide, 3rd ed., SIAM, Philadelphia, 1999.

[AS64] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, National Bureau of Standards Applied Mathematics Series, 55, U.S. Government Printing Office, Washington, DC, 1964. Tenth printing, with corrections, 1972.

[BD02] A. Blum and J. Dunagan, *Smoothed analysis of the perceptron algorithm for linear programming*, in Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms, San Francisco, SIAM, 2002, pp. 905–911.

[Blu89] L. Blum, *Lectures on a theory of computation and complexity over the reals (or an arbitrary ring)*, in Proceedings of the Complex Systems Summer School, Santa Fe, NM, 1989, pp. 1–47.

[Dem88] J. W. Demmel, *The probability that a numerical analysis problem is difficult*, Math. Comput., (1988), pp. 449–480.

[Dem97] J. W. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

[DS01] K. R. Davidson and S. J. Szarek, *Local operator theory, random matrices, and Banach spaces* in Handbook on the Geometry of Banach Spaces, W. B. Johnson and J. Lindenstrauss, eds., North Holland, Amsterdam, 2001, pp. 317–366.

[DST02] J. Dunagan, D. A. Spielman, and S.-H. Teng, *Smoothed analysis of Renegar's condition number for linear programming*, available at http://arxiv.org/abs/cs.DS/0302011, 2003.

[Ede88] A. Edelman, *Eigenvalues and condition numbers of random matrices*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 543–560.

[Ede92] A. Edelman, *Eigenvalue roulette and random test matrices*, in Linear Algebra for Large Scale and Real-Time Applications, Marc S. Moonen, Gene H. Golub, and Bart L. R. De Moor, eds., NATO ASI Series, 1992, pp. 365–368.

[GL83] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins Series in the Mathematical Sciences, The Johns Hopkins University Press, Baltimore, 1983.

[GLS91] M. Grötschel, L. Lovasz, and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, Springer-Verlag, Berlin, 1991.

[Hig90] N. Higham, *How accurate is Gaussian elimination?* in Proceedings of the 13th Dundee Conference, Dundee, Australia, Pitman Res. Notes Math. Ser., Longman Sci. Tech, Harlow, UK, 1990, pp. 137–154.

[JKB95] N. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, Vol. 2, John Wiley & Sons, New York, 1995.

[KJ82] S. Kotz and N. L. Johnson, eds., *Encyclopedia of Statistical Sciences*, Volume 6, John Wiley & Sons, New York, 1982.

[LT91] M. Ledoux and M. Talagrand, *Probability in Banach Spaces*, Springer-Verlag, Berlin, 1991.

[Ren95] J. Renegar, *Incorporating condition measures into the complexity theory of linear programming*, SIAM J. Optim., 5 (1995), pp. 506–524.

[Seg00] Y. Seginer, *The expected norm of random matrices*, Combin. Probab. Comput., 9 (2000), pp. 149–166.

[Sma97] S. Smale, *Complexity theory and numerical analysis*, Acta Numer., 6 (1997), pp. 523–551.

[Smo] D. Spielman, http://www.cs.yale.edu/homes/spielman/SmoothedAnalysis.

[ST03] D. Spielman and S. H. Teng, *Smoothed analysis of termination of linear programming algorithms*, Math. Program., 97 (2003), pp. 375–404.

[ST04] D. A. Spielman and S.-H. Teng, *Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time*, J. ACM, 51 (2004), pp. 385–463.

[TB97] L. N. Trefethen and D. Bau, *Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

[TS90] L. N. Trefethen and R. S. Schreiber, *Average-case stability of Gaussian elimination*, SIAM J. Matrix Anal. Appl., 11 (1990), pp 335–360.

[Wil61] J. H. Wilkinson, *Error analysis of direct methods of matrix inversion*, J. Assoc. Comput. Mach., 8 (1961), pp. 281–330.

[Wsc04] M. Wschebor, *Smoothed analysis of $\kappa(\boldsymbol{a})$*, J. of Complexity, 20 (2004), pp. 97–107.

[YC97] M.-C. Yeung and T. F. Chan, *Probabilistic analysis of Gaussian elimination without pivoting*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 499–517.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 2, pp. 477–502

FAST METHODS FOR ESTIMATING THE DISTANCE TO UNCONTROLLABILITY*

M. GU†, E. MENGI‡, M. L. OVERTON‡, J. XIA†, AND J. ZHU†

Abstract. The distance to uncontrollability for a linear control system is the distance (in the 2-norm) to the nearest uncontrollable system. We present an algorithm based on methods of Gu and Burke–Lewis–Overton that estimates the distance to uncontrollability to any prescribed accuracy. The new method requires $O(n^4)$ operations on average, which is an improvement over previous methods which have complexity $O(n^6)$, where n is the order of the system. Numerical experiments indicate that the new method is reliable in practice.

Key words. distance to uncontrollability, complex controllability radius, trisection, real eigenvalue extraction, shifted inverse iteration, shift-and-invert Arnoldi, Sylvester equation, Kronecker product

AMS subject classifications. 65F15, 93B05, 65K10

DOI. 10.1137/05063060X

1. Introduction.

Given $A \in \mathbb{C}^{n\times n}, B \in \mathbb{C}^{n\times m}$, the linear control system

$$\dot{x} = Ax + Bu \tag{1.1}$$

is controllable if for every pair of states $x_0, x_f \in \mathbb{C}^n$ there exists a continuous control function $u(t)$ to steer the initial state x_0 to the final state x_f within finite time. Equivalently, according to a well-known result by Kalman [8], the system (1.1) is controllable if the matrix $[A - \lambda I \;\; B]$ has full row rank for all $\lambda \in \mathbb{C}$.

To measure the conditioning of (1.1), the distance to uncontrollability was introduced in [12] as

$$\tau(A, B) = \min\{\|\Delta A \; \Delta B\| \; : \; (A + \Delta A, B + \Delta B) \text{ is uncontrollable}\}, \tag{1.2}$$

which was later shown to be equivalent to [4, 5]

$$\tau(A, B) = \min_{\lambda\in\mathbb{C}} \sigma_n([A - \lambda I \;\; B]), \tag{1.3}$$

where $\|\cdot\|$ denotes the 2-norm or Frobenius norm[1] and $\sigma_n([A - \lambda I \;\; B])$ denotes the nth largest singular value of the $n \times (n + m)$ matrix $[A - \lambda I \;\; B]$. This is a global nonsmooth optimization problem in two real variables α and β, the real and imaginary parts of λ. But note that $\sigma_n([A - \lambda I \;\; B])$ is not convex and may have many local minima, so standard optimization methods, which are guaranteed only to converge to a local minimum, will not yield reliable results in general.

*Received by the editors May 3, 2005; accepted for publication (in revised form) by N. Mastronardi December 21, 2005; published electronically June 21, 2006.
http://www.siam.org/journals/simax/28-2/63060.html

†Department of Mathematics, University of California at Berkeley, Berkeley, CA 94720 (mgu@math.berkeley.edu, jxia@math.berkeley.edu, zhujiang@math.berkeley.edu). The research of these authors was supported in part by the National Science Foundation grant DMS-0412049.

‡Courant Institute of Mathematical Sciences, New York University, New York, NY 10012 (mengi@cs.nyu.edu, overton@cs.nyu.edu). The research of these authors was supported in part by the National Science Foundation grant CCR-0204388.

[1]The definitions of $\tau(A, B)$ in terms of the Frobenius norm and the 2-norm are equivalent.

Gu [6] proposed a bisection method which can correctly estimate $\tau(A, B)$ within a factor of 2 in time polynomial in n. Throughout the paper, when we refer to operation counts we assume that the computation of the eigenvalues of a matrix or pencil is an atomic operation whose cost is cubic in the dimension. Burke, Lewis, and Overton [3] suggested a trisection variant to retrieve the distance to uncontrollability to any desired accuracy. The methods in these two papers are based on a simultaneous comparison of two estimates $\delta_1 > \delta_2$ with $\tau(A, B)$. More precisely, Gu derived a scheme that returns one of the inequalities

$$\tau(A, B) \leq \delta_1 \tag{1.4}$$

and

$$\tau(A, B) > \delta_2. \tag{1.5}$$

Even if both of the inequalities are satisfied, Gu's scheme returns information about only one of the inequalities. Gu's method depends on the extraction of the real eigenvalues of a pencil of size $2n^2 \times 2n^2$ and the imaginary eigenvalues of matrices of size $2n \times 2n$. Computationally the verification scheme is dominated by the extraction of the real eigenvalues of the generalized problem of size $2n^2 \times 2n^2$, which requires $O(n^6)$ operations if the standard QZ algorithm is used.

In this paper we present an alternative verification scheme for comparisons (1.4) and (1.5). In the new verification scheme we still need to find real eigenvalues of $2n^2 \times 2n^2$ matrices, so there is no asymptotic gain over Gu's verification scheme when we use the QR algorithm. Nevertheless, we show that the inverse of these $2n^2 \times 2n^2$ matrices shifted by a real number times the identity can be multiplied onto a vector efficiently by solving a Sylvester equation of size $2n$ with a cost of $O(n^3)$. Therefore, given a real number as the shift, by applying shifted inverse iteration or a shift-and-invert preconditioned Arnoldi method, the closest eigenvalue to the real number can be obtained by performing $O(n^3)$ operations. Motivated by the fact that we need only real eigenvalues, we provide two alternative ways to scan the real axis to find the desired eigenvalues. Both of the approaches require an upper bound on the norm of the input matrix (of size $2n^2 \times 2n^2$) as a parameter. For one of the approaches, which is based on a "divide and conquer" idea, choosing this parameter arbitrarily large does not affect the efficiency of the algorithm much. The efficiency of the other approach, which we name "adaptive progress," depends not only on this parameter significantly but also on another parameter that bounds the distance between the closest pair of eigenvalues from below. For the divide and conquer approach, we prove that extracting all of the real eigenvalues requires $O(n^4)$ operations on average and $O(n^5)$ operations in the worst case. For the adaptive progress approach such neat results are not immediate because of the dependence of the performance of the algorithm on the parameters. In practice we observe that the divide and conquer approach is the more efficient and more reliable method.

In section 2 we will review the trisection method for estimating $\tau(A, B)$ and Gu's scheme for verifying which one of (1.4) and (1.5) holds. In section 3 we present our modified eigenvalue problem for the same purpose and fast methods based on the shifted inverse iteration or shift-and-invert Arnoldi for solving it. Specifically, to extract all of the real eigenvalues, we discuss two search strategies: an adaptive progress approach and a divide and conquer approach. The effectiveness and reliability of the methods are demonstrated by the numerical examples in section 4.

2. Trisection and Gu's verification scheme.

2.1. Bisection and trisection. The problem of computing the distance to uncontrollability is equivalent to the minimization of $\sigma_n([A - \lambda I \quad B])$ over the entire complex plane. Gu [6] proposed the first polynomial-time estimation scheme. Burke, Lewis, and Overton [3] later suggested a trisection version to retrieve the distance to uncontrollability to an arbitrary accuracy. Given two real numbers $\delta_1 > \delta_2$, at each iteration both of the algorithms alter an upper bound or a lower bound depending on which of the inequalities (1.4) and (1.5) holds. This test is based on the following theorem [6], which is a consequence of the fact that singular values are well-conditioned (in the absolute sense).

THEOREM 2.1 (see Gu [6]). $\delta > \tau(A,B)$ $\eta \in [0, 2(\delta - \tau(A,B))]$ α β

$$\delta \in \sigma\left([A - (\alpha + \beta i)I, B]\right) \quad \delta \in \sigma\left([A - (\alpha + \eta + \beta i)I, B]\right), \tag{2.1}$$

$\sigma(\cdot)$

We shall describe two alternative ways of verifying the existence of a pair α and β satisfying (2.1) for a given δ and η in subsections 2.2 and 3.1. Suppose we set $\delta_1 = \delta$ and $\delta_2 = \delta - \eta/2$. The theorem above implies that when no pair satisfying (2.1) exists the inequality $\eta > 2(\delta - \tau(A,B))$ is satisfied, so condition (1.5) holds. On the other hand, when a pair exists, then by definition (1.3) we can conclude (1.4).

Gu's bisection algorithm (Algorithm 1) keeps only an upper bound on the distance to uncontrollability. It refines the upper bound until condition (1.5) is satisfied. Notice that in Algorithm 1, $\delta = \eta = \delta_1$. At termination the distance to uncontrollability lies within factor of 2 of δ_1, with $\delta_1/2 < \tau(A,B) \leq 2\delta_1$.

Algorithm 1 Gu's bisection estimation algorithm

Call:	$\delta_1 \leftarrow$ `Bisection`(A,B).
Input:	$A \in \mathbb{C}^{n\times n}$ and $B \in \mathbb{C}^{n\times m}$ with $m \leq n$.
Output:	A scalar δ_1 satisfying $\delta_1/2 < \tau(A,B) \leq 2\delta_1$.

1. Initialize the estimate as $\delta_1 \leftarrow \sigma_n([A \;\; B])/2$.
   repeat
     $\delta_2 \leftarrow \frac{\delta_1}{2}$.
     Apply Gu's test.
     if (1.4) is verified **then**
       $\delta_1 \leftarrow \delta_2$.
       done $\leftarrow$ FALSE.
     else
       % Otherwise (1.5) is verified.
       done $\leftarrow$ TRUE.
     end if
   until done = TRUE
2. Return δ_1.

To obtain the distance to uncontrollability with better accuracy, Burke, Lewis, and Overton [3] proposed a trisection variant. The trisection algorithm (Algorithm 2) bounds $\tau(A,B)$ by an interval $[l,u]$ and reduces the length of this interval by a factor of $\frac{2}{3}$ at each iteration. Thus it can compute $\tau(A,B)$ to any desired accuracy in $O(n^6)$ operations which is the cost of Gu's test, as described next.

Algorithm 2 Trisection variant of Algorithm 1

Call: $[l,\, u] \leftarrow$ `Trisection`(A,B,ϵ).
Input: $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{n\times m}$ with $m \le n$, and a tolerance $\epsilon > 0$.
Output: Scalars l and u satisfying $l < \tau(A,B) \le u$ and $u - l < \epsilon$.

1. Initialize the lower bound as $l \leftarrow 0$ and the upper bound as $u \leftarrow \sigma_n([A\ \ B])$.
   repeat
   $\delta_1 \leftarrow l + \frac{2}{3}(u-l)$
   $\delta_2 \leftarrow l + \frac{1}{3}(u-l)$
   Apply Gu's test.
   if (1.4) is verified **then**
   $u \leftarrow \delta_1$.
   else
   % Otherwise (1.5) is verified.
   $l \leftarrow \delta_2$.
   end if
   until $u - l < \epsilon$
2. Return l and u.

2.2. Gu's verification scheme. By means of Gu's test we can numerically verify whether a real pair of solutions to (2.1) exists. Equation (2.1) in Theorem 2.1 implies that there exist nonzero vectors $\binom{x}{y}$, z, $\binom{x}{\hat{y}}$, and $\hat{z}$ such that

$$(2.2a)\qquad (\,A-(\alpha+\beta i)I \quad B\,)\begin{pmatrix}x\\ y\end{pmatrix} = \delta z,\ \begin{pmatrix}A^*-(\alpha-\beta i)I\\ B^*\end{pmatrix} z = \delta\begin{pmatrix}x\\ y\end{pmatrix},$$

$$(2.2b)\qquad (\,A-(\alpha+\eta+\beta i)I \quad B\,)\begin{pmatrix}\hat{x}\\ \hat{y}\end{pmatrix} = \delta\hat{z},\ \begin{pmatrix}A^*-(\alpha+\eta-\beta i)I\\ B^*\end{pmatrix}\hat{z} = \delta\begin{pmatrix}\hat{x}\\ \hat{y}\end{pmatrix}.$$

These equations can be rewritten as

$$(2.3a)\qquad \begin{pmatrix} -\delta I & A-\alpha I & B\\ A^*-\alpha I & -\delta I & 0\\ B^* & 0 & -\delta I\end{pmatrix}\begin{pmatrix}z\\ x\\ y\end{pmatrix} = \beta i\begin{pmatrix}0 & I & 0\\ -I & 0 & 0\\ 0 & 0 & 0\end{pmatrix}\begin{pmatrix}z\\ x\\ y\end{pmatrix}$$

and

$$(2.3b)\qquad \begin{pmatrix} -\delta I & A-(\alpha+\eta)I & B\\ A^*-(\alpha+\eta)I & -\delta I & 0\\ B^* & 0 & -\delta I\end{pmatrix}\begin{pmatrix}\hat{z}\\ \hat{x}\\ \hat{y}\end{pmatrix} = \beta i\begin{pmatrix}0 & I & 0\\ -I & 0 & 0\\ 0 & 0 & 0\end{pmatrix}\begin{pmatrix}\hat{z}\\ \hat{x}\\ \hat{y}\end{pmatrix}.$$

Furthermore using the QR factorization

$$(2.4)\qquad \begin{pmatrix}B\\ -\delta I\end{pmatrix} = \begin{pmatrix}Q_{11} & Q_{12}\\ Q_{21} & Q_{22}\end{pmatrix}\begin{pmatrix}R\\ 0\end{pmatrix}$$

these problems can be reduced to standard eigenvalue problems of size $2n \times 2n$, i.e., the eigenvalues of the pencils in (2.3a) and in (2.3b) are the same as the eigenvalues of the matrices

$$(2.5a)\qquad \begin{pmatrix} A-\alpha I & BQ_{22}-\delta Q_{12}\\ \delta Q_{12}^{-1} & -Q_{12}^{-1}(A^*-\alpha I)Q_{12}\end{pmatrix}$$

and

$$\begin{pmatrix} A-(\alpha+\eta)I & BQ_{22}-\delta Q_{12} \\ \delta Q_{12}^{-1} & -Q_{12}^{-1}(A^*-(\alpha+\eta)I)Q_{12} \end{pmatrix}, \tag{2.5b}$$

respectively. In order for (2.1) to have at least one real solution (α, β), these two matrices must share a common pure imaginary eigenvalue βi. This requires a $2n^2 \times 2n^2$ generalized eigenvalue problem to have a real eigenvalue α (see [6]). For a given δ and η, we check whether the latter generalized eigenvalue problem has any real eigenvalue α. If it does, then we check the existence of a real eigenvalue α for which the matrices (2.5a) and (2.5b) share a common pure imaginary eigenvalue βi. There exists a pair of α and β satisfying (2.1) if and only if this process succeeds.

3. Modified fast verification scheme. It turns out that Gu's verification scheme can be simplified. In this modified scheme the $2n^2 \times 2n^2$ generalized eigenvalue problems whose real eigenvalues are sought in Gu's scheme are replaced by $2n^2 \times 2n^2$ standard eigenvalue problems, and the $2n \times 2n$ standard eigenvalue problems (2.5a) and (2.5b) whose imaginary eigenvalues are sought are replaced by new $2n \times 2n$ standard eigenvalue problems that do not require the computation of QR factorizations.

The simplification of the problem of size $2n^2 \times 2n^2$ is significant, as the inverse of the new matrix of size $2n^2 \times 2n^2$ (whose real eigenvalues are sought) times a vector can be computed in a cheap manner by solving a Sylvester equation of size $2n \times 2n$ with a cost of $O(n^3)$. As a consequence the closest eigenvalue to a given complex point can be computed efficiently by applying shifted inverse iteration or shift-and-invert Arnoldi. We discuss how this idea can be extended to extract all of the real eigenvalues with an average cost of $O(n^4)$ and a worst case cost of $O(n^5)$, reducing the running time of each iteration of the bisection or the trisection algorithm asymptotically.

3.1. New generalized eigenvalue problem. According to (2.2a)

$$y = \frac{1}{\delta} B^* z$$

and the two equations in (2.2a) can be rewritten as

$$\begin{pmatrix} \hat{B} & A-\alpha I \\ A^*-\alpha I & -\delta I \end{pmatrix} \begin{pmatrix} z \\ x \end{pmatrix} = \beta i \begin{pmatrix} & I \\ -I & \end{pmatrix} \begin{pmatrix} z \\ x \end{pmatrix},$$

where $\hat{B} = \frac{BB^*}{\delta} - \delta I$. That is

$$H(\alpha) \begin{pmatrix} z \\ x \end{pmatrix} = \begin{pmatrix} -(A^*-\alpha I) & \delta I \\ \hat{B} & A-\alpha I \end{pmatrix} \begin{pmatrix} z \\ x \end{pmatrix} = \beta i \begin{pmatrix} z \\ x \end{pmatrix}. \tag{3.1a}$$

Similarly

$$H(\alpha+\eta) \begin{pmatrix} \hat{z} \\ \hat{x} \end{pmatrix} = \begin{pmatrix} -(A^*-(\alpha+\eta)I) & \delta I \\ \hat{B} & A-(\alpha+\eta)I \end{pmatrix} \begin{pmatrix} \hat{z} \\ \hat{x} \end{pmatrix} = \beta i \begin{pmatrix} \hat{z} \\ \hat{x} \end{pmatrix}. \tag{3.1b}$$

Both of the eigenvalue problems above are Hamiltonian, i.e., $JH(\alpha)$ and $JH(\alpha+\eta)$ are Hermitian where

$$J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \tag{3.2}$$

with $n \times n$ blocks. The Hamiltonian property implies that the matrices $H(\alpha+\eta)$ and $-H(\alpha+\eta)^*$ have the same set of eigenvalues. For $H(\alpha)$ and $H(\alpha+\eta)$ or equivalently $H(\alpha)$ and $-H(\alpha+\eta)^*$ to share a common pure eigenvalue βi, the following matrix equation

$$\begin{pmatrix} -(A^* - \alpha I) & \delta I \\ \hat{B} & A - \alpha I \end{pmatrix} X + X \begin{pmatrix} -(A^* - (\alpha+\eta) I) & \delta I \\ \hat{B} & A - (\alpha+\eta) I \end{pmatrix}^* = 0 \tag{3.3}$$

or equivalently

$$\begin{pmatrix} -A^* & \delta I \\ \hat{B} & A \end{pmatrix} X + X \begin{pmatrix} -(A - \eta I) & \hat{B} \\ \delta I & A^* - \eta I \end{pmatrix} = \alpha \left(\begin{pmatrix} -I & 0 \\ 0 & I \end{pmatrix} X + X \begin{pmatrix} -I & 0 \\ 0 & I \end{pmatrix} \right) \tag{3.4}$$

must have a nonzero solution $X \in \mathbb{C}^{2n \times 2n}$. Partition $X = \left(\begin{smallmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{smallmatrix}\right)$, and let $\mathbf{vec}(X)$ denote the vector formed by stacking the column vectors of X. We will use the following properties of Kronecker products:

$$\mathbf{vec}(AX) = (I \otimes A)\mathbf{vec}(X), \ \mathbf{vec}(XA) = (A^T \otimes I)\mathbf{vec}(X).$$

Now we can rewrite (3.4) as

$$\begin{pmatrix} -A_1^* - A_2^T & \delta I & \delta I & 0 \\ B_2^T & -A_1^* + \bar{A}_2 & 0 & \delta I \\ B_1 & 0 & A_1 - A_2^T & \delta I \\ 0 & B_1 & B_2^T & A_1 + \bar{A}_2 \end{pmatrix} \begin{pmatrix} \mathbf{vec}(X_{11}) \\ \mathbf{vec}(X_{12}) \\ \mathbf{vec}(X_{21}) \\ \mathbf{vec}(X_{22}) \end{pmatrix} = \begin{pmatrix} -2\alpha \mathbf{vec}(X_{11}) \\ 0 \\ 0 \\ 2\alpha \mathbf{vec}(X_{22}) \end{pmatrix}, \tag{3.5}$$

where $A_1 = I \otimes A$, $A_2 = (A - \eta I) \otimes I$, $B_1 = I \otimes \hat{B}$, $B_2 = \hat{B} \otimes I$, and $\bar{A}_2$ denotes the matrix obtained by taking the complex conjugate of A_2 entrywise.

The $(1,2)$, $(2,1)$ entries of both sides of (3.4) lead to

$$\begin{pmatrix} B_2^T & \delta I & -A_1^* + \bar{A}_2 & 0 \\ B_1 & \delta I & 0 & A_1 - A_2^T \end{pmatrix} \begin{pmatrix} \mathbf{vec}(X_{11}) \\ \mathbf{vec}(X_{22}) \\ \mathbf{vec}(X_{12}) \\ \mathbf{vec}(X_{21}) \end{pmatrix} = 0.$$

We then have

$$\begin{pmatrix} \mathbf{vec}(X_{12}) \\ \mathbf{vec}(X_{21}) \end{pmatrix} = - \begin{pmatrix} -A_1^* + \bar{A}_2 & 0 \\ 0 & A_1 - A_2^T \end{pmatrix}^{-1} \begin{pmatrix} B_2^T & \delta I \\ B_1 & \delta I \end{pmatrix} \begin{pmatrix} \mathbf{vec}(X_{11}) \\ \mathbf{vec}(X_{22}) \end{pmatrix} \tag{3.6}$$

under the assumption that A does not have two eigenvalues that differ by η, in which case the matrix $A_1 - A_2^T$ is invertible and therefore the inverted matrix in (3.6) exists. This assumption is generically satisfied in practice (numerical troubles that occur when η is small are discussed in section 5). On the other hand the $(1,1)$, $(2,2)$ entries

of both sides of (3.4) give

$$\begin{pmatrix} -A_1^* - A_2^T & 0 & \delta I & \delta I \\ 0 & A_1 + \bar{A}_2 & B_1 & B_2^T \end{pmatrix} \begin{pmatrix} \mathbf{vec}(X_{11}) \\ \mathbf{vec}(X_{22}) \\ \mathbf{vec}(X_{12}) \\ \mathbf{vec}(X_{21}) \end{pmatrix} = 2\alpha \begin{pmatrix} -I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{vec}(X_{11}) \\ \mathbf{vec}(X_{22}) \\ \mathbf{vec}(X_{12}) \\ \mathbf{vec}(X_{21}) \end{pmatrix},$$

which can be simplified with (3.6) to

$$\left[\begin{pmatrix} -A_1^* - A_2^T & 0 \\ 0 & A_1 + \bar{A}_2 \end{pmatrix} - \begin{pmatrix} \delta I & \delta I \\ B_1 & B_2^T \end{pmatrix} \begin{pmatrix} -A_1^* + \bar{A}_2 & 0 \\ 0 & A_1 - A_2^T \end{pmatrix}^{-1} \begin{pmatrix} B_2^T & \delta I \\ B_1 & \delta I \end{pmatrix} \right] \begin{pmatrix} \mathbf{vec}(X_{11}) \\ \mathbf{vec}(X_{22}) \end{pmatrix} = 2\alpha \begin{pmatrix} -I & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} \mathbf{vec}(X_{11}) \\ \mathbf{vec}(X_{22}) \end{pmatrix},$$

i.e.,

$$\mathcal{A}v = \alpha v, \tag{3.7}$$

where

$$\mathcal{A} = \frac{1}{2} \left[\begin{pmatrix} A_1^* + A_2^T & 0 \\ 0 & A_1 + \bar{A}_2 \end{pmatrix} - \begin{pmatrix} -\delta I & -\delta I \\ B_1 & B_2^T \end{pmatrix} \begin{pmatrix} -A_1^* + \bar{A}_2 & 0 \\ 0 & A_1 - A_2^T \end{pmatrix}^{-1} \begin{pmatrix} B_2^T & \delta I \\ B_1 & \delta I \end{pmatrix} \right]. \tag{3.8}$$

For the verification of a pair α and β satisfying (2.1), we first solve the eigenvalue problem (3.7). If there exists a real eigenvalue α of this problem such that the matrices $H(\alpha)$ and $H(\alpha + \eta)$ share a common imaginary eigenvalue, then the verification succeeds.

3.2. Inverse iteration. The eigenvalue problem in (3.7) is a simplified version of the generalized eigenvalue problem in [6]. This is a problem of finding the real eigenvalues of a nonsymmetric matrix. The implementation[2] of the trisection algorithm of [3] uses the MATLAB function `eig` to compute the eigenvalues of that generalized eigenvalue problem with a cost of $O(n^6)$ and therefore does not exploit the fact that we need only the real eigenvalues of the generalized problem. In section 3.3 we discuss two strategies to extract the real eigenvalues of a given matrix $\mathcal{X}$ that is preferable to `eig` when the closest eigenvalue of $\mathcal{X}$ to a given point can be obtained efficiently.

In this section we show how one can compute the closest eigenvalue of $\mathcal{A}$ to a given point in the complex plane in $O(n^3)$ time. This is due to the fact that given a shift ν and a vector $u \in \mathbb{C}^{2n^2}$, the multiplication $(\mathcal{A} - \nu I)^{-1}u$ can be performed by solving a Sylvester equation of size $2n \times 2n$ which is derived next. Therefore, shifted inverse iteration or shift-and-invert Arnoldi can locate the closest eigenvalue efficiently.

[2] http://www.cs.nyu.edu/faculty/overton/software/uncontrol/.

3.2.1. Computing $\mathcal{A}^{-1}u$. We first derive the Sylvester equation whose solution yields $v = \mathcal{A}^{-1}u$, where $u = \binom{u_1}{u_2}$, $v = \binom{v_1}{v_2}$, and $u_1, u_2, v_1, v_2 \in \mathbb{C}^{n^2}$. We can also write

$$\mathcal{A}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}. \tag{3.9}$$

Let

$$w = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} -A_1^* + \bar{A}_2 & 0 \\ 0 & A_1 - A_2^T \end{pmatrix}^{-1} \begin{pmatrix} B_2^T & \delta I \\ B_1 & \delta I \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}. \tag{3.10}$$

Then (3.9) can be rewritten as

$$\begin{pmatrix} A_1^* + A_2^T & 0 \\ 0 & A_1 + \bar{A}_2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} - \begin{pmatrix} -\delta I & -\delta I \\ B_1 & B_2^T \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = 2 \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}. \tag{3.11}$$

Equations (3.10) and (3.11) can then be combined into one linear system

$$\begin{pmatrix} A_1^* + A_2^T & \delta I & \delta I & 0 \\ B_2^T & A_1^* - \bar{A}_2 & 0 & \delta I \\ B_1 & 0 & -A_1 + A_2^T & \delta I \\ 0 & -B_1 & -B_2^T & A_1 + \bar{A}_2 \end{pmatrix} \begin{pmatrix} v_1 \\ w_1 \\ w_2 \\ v_2 \end{pmatrix} = 2 \begin{pmatrix} u_1 \\ 0 \\ 0 \\ u_2 \end{pmatrix}, \tag{3.12}$$

which is analogous to (3.5). By introducing vector forms

$$u = \begin{pmatrix} \mathbf{vec}(U_1) \\ \mathbf{vec}(U_2) \end{pmatrix}, \; v = \begin{pmatrix} \mathbf{vec}(V_1) \\ \mathbf{vec}(V_2) \end{pmatrix}, \; w = \begin{pmatrix} \mathbf{vec}(W_1) \\ \mathbf{vec}(W_2) \end{pmatrix},$$

we get a matrix equation similar to (3.4),

$$\begin{pmatrix} A^* & \delta I \\ \hat{B} & -A \end{pmatrix} Z + Z \begin{pmatrix} A - \eta I & \hat{B} \\ \delta I & -A^* + \eta I \end{pmatrix} = 2 \begin{pmatrix} U_1 & 0 \\ 0 & -U_2 \end{pmatrix}, \tag{3.13}$$

where

$$Z = \begin{pmatrix} V_1 & W_1 \\ W_2 & V_2 \end{pmatrix}. \tag{3.14}$$

Equation (3.13) is a $2n \times 2n$ Sylvester equation. By using a Sylvester equation solver (such as the LAPACK routine `dtrsyl` [1]) we can solve for Z at $O(n^3)$ cost and thus obtain $v = \mathcal{A}^{-1}u$.

3.2.2. Computing $(\mathcal{A} - \nu I)^{-1}u$. The derivation of the Sylvester equation for the multiplication $\mathcal{A}^{-1}u$ easily extends to the multiplication $(\mathcal{A} - \nu I)^{-1}u$ for a given shift ν. We alternatively rewrite the multiplication as

$$(\mathcal{A} - \nu I)\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \tag{3.15}$$

and introduce $w = \binom{w_1}{w_2}$ as in the previous section. We end up with

$$\begin{pmatrix} A_1^* + A_2^T - 2\nu I & \delta I & \delta I & 0 \\ B_2^T & A_1^* - \bar{A}_2 & 0 & \delta I \\ B_1 & 0 & -A_1 + A_2^T & \delta I \\ 0 & -B_1 & -B_2^T & A_1 + \bar{A}_2 - 2\nu I \end{pmatrix} \begin{pmatrix} v_1 \\ w_1 \\ w_2 \\ v_2 \end{pmatrix} = 2 \begin{pmatrix} u_1 \\ 0 \\ 0 \\ u_2 \end{pmatrix}, \tag{3.16}$$

which is analogous to (3.12). In terms of a matrix equation, we obtain

$$\begin{pmatrix} A^* - \nu I & \delta I \\ \hat{B} & -A + \nu I \end{pmatrix} Z + Z \begin{pmatrix} A - (\eta + \nu) I & \hat{B} \\ \delta I & -A^* + (\eta + \nu) I \end{pmatrix} = 2 \begin{pmatrix} U_1 & 0 \\ 0 & -U_2 \end{pmatrix}, \tag{3.17}$$

where Z is as defined in (3.14). Equation (3.17) is identical to (3.13) except that A is replaced by $A - \nu I$ in (3.13) and thus can be solved at $O(n^3)$ cost.

3.3. Real eigenvalue searching strategies. In this section we seek the real eigenvalues of a given matrix $\mathcal{X} \in \mathbb{C}^{q \times q}$. The iterative methods here are preferable to the standard ways of computing eigenvalues such as the QR algorithm when $(\mathcal{X} - \nu I)^{-1} u$ for a given shift $\nu \in \mathbb{R}$ and a given vector $u \in \mathbb{C}^q$ is efficiently computable. In particular, as discussed in the previous section, this is the case for $\mathcal{A}$.

Throughout this section we will assume the existence of a reliable implementation of the shifted inverse iteration or a shift-and-invert Arnoldi method that returns the closest eigenvalue to a given shift accurately. In practice we make use of the MATLAB function `eigs` (based on ARPACK [9, 10]). Additionally, we assume that an upper bound, D, on the norm of $\mathcal{X}$ is available and therefore we know that all of the real eigenvalues lie in the interval $[-D, D]$. A straightforward approach would be to partition the interval $[-D, D]$ into equal subintervals and find the closest eigenvalue to the midpoint of each interval. This approach must work as long as the subintervals are chosen small enough. Nevertheless, partitioning $[-D, D]$ into very fine subintervals is not desirable, since this will require an excessive number of closest eigenvalue computations. Next we present two viable approaches that are both reliable and efficient.

3.3.1. Adaptive progress. The first approach we present here is rather brute-force. In addition to the existence of an upper bound D on the norm of $\mathcal{X}$, we assume that a positive number

$$d \leq \min_{\lambda_i, \lambda_j : \text{distinct eigenvalues}} |\lambda_i - \lambda_j| \tag{3.18}$$

is known [illegible]. We start from the right endpoint D as our initial shift. At each iteration we compute the closest eigenvalue to the current shift and decrement the current shift by an amount depending on the distance from the computed eigenvalue to the shift. We keep decrementing the shift until we reach the left endpoint.

The way the shift ν is updated depends on the closest eigenvalue λ that is found. If λ is real and already discovered, then λ must be larger than ν. In this case there is no real eigenvalue in the interval $(\nu - (\lambda - \nu), \nu]$. Additionally there is no real eigenvalue in the interval $(\lambda - d, \lambda]$. The corresponding update rule in Algorithm 3 combines these two conditions. When λ is real and not discovered, the shift ν is $\lambda - d$. Finally when the closest eigenvalue is not real, the new shift is set to the leftmost of the intersection points of two circles with the real line. One of the circles is centered at ν and has radius $|\lambda - \nu|$. The second circle is centered at λ and has radius d.

In Figure 3.1 the progress of Algorithm 3 on an example is shown. The algorithm iterates 10 times to investigate the part of the real axis where the eigenvalues are known to lie. In particular notice that the algorithm locates the same eigenvalue near the real axis at the second, third, and fourth iterations and another on the real axis at the fifth, sixth, and seventh iterations. Locating the same eigenvalue a few times is a deficiency of this algorithm.

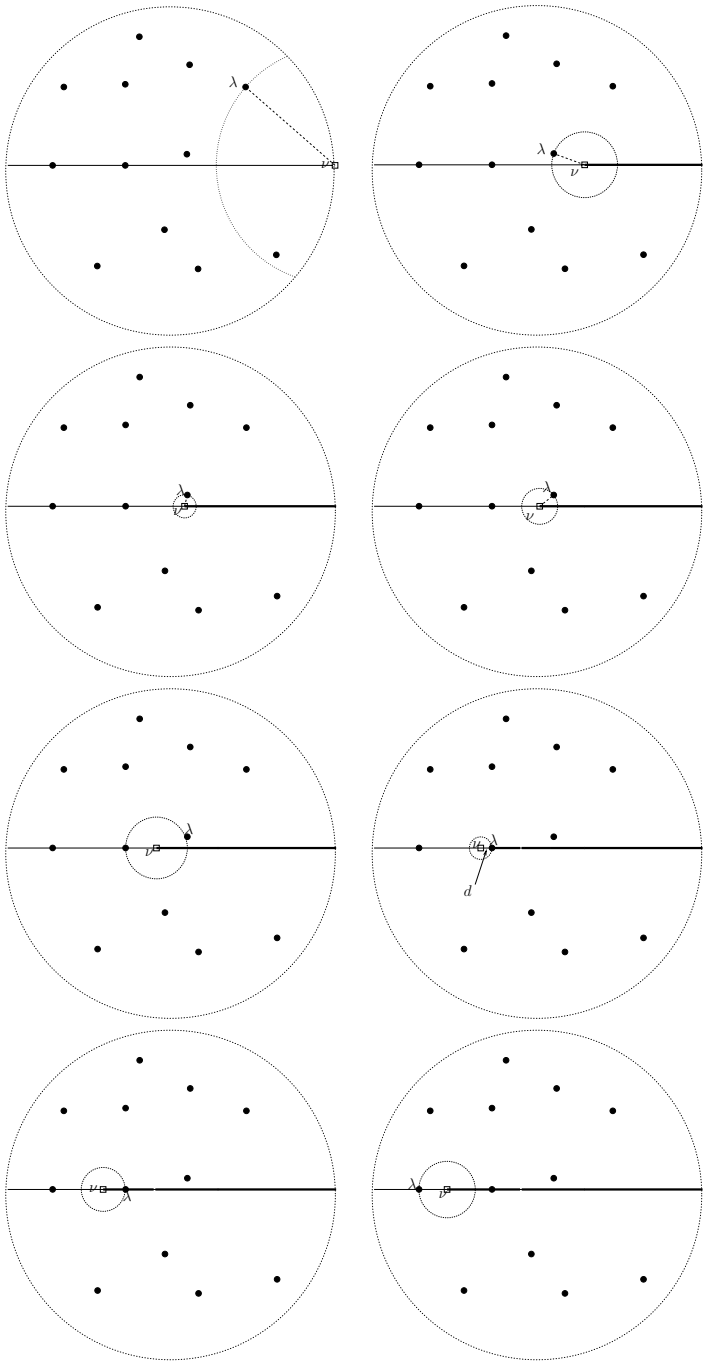


FIG. 3.1. *The first eight iterations (top leftmost is the first iteration; iteration numbers increase from left to right and top to bottom) of the adaptive progress algorithm on an example are displayed. Black dots denote the eigenvalues. Squares mark the location of the shift ν. The closest eigenvalue to ν is denoted by λ. The part of the real axis already investigated is marked by a thicker line. Iterations $2, 3, 4$ locate the same eigenvalue close to the real axis and iterations $5, 6, 7$ locate the same real eigenvalue repeatedly. Little progress is achieved in moving the shift toward the left during these iterations. When an undiscovered real eigenvalue is located, the next shift is obtained by subtracting d, a lower bound on the distance of the closest two eigenvalues, from the real eigenvalue.*

Algorithm 3 Adaptive progress real eigenvalue search algorithm

Call:	$\Lambda \leftarrow$ `Adaptive_Progress`($\mathcal{X}$,D,d).
Input:	$\mathcal{X} \in \mathbb{C}^{q\times q}$, D, an upper bound on the norm of $\mathcal{X}$, and d, a lower bound for the distance between the closest two eigenvalues of $\mathcal{X}$.
Output:	$\Lambda \in \mathbb{R}^l$ with $l \leq q$ containing all of the real eigenvalues of $\mathcal{X}$ in the interval $[-D, D]$.

1. Initially set the shift $\nu \leftarrow D$ and the vector of real eigenvalues $\Lambda \leftarrow [\,]$.
   while $\nu \geq -D$ **do**
     Compute the closest eigenvalue λ to the shift ν.
     if λ is real **then**
       if $\lambda \in \Lambda$ **then**
         % λ is real and already discovered.
         $\nu \leftarrow \nu - \max(|\lambda - \nu|, d - |\lambda - \nu|)$.
       else
         % λ is real but not discovered yet.
         Add λ to Λ.
         $\nu \leftarrow \lambda - d$.
       end if
     else
       % Otherwise λ is not purely real. Choose the leftmost
       % intersection point of the circle centered at ν and with
       % radius $|\nu - \lambda|$ and the circle centered at λ and with
       % radius d with the real line as the new shift.
       if $d \geq \operatorname{Im}\lambda$ **then**
         % Both of the circles intersect the real line.
         $\nu \leftarrow \min(\operatorname{Re}\lambda - \sqrt{d^2 - \operatorname{Im}\lambda^2}, \nu - |\nu - \lambda|)$.
       else
         % Only the circle centered at ν intersects the real line.
         $\nu \leftarrow \nu - |\nu - \lambda|$.
       end if
     end if
   end while
2. Return the real eigenvalue list Λ.

From the description of the algorithm it is not clear whether it terminates. The next theorem shows that the adaptive progress algorithm indeed terminates.

THEOREM 3.1. [illegible] $\nu - \frac{d}{2}$ [illegible] $O(\frac{D}{d})$ [illegible]. Clearly when the closest eigenvalue λ is real, the shift is decremented by at least $d/2$. Therefore let us focus on the case when the eigenvalue λ is not real.

We will find a lower bound for the progress h such that the next shift is $\nu - h$. When only the circle centered at ν intersects the real line (i.e., the imaginary part of λ is greater than d), it is apparent from Figure 3.2c that the distance $|\lambda - \nu|$ is greater than or equal to d. Since the next shift is set to the intersection point $\nu - |\lambda - \nu|$, we have $h = |\lambda - \nu| \geq d$. When both of the circles intersect the real line, as in Figure 3.2a and Figure 3.2b, the progress h is the maximum of the lengths of the line segment

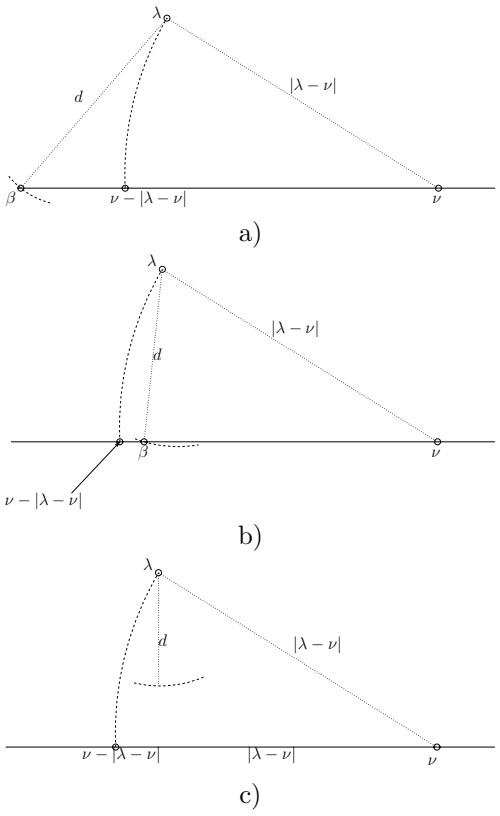


FIG. 3.2. *Three possibilities for the adaptive progress algorithm when the closest eigenvalue λ to the shift ν is not real. The circular arcs are arcs of the circle centered at ν with radius $|\lambda-\nu|$ and the circle centered at λ with radius d. The point β is the intersection point of the second circle with the real line.* a) *Both of the circles intersect the real line, and the circle centered at λ intersects the real line further to the left.* b) *Both of the circles intersect the real line, and the circle centered at ν intersects the real line further to the left.* c) *Only the circle centered at ν intersects the real line.*

from ν to λ and the line segment from ν to β. In both cases, since in the triangle with vertices ν, β and λ, the length of the third edge from β to λ is d, the triangular inequality yields

$$\max(|\nu - \lambda|, |\nu - \beta|) \geq \frac{d}{2}.$$

As the length of the interval to be searched is $2D$, the maximum number of closest eigenvalue computations is bounded by $\frac{4D}{d}$. ∎

We point out a few disadvantages of the adaptive progress approach. The most obvious is the need for a lower bound on the distance between the closest two eigenvalues of $\mathcal{X}$. For reliability, one needs to set d small. The consequence of choosing d small, however, is too many closest eigenvalue computations. A second disadvantage of the adaptive progress approach is that once it detects a new real eigenvalue, it will typically continue to converge to the same eigenvalue with the next few shifts that are close to the eigenvalue. Finally, when an upper bound on $\|\mathcal{X}\|$ is not available, in a robust implementation D must be chosen large, and clearly this degrades the

performance of the algorithm.

3.3.2. Divide and conquer. As an alternative to the adaptive progress approach it is possible to apply a divide and conquer algorithm. Unlike the adaptive progress approach, the divide and conquer algorithm does not require the knowledge of a lower bound on the distance between the closest two eigenvalues of $\mathcal{X}$. In addition it chooses shifts that are away from the computed eigenvalues in order to avoid the discovery of the same eigenvalue too many times (to be precise, each eigenvalue can be discovered at most three times). Even though an upper bound D on the norm of $\mathcal{X}$ is required and may not be available in general, in practice we can choose D very large so that the interval $[-D, D]$ contains all of the real eigenvalues and, as we discuss, this affects the number of closest eigenvalue computations insignificantly. The factor most affecting the efficiency of the algorithm is the eigenvalue distribution.

Algorithm 4 Divide and conquer real eigenvalue search algorithm

Call:	$\Lambda \leftarrow$ `Divide_And_Conquer`$(\mathcal{X},L,U)$.
Input:	$\mathcal{X} \in \mathbb{C}^{q\times q}$, a lower bound L for the smallest real eigenvalue desired and an upper bound U for the largest real eigenvalue desired.
Output:	$\Lambda \in \mathbb{R}^l$ with $l \leq q$ containing all of the real eigenvalues of $\mathcal{X}$ in the interval $[L, U]$.

1. Set the shift $\nu \leftarrow \frac{(U+L)}{2}$.
2. Compute the eigenvalue λ closest to the shift ν.

if $U - L < 2|\lambda - \nu|$ **then**
  % Base case: there is no real eigenvalue in the interval $[L, U]$.
  Return [ ].
else
  % Recursive case: Search the left and right intervals.
  $\Lambda_L \leftarrow$ `Divide_And_Conquer`$(\mathcal{X},L,\nu - |\lambda - \nu|)$
  $\Lambda_R \leftarrow$ `Divide_And_Conquer`$(\mathcal{X},\nu + |\lambda - \nu|,U)$
  % Combine all of the real eigenvalues.
  if λ is real **then**
    Return $\lambda \cup \Lambda_L \cup \Lambda_R$.
  else
    Return $\Lambda_L \cup \Lambda_R$.
  end if
end if

In this approach, given an interval $[L, U]$ we compute the eigenvalue of $\mathcal{X}$ closest to the midpoint of the interval $\nu = \frac{U+L}{2}$. If the modulus of the difference between the computed eigenvalue λ and the midpoint ν is greater than half of the length of the interval, then we terminate. Otherwise we apply the same procedure to the subintervals $[L, \nu - |\lambda - \nu|]$ and $[\nu + |\lambda - \nu|, U]$. Initially we apply the algorithm to the whole interval $[-D, D]$.

Figure 3.3 illustrates the first six iterations of the divide and conquer algorithm on the same example used in Figure 3.1. The divide and conquer algorithm completes the investigation of the real interval where the real eigenvalues reside after iterating 7 times as opposed to the 10 iterations required by the adaptive progress algorithm.

For reliability the parameter D must be chosen large. Suppose all the eigenvalues

are contained in the disk of radius D' with $D' \ll D$. To discover that there is no real eigenvalue in the interval $[D', D]$, at most two extra closest eigenvalue computations are required. If the first shift tried in the interval $[D', D]$ is closer to D' than D, then the distance from the closest eigenvalue to this shift may be less than half the length of the interval $[D', D]$, so a second closest eigenvalue computation may be needed. Otherwise the interval $[D', D]$ will be investigated in one iteration. Similar remarks hold for the interval $[-D, -D']$. However, the larger choices of D may slightly increase or decrease the number of shifts required to investigate $[-D', D']$. The important point is that regardless of how large D is compared to the radius of the smallest disk containing the eigenvalues, the cost is limited to approximately four extra iterations.

Next we show that the number of closest eigenvalue computations cannot exceed $2q + 1$ (recall that $\mathcal{X} \in \mathbb{C}^{q \times q}$).

THEOREM 3.2 (worst case for Algorithm 4). [illegible] 3.2 [illegible] $2q + 1$ [illegible]. We can represent the progress of the algorithm by a full binary tree, i.e., a tree with each node having either two children or no children. Each node of the tree corresponds to an interval. The root of the tree corresponds to the whole interval $[-D, D]$. At each iteration of the algorithm the interval under consideration is either completely investigated or replaced by two disjoint subintervals that need to be investigated. In the first case, the node corresponding to the current interval is a leaf. In the second case, the node has two children, one for each of the subintervals.

We claim that the number of leaves in this tree cannot exceed $q+1$. The intervals corresponding to the leaves are disjoint. Each such interval has a closest left interval (except the leftmost interval) and a closest right interval (except the rightmost interval) represented by two of the leaves in the tree. Each interval is separated from the closest one on the left by the part of a disk on the real axis in which an eigenvalue lies, and similarly for the closest interval on the right. Since the matrix $\mathcal{X}$ has q eigenvalues, there can be at most q separating disks and therefore at most $q + 1$ disjoint intervals represented by the leaves of the tree. A full binary tree with $q+1$ leaves has q internal nodes. Therefore, the total number of the nodes in the tree, which is the same as the number of closest eigenvalue computations, cannot exceed $2q + 1$. ∎

The upper bound $2q + 1$ on the worst case performance of the algorithm is tight, as illustrated by the following example. Consider a matrix with the real eigenvalues $\frac{2^{j-1}-1}{2^{j-1}}$, $j = 1, \ldots, q$, and suppose we search over the interval $[-1, 1]$. Clearly, the algorithm discovers each eigenvalue twice except the largest one, which it discovers three times (assuming that when there are two eigenvalues equally close to a midpoint, the algorithm locates the eigenvalue on the right). Therefore, the total number of closest eigenvalue computations is $2q + 1$.

Next we aim to show that the average case performance of the algorithm is much better than the worst case. First we note the following elementary result that is an immediate consequence of the fact that the square-root function is strictly concave.

LEMMA 3.3. [illegible] l [illegible] $k_1, k_2, \ldots, k_l$ [illegible] l [illegible] $p_1, p_2, \ldots, p_l \in (0, 1)$ [illegible] $\sum_{j=1}^{l} p_j = 1$ [illegible]

$$\sqrt{\sum_{j=1}^{l} p_j k_j} > \sum_{j=1}^{l} p_j \sqrt{k_j} \tag{3.19}$$

[illegible]

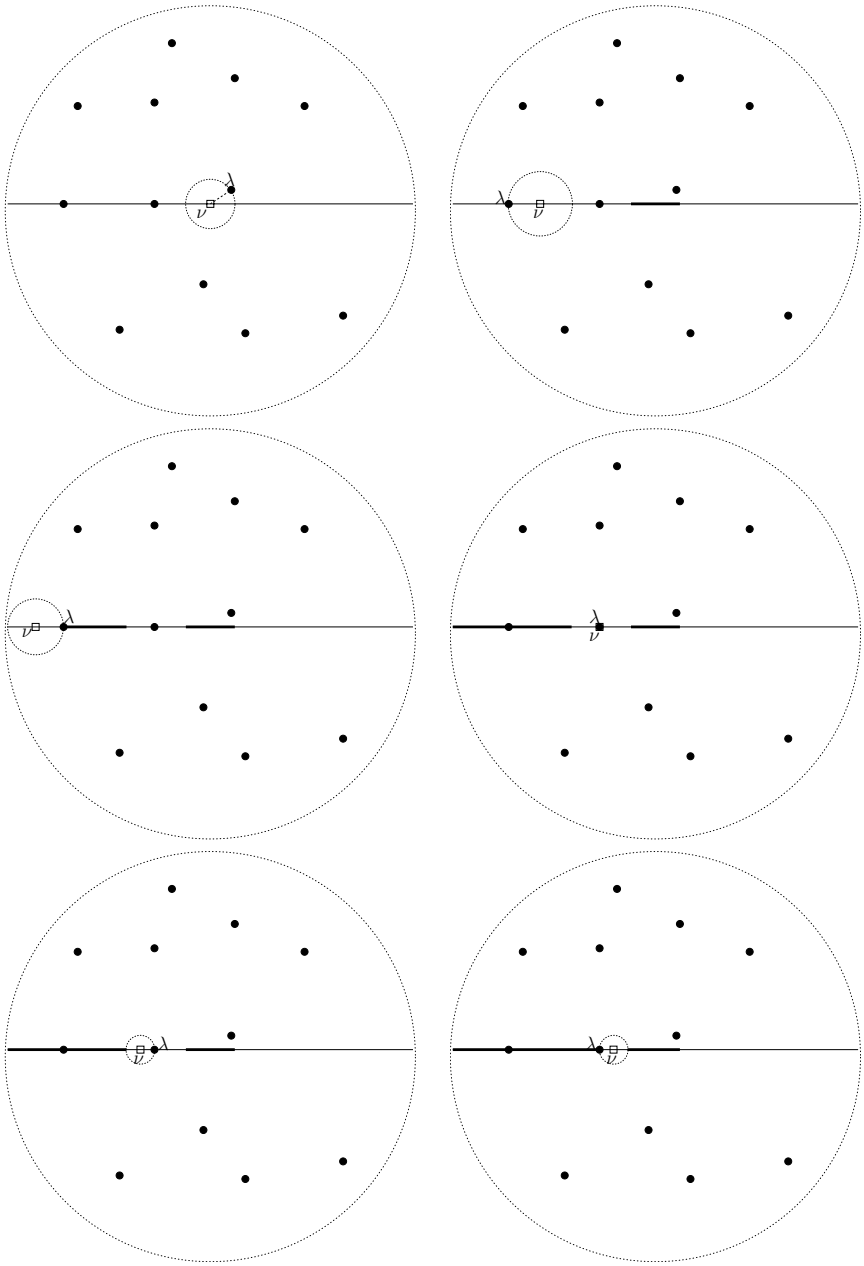


FIG. 3.3. *First six iterations of the divide and conquer algorithm on the same example used in Figure* 3.1.

In the average case analysis we let the eigenvalues of $\mathcal{X}$, say $\xi_1, \xi_2, \ldots, \xi_q$, vary. We assume that the eigenvalues are independently selected from a uniform distribution inside the circle centered at the origin with radius μ. We use Algorithm 4 to compute the real eigenvalues lying inside the circle of radius $D = 1 \leq \mu$ (the value of the radius D is irrelevant for the average case analysis as discussed below; we choose $D = 1$ for simplicity). In Table 3.1 the random variables and the probability density functions

TABLE 3.1
Notation for Theorem 3.4.

X	:	Number of iterations performed by Algorithm 4.
N	:	Number of eigenvalues lying inside the unit circle.
H	:	Modulus of the eigenvalue closest to the origin.
X_l	:	Number of iterations performed by Algorithm 4 on the left interval $[-1, -H]$.
X_r	:	Number of iterations performed by Algorithm 4 on the right interval $[H, 1]$.
N_l	:	Number of eigenvalues lying inside the left circle centered at $-(1+H)/2$ with radius $(1-H)/2$.
$h(H\|N=j)$	:	The probability density function of the variable H given there are j eigenvalues inside the unit circle.
$g_l(N_l\|N=j, H=\beta)$	:	The probability density function of the variable N_l given there are j eigenvalues inside the unit circle and the smallest of the moduli of the eigenvalues is β.

referenced by the proof of the next theorem are summarized.

The quantity we are interested in is $E(X|N=j)$, the expected number of iterations required by Algorithm 4 given that there are j eigenvalues inside the unit circle.

We list a few observations.

- [illegible] $\omega_1, \omega_2, \ldots, \omega_j$ [illegible] $D = 1$ [illegible] This is a simple consequence of the assumption that the eigenvalues are selected from the uniform distribution mutually independently. Let the eigenvalues inside the unit circle be $\xi_{i_1}, \xi_{i_2}, \ldots, \xi_{i_j}$ with $i_1 < i_2 < \cdots < i_j$. We associate ω_k with the location of the kth smallest indexed eigenvalue inside the unit circle, i.e., $\omega_k = \xi_{i_k}$. Let C_1 denote the unit circle. The variable ω_k is uniformly distributed because

$$\begin{aligned} & p(\omega_k | j \text{ of } \xi_1, \ldots, \xi_q \in C_1) \\ &= \sum_{i_1,\ldots,i_j} p(\xi_{i_1}, \ldots, \xi_{i_j} \in C_1 | j \text{ of } \xi_1, \ldots, \xi_q \in C_1) p(\omega_k | \xi_{i_1}, \ldots, \xi_{i_j} \in C_1) \\ &= \sum_{i_1,\ldots,i_j} \binom{q}{j}^{-1} p(\omega_k | \xi_{i_k} \in C_1) \\ &= \sum_{i_1,\ldots,i_j} \binom{q}{j}^{-1} \frac{1}{\pi} = \frac{1}{\pi}. \end{aligned}$$

  Above the summation is over the subsets of $\xi_1, \xi_2, \ldots, \xi_q$ consisting of j elements. Similarly we can show that for $k \neq l$,

$$p(\omega_k, \omega_l | j \text{ of } \xi_1, \ldots, \xi_q \in C_1) = \frac{1}{\pi^2}.$$

  Therefore the variables $\omega_1, \omega_2, \ldots, \omega_j$ are mutually independent.

- [illegible] $\phi_1, \phi_2, \ldots, \phi_{j-1}$ [illegible] H [illegible] Suppose $\omega_{i_1}, \omega_{i_2}, \ldots, \omega_{i_{j-1}}$ with $i_1 < i_2 < \cdots < i_{j-1}$ are the eigenvalues inside the desired area. When we map ω_{i_k} to ϕ_k, the argument above applies to prove the uniformity and mutual independence of the variables $\phi_1, \phi_2, \ldots, \phi_{j-1}$.

- [illegible] c [illegible] $\vartheta_1, \vartheta_2, \ldots, \vartheta_c$ [illegible] $\frac{1-H}{2}$ [illegible] $(\frac{-(1+H)}{2}, 0)$ [illegible] This again follows from the arguments above by mapping ϕ_{i_k} to ϑ_k, where $\phi_{i_1}, \phi_{i_2}, \ldots, \phi_{i_c}$ are the eigenvalues inside the desired region with $i_1 < i_2 < \cdots < i_c$.
- [illegible] D [illegible] D Consider the variables $\hat{\omega}_1$ denoting the locations of the j eigenvalues all inside the circle of radius D_1, and $\hat{\omega}_2 = \frac{D_2\hat{\omega}_1}{D_1}$ denoting the locations of the j eigenvalues inside the circle of radius D_2. Let us denote the number of iterations by Algorithm 4 with input $\hat{\omega}_1$ over the interval $[-D_1, D_1]$ by $X_1(\hat{\omega}_1)$ and the number of iterations with input $\hat{\omega}_2$ over the interval $[-D_2, D_2]$ by $X_2(\hat{\omega}_2)$. It immediately follows that $X_1(\hat{\omega}_1) = X_1(\frac{D_1\hat{\omega}_2}{D_2}) = X_2(\hat{\omega}_2)$. By exploiting this equality we can deduce $E(X_1|N_1 = j) = E(X_2|N_2 = j)$,

$$\begin{aligned} E(X_1|N_1 = j) &= \int_{C_{D_1}} X_1(\hat{\omega}_1)p(\hat{\omega}_1)\, d\hat{\omega}_1 \\ &= \left(\frac{1}{\pi D_1^2}\right)^j \int_{C_{D_1}} X_1(\hat{\omega}_1)\, d\hat{\omega}_1 \\ &= \left(\frac{1}{\pi D_1^2}\right)^j \int_{C_{D_2}} X_1\left(\frac{D_1\hat{\omega}_2}{D_2}\right) \frac{D_1^{2j}}{D_2^{2j}} d\hat{\omega}_2 \\ &= \int_{C_{D_2}} X_2(\hat{\omega}_2)p(\hat{\omega}_2)\, d\hat{\omega}_2 \\ &= E(X_2|N_2 = j), \end{aligned}$$

where N_1 and N_2 are the number of eigenvalues inside the circle of radius C_{D_1} of radius D_1 and the circle of radius C_{D_2} of radius D_2, respectively. Note that the eigenvalues inside both the circle C_{D_1} and the circle C_{D_2} are uniformly distributed and independent, as we discussed above.

By combining these remarks we conclude the equality $E(X|N = j) = E(X_l|N_l = j)$, since the eigenvalues are uniformly distributed and independent inside the circles, and the sizes of the circles do not affect the expected number of iterations given that there are j eigenvalues inside the circles.

The next theorem establishes a recurrence equation for $E(X|N = j)$ in terms of $E(X|N = k)$, $k = 0 \ldots j-1$. Using the recurrence equation we will show $E(X|N = j) = O(\sqrt{j})$ by induction. For convenience let us use the shorthand notation $E_j(X)$ for $E(X|N = j)$.

THEOREM 3.4. [illegible] q [illegible] μ [illegible] 4 [illegible] $[-1, 1]$ [illegible] $E_j(X)$ [illegible]

$$E_0(X) = 1 \tag{3.20}$$

[illegible] $0 < j < q$

$$E_j(X) = 2F_{j-1}(X) + 1, \tag{3.21}$$

[illegible] $F_{j-1}(X)$ [illegible] $E_0(X), \ldots, E_{j-1}(X)$

$$F_{j-1}(X) = \int_0^1 \left(\sum_{k=0}^{j-1} E_k(X) g_l(N_l = k | N = j, H = \beta) \right) h(H = \beta | N = j)\, d\beta. \tag{3.22}$$

[illegible]. Equation (3.20) is trivial; when there is no eigenvalue inside the unit circle, the algorithm will converge to an eigenvalue on or outside the unit circle and terminate.

For $j > 0$ at the first iteration of the algorithm, we compute the closest eigenvalue to the midpoint and repeat the same procedure with the left interval and with the right interval, so the equality

$$X = X_l + X_r + 1$$

and therefore the equality

$$E_j(X) = E(X_l | N = j) + E(X_r | N = j) + 1 \tag{3.23}$$

follow. Clearly the number of iterations on the left and right intervals depend on the modulus of the computed eigenvalue. By the definition of conditional expectations, we deduce

$$E(X_l | N = j) = \int_0^1 E(X_l | N = j, H = \beta) h(H = \beta | N = j)\, d\beta \tag{3.24}$$

and similarly

$$E(X_r | N = j) = \int_0^1 E(X_r | N = j, H = \beta) h(H = \beta | N = j)\, d\beta. \tag{3.25}$$

Now we focus on the procedures applied on the left and right intervals. Let the modulus of the eigenvalue computed at the first iteration be β. There may be up to $j - 1$ eigenvalues inside the circle centered at the midpoint of the left interval $[-1, -\beta]$ and with radius $\frac{1-\beta}{2}$. The expected number of iterations on the left interval is independent of the radius $\frac{1-\beta}{2}$ and the number of eigenvalues lying outside this circle. Therefore given the number of eigenvalues inside this circle, by the definition of conditional expectations, the equality

(3.26)

$$\begin{aligned} E(X_l | N = j, H = \beta) &= \sum_{k=0}^{j-1} E(X_l | N_l = k, N = j, H = \beta) g_l(N_l = k | N = j, H = \beta) \\ &= \sum_{k=0}^{j-1} E(X_l | N_l = k) g_l(N_l = k | N = j, H = \beta) \\ &= \sum_{k=0}^{j-1} E_k(X) g_l(N_l = k | N = j, H = \beta) \end{aligned}$$

is satisfied. A similar argument applies to the right interval to show the analogous equality

$$E(X_r | N = j, H = \beta) = \sum_{k=0}^{j-1} E_k(X) g_l(N_l = k | N = j, H = \beta). \tag{3.27}$$

By substituting (3.26) into (3.24), (3.27) into (3.25), and combining these with (3.23), we deduce the result. ∎

COROLLARY 3.5 (average case for Algorithm 4). [illegible] $E_j(X)$ [illegible] $c\sqrt{j+f}-1$ [illegible] $c \geq \sqrt{12}$ [illegible] $f \in [4/c^2, 1/3]$ [illegible]. The proof is by induction. In the base case, when there is no eigenvalue inside the unit circle, the algorithm iterates only once, i.e., $E_0(X) = 1 \leq c\sqrt{f}-1$.

Assume for all $k < j$, that the claim $E_k(X) \leq c\sqrt{k+f}-1$ holds. We need to show the inequality $E_j(X) \leq c\sqrt{j+f}-1$ is satisfied under this assumption. By definition (3.22) in Theorem 3.4 we have

(3.28)
$$F_{j-1}(X) \leq \int_0^1 \left(\sum_{k=0}^{j-1} (c\sqrt{k+f}-1) g_l(N_l = k | N = j, H = \beta) \right) h(H = \beta | N = j)\, d\beta.$$

As we argued before, the uniformity and independence of each of the $j-1$ eigenvalues inside the unit circle but outside the circle of radius $H = \beta$ is preserved. In other words $g_l(N_l|N = j, H = \beta)$ is a binomial density function, and we can explicitly write $g_l(N_l = k|N = j, H = \beta)$, the probability that there are k eigenvalues inside the left circle given that there are $j-1$ eigenvalues contained in the unit circle and outside the circle of radius β, as

$$g_l(N_l = k|N = j, H = \beta) = \binom{j-1}{k} \left(\frac{1-\beta}{4(1+\beta)} \right)^k \left(1 - \frac{1-\beta}{4(1+\beta)} \right)^{j-1-k}.$$

Now the expected value of the binomial distribution above is $(j-1)\frac{1-\beta}{4(1+\beta)}$. From Lemma 3.3, we deduce

$$\begin{aligned} \frac{\sqrt{j+f}}{2} &\geq \frac{\sqrt{j-1+4f}}{2} \\ &\geq \sqrt{\frac{(1-\beta)(j-1)}{4(1+\beta)} + f} \\ &= \sqrt{\sum_{k=0}^{j-1} (k+f) g_l(N_l = k|N = j, H = \beta)} \\ &> \sum_{k=0}^{j-1} \sqrt{k+f}\, g_l(N_l = k|N = j, H = \beta). \end{aligned}$$

Substituting the upper bound $\frac{\sqrt{j+f}}{2}$ for $\sum_{k=0}^{j-1} \sqrt{k+f}\, g_l(N_l = k|N = j, H = \beta)$ in (3.28) yields

$$F_{j-1}(X) \leq \int_0^1 \left(\frac{c\sqrt{j+f}}{2} - 1 \right) h(H = \beta | N = j)\, d\beta = \frac{c\sqrt{j+f}}{2} - 1. \tag{3.29}$$

Now it follows from (3.21) that

$$E_j(X) \leq c\sqrt{j+f}-1 \tag{3.30}$$

as desired. ∎

Recall that we intend to apply the divide and conquer approach to $\mathcal{A}$ which has size $2n^2 \times 2n^2$. Assume that the conditions of Corollary 3.5 hold for the eigenvalues of $\mathcal{A}$ and the circle of radius D contains all of the eigenvalues. Suppose also that for any shift ν, convergence of the shifted inverse iteration or shift-and-invert Arnoldi method to the closest eigenvalue requires the matrix vector multiplication $(\mathcal{A} - \nu I)^{-1}u$ for various u only a constant number of times. Then the average running time of each trisection step is $O(n^4)$, since finding the closest eigenvalue takes $O(n^3)$ time (which is the cost of solving a Sylvester equation of size $2n$ a constant number of times) and we compute the closest eigenvalue $O(n)$ times at each trisection step on average. Because of the special structure of the Kronecker product matrix $\mathcal{A}$, even if the input matrices have eigenvalues uniformly distributed and mutually independent, the eigenvalues of $\mathcal{A}$ may not have this property. However, the numerical examples in the next section suggest that the number of closest eigenvalue computations as a function of the size of the Kronecker product matrices is still bounded by $O(\sqrt{q})$. According to Theorem 3.2, in the worst case scenario, each trisection step requires $O(n^5)$ operations, which is an improvement over computing all of the eigenvalues of $\mathcal{A}$.

4. Numerical experiments. We first compare the accuracy of the new algorithm with the divide and conquer approach, and Gu's algorithm in [6] on a variety of examples. Second, we discuss why in general we prefer the divide and conquer approach over the adaptive progress approach. In our final example we aim to show the asymptotic running time difference between the new method and Gu's method. All the tests are run using MATLAB 6.5 under Linux on a PC.

4.1. Accuracy of the new algorithm and the old algorithm. We present results comparing the accuracy of the new method using the divide and conquer approach with Gu's method in [6]. In exact arithmetic both the method in [6] and the new method using the divide and conquer approach must return the same interval, since they perform the same verification by means of different but equivalent eigenvalue problems. Our data set consists of pairs (A, B), where A is provided by the software package `EigTool` [13] and B has entries selected independently from the normal distribution with zero mean and variance one. The data set is available on the web.[3] In all of the tests the initial interval is set $[0, \sigma_n([A \quad B])]$ and the trisection step is repeated until an interval $(l, u]$ with $u - l \leq 10^{-4}$ is obtained.

When the second and third columns in Table 4.1 are considered, on most of the examples the methods return the same interval with the exception of the companion, Demmel, Godunov, and gallery5 examples. The common property of these matrices is that they have extremely ill-conditioned eigenvalues. As we discuss in section 5, when the matrix A has an ill-conditioned eigenvalue, the new method is not expected to produce accurate small intervals containing the distance to uncontrollability. One false conclusion that one may draw from Table 4.1 is that Gu's method is always more accurate than the new method. Indeed for the Basor–Morrison, Grcar, or Landau examples with $n = 5$ (for which the eigenvalues are fairly well conditioned) the new method generates more accurate results than Gu's method when one seeks intervals of length around 10^{-6}. In terms of accuracy these two methods have different weaknesses.

[3] http://www.cs.nyu.edu/∼mengi/robust_stability/data_dist_uncont.mat.

TABLE 4.1

For pairs (A, B) with A chosen from `EigTool` *as listed above in the leftmost column and B normally distributed, intervals $(l, u]$ that are supposed to contain the distance to uncontrollability of the system (A, B) are computed with $u - l \leq 10^{-4}$ by both of the methods. The size of the system (n, m) is provided next to the name of the matrix A in the leftmost column. In the fourth column the norm of $\mathcal{A}$ computed by* MATLAB *at the last trisection step is given. In the rightmost column the norm of $\mathcal{A}$ is approximated using* (5.1).

Example	New method	Gu's method	$\|\mathcal{A}\|$	$\approx \|\mathcal{A}\|$
Airy(5,2)	(0.03759,0.03767]	(0.03759,0.03767]	8×10^7	7×10^8
Airy(10,4)	(0.16337,0.16345]	(0.16337,0.16345]	4×10^7	6×10^8
Basor–Morrison(5,2)	(0.68923,0.68929]	(0.68923,0.68929]	2×10^6	2×10^7
Basor–Morrison(10,4)	(0.60974,0.60980]	(0.60974,0.60980]	2×10^7	5×10^8
Chebyshev(5,2)	(0.75026,0.75034]	(0.75026,0.75034]	3×10^7	5×10^8
Chebyshev(10,4)	(0.82703,0.82711]	(0.82703,0.82711]	6×10^{10}	3×10^{12}
Companion(5,2)	(0.42431,0.42438]	(0.42431,0.42438]	2×10^8	4×10^9
Companion(10,4)	(0.46630,0.46637]	(0.46610,0.46616]	5×10^{13}	5×10^{18}
Convection diffusion(5,2)	(0.69829,0.69836]	(0.69829,0.69836]	7×10^5	1×10^7
Convection diffusion(10,4)	(1.48577,1.48586]	(1.48577,1.48586]	9×10^6	1×10^8
Davies(5,2)	(0.23170,0.23176]	(0.23170,0.23176]	2×10^6	2×10^7
Davies(10,4)	(0.70003,0.70012]	(0.70003,0.70012]	1×10^6	1×10^7
Demmel(5,2)	(0.09090,0.09097]	(0.09049,0.09056]	8×10^{49}	`Inf`
Demmel(10,4)	(0.12049,0.12057]	(0.11998,0.12006]	9×10^{80}	`Inf`
Frank(5,2)	(0.45907,0.45916]	(0.45907,0.45916]	1×10^7	7×10^8
Frank(10,4)	(0.67405,0.67414]	(0.67405,0.67414]	3×10^{16}	2×10^{18}
Gallery5(5,2)	(0.17468,0.17474]	(0.02585,0.02592]	1×10^{16}	1×10^{29}
Gauss–Seidel(5,2)	(0.06279,0.06288]	(0.06279,0.06288]	2×10^{20}	`Inf`
Gauss–Seidel(10,4)	(0.05060,0.05067]	(0.05060,0.05067]	1×10^{30}	3×10^{40}
Godunov(7,3)	(1.23802,1.23810]	(1.23764,1.23773]	1×10^{14}	8×10^{31}
Grcar(5,2)	(0.49571,0.49579]	(0.49571,0.49579]	2×10^5	3×10^6
Grcar(10,4)	(0.44178,0.44185]	(0.44178,0.44185]	4×10^7	6×10^8
Hatano(5,2)	(0.39570,0.39578]	(0.39570,0.39578]	4×10^6	2×10^7
Hatano(10,4)	(0.23297,0.23304]	(0.23297,0.23304]	4×10^8	1×10^{10}
Kahan(5,2)	(0.18594,0.18601]	(0.18594,0.18601]	8×10^8	2×10^{10}
Kahan(10,4)	(0.05587,0.05594]	(0.05587,0.05594]	8×10^{13}	7×10^{14}
Landau(5,2)	(0.41766,0.41773]	(0.41766,0.41773]	1×10^5	1×10^6
Landau(10,4)	(0.28166,0.28174]	(0.28166,0.28174]	1×10^7	3×10^8
Markov chain(6,2)	(0.04348,0.04358]	(0.04348,0.04358]	3×10^7	5×10^8
Markov chain(10,4)	(0.07684,0.07693]	(0.07684,0.07693]	8×10^8	6×10^{10}
Orr–Sommerfield(5,2)	(0.04789,0.04796]	(0.04789,0.04796]	1×10^9	8×10^9
Orr–Sommerfield(10,4)	(0.07836,0.07843]	(0.07836,0.07843]	2×10^{10}	2×10^{12}
Skew–Laplacian(8,3)	(0.01001,0.01011]	(0.01001,0.01011]	3×10^{10}	4×10^{13}
Supg(4,2)	(0.06546,0.06554]	(0.06546,0.06554]	7×10^8	4×10^9
Supg(9,4)	(0.03627,0.03634]	(0.03627,0.03634]	2×10^{13}	1×10^{14}
Transient(5,2)	(0.11027,0.11036]	(0.11027,0.11036]	3×10^7	4×10^8
Transient(10,4)	(0.13724,0.13731]	(0.13724,0.13731]	6×10^8	6×10^9
Twisted(5,2)	(0.14929,0.14936]	(0.14929,0.14936]	2×10^7	1×10^8
Twisted(10,4)	(0.77178,0.77185]	(0.77178,0.77185]	1×10^7	2×10^8

TABLE 4.2
A comparison of the real eigenvalue extraction techniques when the matrix A has eigenvalues squeezed in a small real interval.

Method	d	Computed interval	No. of calls to eigs
Adaptive progress	10^{-1}	(0.00702,0.00711]	19
Adaptive progress	10^{-2}	(0.00474,0.00483]	35
Adaptive progress	10^{-3}	(0.00476,0.00484]	81
Divide and conquer	-	(0.00476,0.00484]	13
Gu's algorithm	-	(0.00476,0.00484]	-

TABLE 4.3
A comparison of the real eigenvalue extraction techniques for an uncontrollable pair.

Method	d	Computed interval	No. of calls to eigs
Adaptive progress	1	(0.13538,0.13546]	17
Adaptive progress	0.5	(0.00025,0.00033]	19
Adaptive progress	0.1	(0.00000,0.00008]	50
Divide and conquer	-	(0.00000,0.00008]	12
Gu's algorithm	-	(0.00000,0.00008]	-

4.2. Comparison of the real eigenvalue extraction techniques. When the Kronecker product matrix $\mathcal{A}$ has too many eigenvalues close to the real axis, the adaptive progress method is not the ideal real eigenvalue extraction technique. A remedy for the efficiency problems in this case is choosing d large, which may cause accuracy problems. Suppose we choose $A = Q\,\mathrm{diag}(v)\,Q^*$, where

$$v = [-1 \;\; -0.99 \;\; -0.98 \;\; -0.97 \;\; -0.96]$$

and Q is a unitary matrix whose columns form an orthonormal basis for the column space of a normally distributed matrix. The matrix B is chosen from a normal distribution. The eigenvalue pattern of A is also reflected in $\mathcal{A}$, as it also has eigenvalues tightly squeezed around -1. Among the values listed in Table 4.2, $d = 10^{-3}$ is the one for which the adaptive progress approach returns the correct interval. But the average number of calls to `eigs` for $d = 10^{-3}$ by the adaptive progress approach is more than six times the number of calls made by the divide and conquer approach.

In a second example we choose the pair

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0.95 & 1 \\ 0 & 0 & 0.9 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 0.1 \\ 0 \end{pmatrix},$$

which is uncontrollable since $\mathrm{rank}([A - 0.9I \;\; B]) = 2$. In Table 4.3 we see that the adaptive progress approach with $d = 0.1$ yields the correct interval. Nevertheless the number of calls to `eigs` is once again excessive compared to the number of calls by the divide and conquer approach.

These examples illustrate that there may not exist a d value such that the adaptive progress returns an accurate result with fewer calls to `eigs` than the number of calls required by the divide and conquer approach.

4.3. Running times of the new algorithm on large matrices. To observe the running time differences between Gu's method and the new method with the divide and conquer approach, we run the algorithms on pairs (A, B) of various size, where A is a Kahan matrix available through `EigTool` and B is a normally distributed matrix.

TABLE 4.4

Running times of Gu's method/new method in seconds and the average number of calls to eigs by the new method for Kahan-random matrix pairs of various size.

Size (n,m)	t_{cpu} (Gu's method)	t_{cpu} (new method)	No. of calls to eigs
(10,6)	47	171	34
(20,12)	3207	881	63
(30,18)	46875	3003	78
(40,24)	263370	7891	92

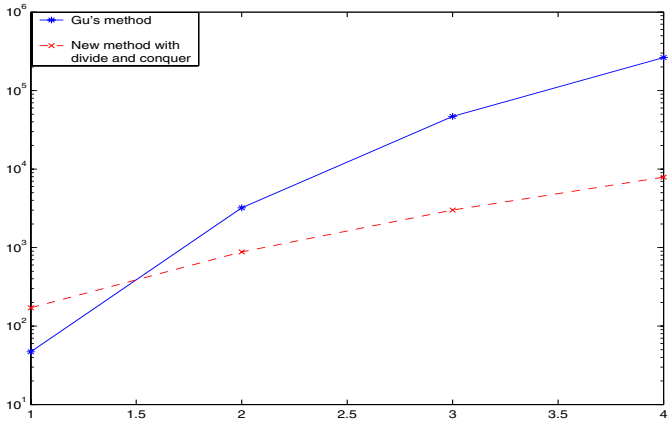


FIG. 4.1. *Running times of the methods on Kahan-random matrix pairs are displayed as functions of the size of the matrix in logarithmic scale.*

We normalized the pairs (by dividing them by $\sigma_n([A\ B])$) so that the same number of trisection steps are required. For $(n, m) = (40, 24)$ we didn't run Gu's method, since it takes an excessive amount of time. Instead we extrapolated its running time. For all other sizes both methods return the same interval of length approximately 10^{-4}. In Table 4.4 the running times of both the algorithms and the average number of calls to eigs made by the divide and conquer approach are provided for various sizes. For small pairs Gu's method is faster. However, for matrices of size 20 and larger the new method is more efficient and the difference in the running times increases drastically as a function of n. In the third column the average number of calls to `eigs` is shown and apparently varies linearly with n. Figure 4.1 displays plots of the running times as functions of n using a log scale. The asymptotic difference in the running times agrees with the plots.

5. Concluding remarks. Based on the results in the previous section, among all the methods discussed, the most reliable and efficient appears to be the new verification scheme with the divide and conquer approach to extract the real eigenvalues. The divide and conquer approach requires only an upper bound on the norm of $\mathcal{A}$. In practice this parameter may be set arbitrarily large and the efficiency of the algorithm is affected insignificantly. Alternatively the upper bound on $\|\mathcal{A}\|$ in [11], or when A has simple eigenvalues, the formula (5.1) discussed below can be used.

Improvements to the divide and conquer approach still seem possible. As the upper and lower bound become closer, the Kronecker product matrices $\mathcal{A}$ in two successive iterations differ only slightly. Therefore it is desirable to benefit from the eigenvalues computed in the previous iteration in the selection of the shifts. We

address further details of the new algorithm below.

5.1. Sylvester equation solvers. The Sylvester equations needed to perform the multiplication $(\mathcal{A}-\nu I)^{-1}u$ are not sparse in general. We solve them by first reducing the coefficient matrices on the left-hand side of (3.17) to upper quasi-triangular forms (block upper triangular matrices with 1×1 and 2×2 blocks on the diagonal). Then the algorithm of Bartels and Stewart can be applied [2]. In our implementation we used the LAPACK routine `dtrsyl` [1], which is similar to the method of Bartels and Stewart, but rather than computing the solution column by column it generates the solution row by row, bottom to top. A more efficient alternative may be the recursive algorithm of Jonsson and Kågström [7].

5.2. Difficulties in computing to a high precision. Gu's method in [6] suffers from the fact that the matrix Q_{12} in (2.4) becomes highly ill-conditioned as $\delta \to 0$ and is not invertible at the limit. This is an issue if the input pair is uncontrollable or nearly uncontrollable.

For the new method instability is caused by small η. The accuracy of the algorithm depends on the ability to extract the real eigenvalues of $\mathcal{A}$ successfully. (The imaginary eigenvalues of $H(\alpha)$ can be obtained reliably by using a Hamiltonian eigenvalue solver.) A computed eigenvalue of $\mathcal{A}$ differs from the exact one by a quantity with modulus on the order of $\|\mathcal{A}\|\epsilon_{mach}/|w^*z|$, where w and z are the corresponding unit left and right eigenvectors, respectively. In general the more dominant factor in the formation of this numerical error is the norm $\|\mathcal{A}\|$ rather than the absolute condition number of the eigenvalue (appearing in the denominator), since the inverted matrix in the definition of $\mathcal{A}$ is the inverse of a matrix that is nearly singular for small η, and therefore the norm of $\mathcal{A}$ is big. There is another numerical trouble caused by big $\|\mathcal{A}\|$. We cannot expect to solve the linear system $(\mathcal{A}-\nu I)x = u$ accurately for $\mathcal{A}$ with large norm. This obviously has an effect on the convergence of shifted inverse iteration and shift-and-invert preconditioned Arnoldi especially considering the fact that the shift ν is not close to an eigenvalue in general. (Because of this, computing the eigenvalues of $\mathcal{A}$ using the QR algorithm may be superior to computing them using shifted inverse iteration or shift-and-invert preconditioned Arnoldi, as indeed we observed in practice.) In our experience `eigs` has convergence problems typically when the norm of $\mathcal{A}$ reaches the order of 10^{10}. Smaller η contributes to the increase in the norm of $\mathcal{A}$; however, it is not the only factor. Indeed for certain pairs (A, B) the norm $\|\mathcal{A}\|$ is large even when η is not small. Under the assumption that A is diagonalizable, an upper bound on $\|\mathcal{A}\|$ is derived in [11]. Specifically when A has simple eigenvalues, the upper bound on $\|\mathcal{A}\|$ in [11] simplifies to

$$2\|A\| + \frac{(2\|BB^*/\delta - \delta I\| + \delta)^2}{\eta \inf_{\det(A-\lambda I)=0} |y_\lambda^* x_\lambda|^2} \tag{5.1}$$

with x_λ and y_λ denoting the unit right and unit left eigenvectors, respectively, corresponding to λ. Notice that the upper bound given by (5.1) can be efficiently computed in $O(n^3)$ time, and therefore in an implementation it can be used to estimate the length of the smallest interval containing the distance to uncontrollability that can possibly be computed accurately. Surprisingly the norm of $\mathcal{A}$ heavily depends on the worst conditioned eigenvalue of A, but it has little to do with the norm of A. For instance, when A is normal and $\|B\|$ is not very large, we expect that $\|\mathcal{A}\|$ exceeds 10^{10} only when η is smaller than 10^{-10} unless the pair (A, B) is nearly uncontrollable. This in turn means we can reliably compute an interval of length 10^{-10} containing

the distance to uncontrollability. On the other hand when A is far from being normal or the pair (A, B) is close to being uncontrollable and a small interval is required, the new method performs poorly. The accuracy of the intervals generated on various examples in Table 4.1 in the second column is also justified by (5.1). All of the examples for which the method performs poorly are highly nonnormal. In the fourth column in Table 4.1 the norms of $\mathcal{A}$ computed by calling MATLAB's `norm` at the last trisection step (approximately when η is the difference between the upper and lower bounds of the interval in the second column and δ is the upper bound of the interval) are listed. In the rightmost column the upper bounds on the norm of $\mathcal{A}$ using (5.1) are provided. For most of the pairs in Table 4.1 the upper bound on $\|\mathcal{A}\|$ in the rightmost column is tight.

5.3. Alternative eigenvalue problem. To see whether there exists an α such that $H(\alpha)$ and $H(\alpha + \eta)$ share an eigenvalue, we extract the real eigenvalues of $\mathcal{A}$. Alternatively we can solve the generalized eigenvalue problem

(5.2)
$$P - \lambda M = \begin{pmatrix} -A_1^* - A_2^T & \delta I & \delta I & 0 \\ B_2^T & -A_1^* + \bar{A}_2 & 0 & \delta I \\ B_1 & 0 & A_1 - A_2^T & \delta I \\ 0 & B_1 & B_2^T & A_1 + \bar{A}_2 \end{pmatrix} - \lambda \begin{pmatrix} -I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I \end{pmatrix}.$$

The real eigenvalue extraction techniques are applicable to this problem as well, since the scalar λ is an eigenvalue of the pencil above if and only if $\frac{1}{\lambda - \nu}$ is an eigenvalue of the matrix $(P - \nu M)^{-1} M$. The multiplication $x = (P - \nu M)^{-1} M y$ can be performed efficiently by solving the linear system $(P - \nu M)x = My$. When we write this linear system in matrix form, we obtain the Sylvester equation (3.3) but with α replaced by ν and with the matrix

$$\begin{pmatrix} -Y_{11} & 0 \\ 0 & Y_{22} \end{pmatrix}$$

replacing 0 on the right-hand side, where $y = [\mathbf{vec}(Y_{11})\ y_{12}\ y_{21}\ \mathbf{vec}(Y_{22})]^T$ with equal sized block components. Notice that the fact that the eigenvalue problem (5.2) is of double size compared to the eigenvalue problem $\mathcal{A}x = \lambda x$ is not an efficiency concern. We still solve Sylvester equations of the same size. The real issue is that these two eigenvalue problems have different conditioning. Theoretically either of them can be better conditioned than the other in certain situations. In practice we retrieved more accurate results with the eigenvalue problem $\mathcal{A}x = \lambda x$ most of the time, even though there are also examples on which the algorithm using (5.2) yields more accurate results.

6. Software. By combining the new verification scheme and BFGS, it is possible to come up with a more efficient and accurate algorithm. A local minimum of the function $\sigma_n([A - \lambda I \quad B])$ can be found in a cheap manner by means of the BFGS optimization algorithm. Notice that the cost of this local optimization step is $O(1)$, since we are searching over two unknowns, namely, the real and the imaginary parts of λ. Using the new verification scheme we can check whether the local minimum is indeed a global minimum as described in [3, Algorithm 5.3]. If the local minimum is not a global minimum, the new verification scheme also provides us with a point λ', where the value of the function $\sigma_n([A - \lambda I \quad B])$ is less than the local minimum.

Therefore we can repeat the application of BFGS followed by the new scheme until we verify that the local minimum is a global minimum.

An efficient implementation of the new method is freely available.[4] In this implementation, by setting an input parameter appropriately, one can either run the trisection method or the hybrid method just described. Typically, the new scheme is faster than the previous implementation of the trisection method of [3] for matrices of size larger than 20.

Acknowledgements. Many thanks to A. Yılmaz for carefully reading the average case analysis for the divide and conquer approach. We are also grateful to two anonymous referees for their invaluable suggestions.

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. Sorenson, *LAPACK Users' Guide*, 3rd ed., SIAM, Philadelphia, 1999.

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

[3] 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.

[4] R. Eising, *The distance between a system and the set of uncontrollable systems*, in Mathematical theory of networks and systems, Lecture Notes in Control and Inform. Sci., 58, Springer, London, 1984, pp. 303–314.

[5] R. Eising, *Between controllable and uncontrollable*, System Control Lett., 4 (1984), pp. 263–264.

[6] M. Gu, *New methods for estimating the distance to uncontrollability*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 989–1003.

[7] I. Jonsson and B. Kågström, *Recursive blocked algorithm for solving triangular systems, Part* I: *One-sided and coupled Sylvester-type matrix equations*, ACM Trans. Math. Software, 28 (2002), pp. 392–415.

[8] R. E. Kalman, *Mathematical description of linear dynamical systems*, J. SIAM Control Ser. A, 1 (1963), pp. 152–192.

[9] R. B. Lehoucq, K. Maschoff, D. Sorensen, and C. Yang, *ARPACK Software Package*, http://www.caam.rice.edu/software/ARPACK/, 1996.

[10] R. B. Lehoucq, D. Sorensen, and C. Yang, *ARPACK Users' Guide*, SIAM, Philadelphia, 1998.

[11] E. Mengi, *Measures for Robust Stability and Controllability*, Ph.D. thesis, Courant Institute of Mathematical Sciences, New York, NY, 2006.

[12] C. C. Paige, *Properties of numerical algorithms relating to computing controllability*, IEEE Trans. Automat. Control, 26 (1981), pp. 130–138.

[13] T. G. Wright, *EigTool: A graphical tool for nonsymmetric eigenproblems*, Oxford University Computing Laboratory, Oxford, UK, http://www.comlab.ox.ac.uk/pseudospectra/eigtool/, 2002.

[4] http://www.cs.nyu.edu/∼mengi/robust_stability/dist_uncont.html.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 2, pp. 503–523

ON THE NEWTON METHOD FOR THE MATRIX *P*TH ROOT*

BRUNO IANNAZZO†

Abstract. Stable versions of Newton's iteration for computing the principal matrix pth root $A^{1/p}$ of an $n \times n$ matrix A are provided. In the case in which X_0 is the identity matrix, it is proved that the method converges for any matrix A having eigenvalues with modulus less than 1 and with positive real parts. Based on these results we provide a general algorithm for computing the principal pth root of any matrix A having no nonpositive real eigenvalues. The algorithm has quadratic convergence, is stable in a neighborhood of the solution, and has a cost of $O(n^3 \log p)$ operations per step. Numerical experiments and comparisons are performed.

Key words. matrix pth root, Newton's method, numerical stability, matrix iterations, rational iterations, Julia set

AMS subject classifications. 15A24, 65F30

DOI. 10.1137/050624790

1. Introduction.

A useful tool for solving nonlinear equations is the Newton method,

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

for an initial value x_0. For the algebraic equation $x^p - a = 0$, $a \in \mathbb{C}$, it turns into

$$x_{k+1} = \frac{(p-1)x_k + a x_k^{1-p}}{p}. \tag{1.1}$$

As pointed out by Cayley in 1879 [4], the study of the convergence of this iteration is very hard for $p > 2$. In fact the set of initial values such that the iteration (1.1) converges to a specific root is a [illegible] but [illegible] set, and its boundary is the so-called Julia set of the iteration.

For $A \in \mathbb{C}^{n\times n}$, one can consider the matrix iteration

$$X_{k+1} = \frac{(p-1)X_k + A X_k^{1-p}}{p} \tag{1.2}$$

for solving the matrix equation

$$X^p - A = 0. \tag{1.3}$$

One of the most interesting solutions of (1.3) is the [illegible] p[illegible] $A^{1/p}$ of A whose eigenvalues lie in the sector

$$\mathcal{S}_p = \{z \in \mathbb{C}\backslash\{0\}\,,\ -\pi/p < \arg z < \pi/p\}. \tag{1.4}$$

*Received by the editors February 19, 2005; accepted for publication (in revised form) by A. Frommer January 26, 2006; published electronically June 21, 2006. This work was supported by MIUR grant 2002014121.

http://www.siam.org/journals/simax/28-2/62479.html

†Dipartimento di Matematica "L.Tonelli," Università di Pisa, Largo B. Pontecorvo 5, 56127 Pisa, Italy (iannazzo@mail.dm.unipi.it).

If A has no nonpositive real eigenvalues, then there exists a unique principal pth root. Here and hereafter we refer to (1.2) as the [illegible].

The main applications of the matrix pth root are for the computation of the logarithm of a matrix and the sector function; for other applications, see [7, 11].

Convergence and stability properties of (1.2) are important issues which play a fundamental role in the design of an algorithm for the matrix pth root. Hoskins and Walton [12] and Smith [18] take as initial value the matrix A. Unfortunately, as discussed in [18], this choice leads to a convergence region not nice enough to design a simple global convergent method.

Concerning stability, Higham [8] and Smith [18] have shown that the simplified Newton iteration is unstable. That is, a small perturbation Δ in X_k, say the one generated by roundoff, may lead to divergence of the sequence obtained by replacing X_k by $X_k + \Delta$. Thus divergence may occur even though the computation of X_k is performed with a numerically stable algorithm. This makes the iteration of almost no practical use.

In this paper we present a suitable modification of the simplified Newton iteration which guarantees stability. Moreover we prove that, choosing $X_0 = I$, convergence occurs for any A having eigenvalues in the set $\mathcal{D} = \{z \in \mathbb{C} : |z| = 1, \operatorname{Re} z > 0\}$. This restriction can be relaxed by means of a suitable scaling, and we provide an algorithm which converges for any A for which $A^{1/p}$ is defined. The iteration that we obtain in this way has quadratic convergence and a cost per step of $O(n^3 \log p)$ arithmetic operations (ops).

Regarding available algorithms, an efficient numerical method for the principal pth root uses the Schur form and was originally proposed by Björck and Hammarling [3] for the square root, then extended by Higham [9] who suggested using the real Schur form for real matrices, and generalized by Smith [18] to the matrix pth root. This method, implemented in the MATLAB toolbox [10], is numerically stable and requires $O(n^3 p)$ ops [11]. The p factor in the operation count is a drawback for large p, and it is desirable to have methods whose cost grows more slowly with p. An interesting analysis of computing the principal matrix pth root has been performed in [2], where the problem is investigated in terms of structured matrix computations and where the Newton iteration for the equation $X^p - A^{-1}$ is proposed. Other methods can be designed based on the identities $A^{1/p} = \exp(\frac{1}{p} \log A)$, where the functions $\log(\cdot)$ and $\exp(\cdot)$ are the matrix generalizations of the customary log and exp functions, respectively [16].

The paper is organized in the following way. In section 2 we show that for $X_0 = I$, Newton's iteration converges for any matrix A with eigenvalues in $\mathcal{D}$. In section 3 we discuss instability issues and propose new variants of (1.2) which, while keeping the same cost of $O(n^3 \log p)$ ops, are proved to be stable in a neighborhood of the solution. In section 4 we describe our general algorithm and discuss some related computational issues. Finally in section 5 we present some numerical experiments and compare our method with the Schur method and with the method based on logarithm and exponential. These results confirm the numerical stability and the overall good performance of the new algorithms.

In the rest of the paper we use the notation $\pi/2p$ instead of $\pi/(2p)$ for the sake of readability.

[illegible] 1. It was observed in [12, 18] that if A has no nonpositive real eigenvalues and if X_0 commutes with A, then the iterates generated by (1.2) coincide with the ones generated by the Newton method in the Banach algebra of the matrices $n \times n$

for the equation $F(X) = X^p - A = 0$; that is

$$X_{k+1} = X_k - F'^{-1}_{X_k}\left(F(X_k)\right), \tag{1.5}$$

provided that the X_k are well defined. The symbol F'_{X_k} here denotes the Fréchet derivative computed at the point X_k. Unfortunately, even if the Fréchet derivative is nonsingular in a neighborhood of $A^{1/p}$, for some choice of A and X_0 the Newton method (1.5) may break down while the simplified one (1.2) still can be applied. See, for instance, [11].

For this reason we will not consider the general theory of the Newton method in Banach algebras, but only the theory of rational iterations. In fact, this approach is easily generalizable to root-finding algorithms different from the Newton method.

2. Convergence. For $p > 2$ rational iterations such as (1.1) have a complicated behavior [14], and it is very difficult to describe the set of initial values for which the iteration converges to a root. The matrix case has a similar behavior; indeed it can be reduced to the scalar one.

Our goal is to determine the set of $A \in \mathbb{C}^{n\times n}$ for which Newton's iteration converges to $A^{1/p}$ for an initial value X_0. The usual choice $X_0 = A$ [12, 18] gives a complicated convergence region; here we show that with $X_0 = I$ the convergence region is more suitable for designing a globally convergent algorithm.

First, we consider A diagonalizable, i.e., $A = M^{-1}DM$ with D diagonal and M nonsingular. The general case has similar behavior and will be discussed later. Since $X_0 = I$, all the iterates are diagonalizable and we may define $D_k = MX_kM^{-1}$ so that (1.2) becomes

$$D_{k+1} = \frac{(p-1)D_k + DD_k^{1-p}}{p}, \tag{2.1}$$

which involves only diagonal matrices, and is essentially n uncoupled scalar iterations of the type

$$\begin{cases} x_{k+1} = \dfrac{(p-1)x_k + \lambda x_k^{1-p}}{p}, \\ x_0 = 1, \end{cases} \tag{2.2}$$

with λ being an eigenvalue of A.

Thus our main problem is to determine the set $\mathcal{B}_p$ of λ such that the iteration (2.2) with $x_0 = 1$ is well defined and converges to the principal pth root $\lambda^{1/p}$, i.e., a pth root of λ whose argument lies in the sector $\mathcal{S}_p$ of (1.4).

For any diagonalizable matrix A having eigenvalues in $\mathcal{B}_p$, the Newton iteration, with $X_0 = I$, converges to $A^{1/p}$. It is not surprising that the sets $\mathcal{B}_p$, for $p > 2$, are bounded by fractals similar to the Julia set of Newton's iteration.

Some of these sets are sketched in Figure 2.1, in which we made a grid of 400×400 points corresponding to a discretization $\widehat{Q}$ of the square

$$Q = \{z \in \mathbb{C}, -3 \le \operatorname{Re} z \le 3, -3 \le \operatorname{Im} z \le 3\}$$

and computed some steps of the Newton sequence (2.2) for $\lambda \in \widehat{Q}$. We plotted in light gray the points λ for which the sequence x_k [illegible] to the principal pth root of λ, and in dark gray the others.

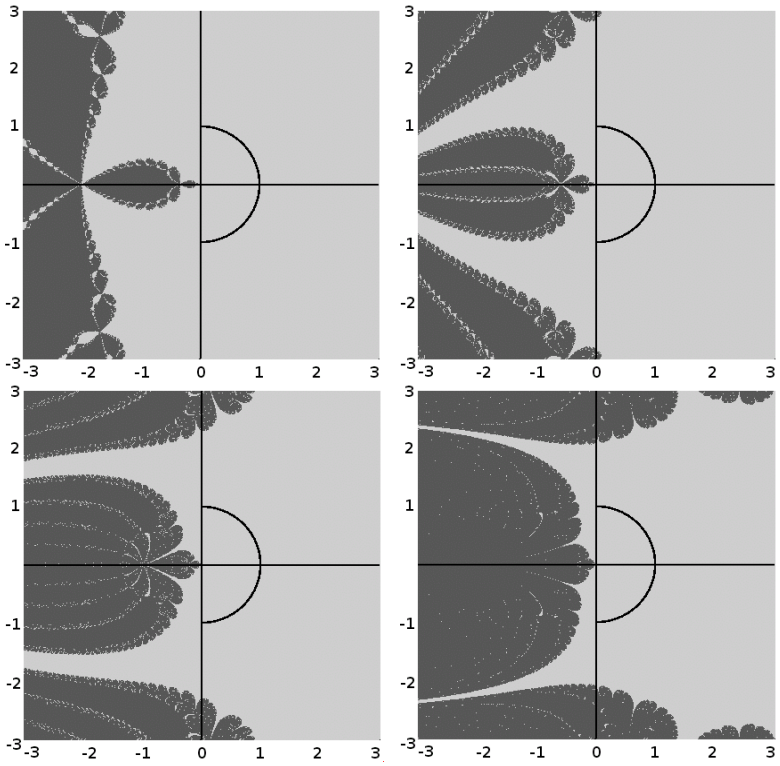


FIG. 2.1. *A sketch of the region of convergence for* $p = 3, 5, 10, 100$ *and the set* $\mathcal{D}$ *of* (2.3). *In light gray the points for which iteration* (2.2) *converges to their principal pth root.*

It is easy to show, by means of standard arguments on the real Newton method, that the positive real axis belongs to $\mathcal{B}_p$ for every p.

The following theorem synthesizes our main convergence result.

THEOREM 2.1. $\mathcal{B}_p$

$$\mathcal{D} = \{z \in \mathbb{C}\,, \operatorname{Re} z > 0, |z| \leq 1\} \tag{2.3}$$

$p > 1$

Consequently if A has its eigenvalues in the set $\mathcal{D}$, then the iteration (1.2), with initial value $X_0 = I$, is well defined and converges to $A^{1/p}$.

For a general matrix A with no nonpositive real eigenvalues, the normalized matrix square root $B = A^{1/2}/\|A^{1/2}\|$, where $\|\cdot\|$ is a generic matrix operator norm, has eigenvalues in the set $\mathcal{D}$. In fact for the spectral radius of B one has $\rho(B) \leq \|B\| = 1$ and since the spectrum of $A^{1/2}$ belongs to the right half-plane, the spectrum of B belongs to the set $\mathcal{D}$. Thus the Newton method applied to the matrix equation $X^p - B = 0$, starting with $X_0 = I$, converges to $B^{1/p}$. Moreover, it is possible to recover $A^{1/p} = (B^{1/p})^2$.

To prove Theorem 2.1, we use the following property.

PROPOSITION 2.2. λ

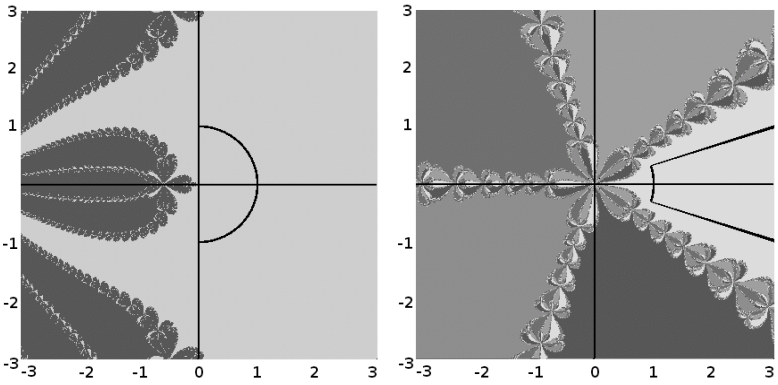


FIG. 2.2. *For $p = 5$ the set $\mathcal{D}$ of Theorem 2.1 (left) and the set $\mathcal{D}_5$ of Theorem 2.3 (right).*

(2.2) $\lambda^{1/p}$

$$\begin{cases} z_{k+1} = \dfrac{(p-1)z_k + z_k^{1-p}}{p}, \\ z_0 = \lambda^{-1/p} \end{cases} \tag{2.4}$$

1

. The proof follows from the equation $z_k = x_k\lambda^{-1/p}$, which can be proved by induction. □

Observe that (2.4) is the Newton method applied to the equation $x^p - 1 = 0$. A similar was used in [2]. The above property provides a connection between the set $\mathcal{B}_p$ and the basin of attraction of the root $x = 1$, which we denote by $\mathcal{A}_p(1)$. In fact, a complex number $a \neq 0$ belongs to $\mathcal{B}_p$ if and only if $a^{-1/p}$ belongs to $\mathcal{A}_p(1) \cap \mathcal{S}_p$.

In this way, we can restate Theorem 2.1 in the following form.

THEOREM 2.3. $\mathcal{A}_p(1)$ $\mathcal{D}_p = \{z \in \mathcal{S}_{2p}, |z| \geq 1\}$ $p > 1$ $\mathcal{S}_p$ (1.4)

A graphical example of the swap between the two theorems is given in Figure 2.2.

2.1. Proof of Theorem 2.3. Define

$$N_p(z) = \frac{(p-1)z^p + 1}{pz^{p-1}} \tag{2.5}$$

for the Newton step and denote by $N_p^{(k)}$ the k-fold composition $N \circ N \circ \cdots \circ N$. Observe also that the function $N_p(z)$ is well defined in $\mathcal{D}_p$.

The proof can be divided into two stages. First, we show that Theorem 2.3 holds if two inequalities are satisfied. Second, we show the validity of such inequalities.

We consider three sets depending on the positive values ξ_p and R_p (see Figure 2.3):

1. a $E_p = \{z \in \mathbb{C}, |z - 1| < R_p\}$ of center 1 and radius R_p;
2. a , $F_p = \{z \in \mathbb{C}, 1 \leq |z| < \xi_p, |\arg(z)| \leq \pi/2p\}$;
3. a , $G_p = \{z \in \mathbb{C}, |z| \geq \xi_p, |\arg(z)| \leq \pi/2p\}$.

We provide an algebraic equation with real solution $s_p = 1 - R_p$ and such that the disk E_p is contained in $\mathcal{A}_p(1)$; then we provide a second algebraic equation with real solution ξ_p and such that each point of the set G_p is transformed by $N_p^{(k)}$ into a point in F_p for some $k \geq 1$. Finally we show that given a point z in F_p, the supremum

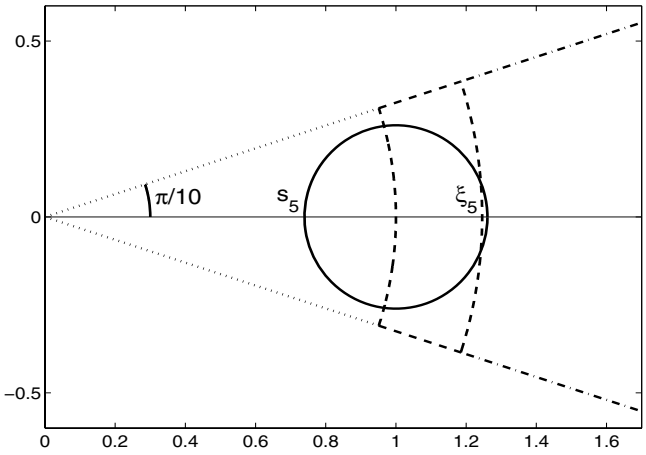


FIG. 2.3. *The three sets used in the proof of Theorem* 2.3 *for the case* $p = 5$*: the circle of radius* $1 - s_5$*, the mincing knife (dash contour), and the blunt wedge (dash-dot contour).*

of the distance of $N_p(z)$ from 1 is reached in the corners of the mincing knife. So, in order to prove that the points in F_p are transformed into points in E_p, it is enough to compute $|N_p(z)-1|$ in the corners of F_p and prove that these values are less than R_p. These are the desired inequalities. In fact, by verifying such inequalities for a specific value of p, one can easily verify that $\mathcal{D}_p \subset F_p \cup G_p$ is a subset of $\mathcal{A}_p(1)$, which is the statement of Theorem 2.3.

We start by giving a way to find a disk centered at the point $z = 1$, such that Newton's iteration converges if x_0 is in this disk.

LEMMA 2.4.

$$(2p-1)s^p - 2ps^{p-1} + 1 = 0 \tag{2.6}$$

s_p $(0,1)$ z $|z-1| < R_p = 1 - s_p$ $z \neq 1$ $|N_p(z)-1| < |z-1|$

. Let us start from inequality $|N_p(z)-1| < |z-1|$, namely

$$\left| \frac{(p-1)z^p - pz^{p-1} + 1}{pz^{p-1}} \right| < |z-1|. \tag{2.7}$$

The polynomial $\phi_p(z) = (p-1)z^p - pz^{p-1} + 1$ can be factorized as $\phi_p(z) = ((p-1)z^{p-2} + \cdots + 3z^2 + 2z + 1)(z-1)^2$, and the inequality (2.7) becomes

$$\frac{|(p-1)z^{p-2} + \cdots + 3z^2 + 2z + 1||z-1|}{|pz^{p-1}|} < 1, \quad z \neq 1. \tag{2.8}$$

Now,

$$\frac{|(p-1)z^{p-2} + \cdots + 2z + 1||z-1|}{|pz^{p-1}|} \leq \frac{1}{p}\left(\frac{p-1}{|z|} + \cdots + \frac{2}{|z|^{p-2}} + \frac{1}{|z|^{p-1}} \right) |z-1|.$$

If $|z-1| < 1-s$, then $|z|^n > s^n$ for every n and the inequality (2.7) holds if

$$\frac{1}{p}\left(\frac{p-1}{s} + \cdots + \frac{2}{s^{p-2}} + \frac{1}{s^{p-1}} \right)(1-s) - 1 \leq 0. \tag{2.9}$$

Multiplying both sides of the above inequality by $ps^{p-1}(1-s)$, with $0 < s < 1$, yields $\phi_p(s) - ps^{p-1}(1-s) \leq 0$, that is,

$$(2p-1)s^p - 2ps^{p-1} + 1 \leq 0. \tag{2.10}$$

It is not difficult to show that the function $f_p(s) = (2p-1)s^p - 2ps^{p-1} + 1$ has the following properties: $f_p(0) > 0$, $f_p(1) = 0$, $f_p'(1) > 0$, and f_p has only a relative minimum in the interval $(0,1)$. All these facts guarantee that the equation $f_p(s) = 0$ has a unique solution s_p in the interval $(0,1)$ and that the inequality (2.10) holds for every $s_p \leq s \leq 1$.

To conclude, we recall that $|z-1| < 1-s$, and so for $0 < |z-1| < R_p = 1 - s_p$, it holds that $|N_p(z)-1| < |z-1|$, which was what we wanted to show. Moreover, we have a constructive way to find R_p by solving the polynomial equation of (2.10) in the interval $(0,1)$. □

This lemma guarantees that for each positive real value $R < R_p$ the closed disk of center 1 and radius R belongs to $\mathcal{A}_p(1)$. Moreover, it holds that $|N_p(s_p)-1| = |s_p - 1|$ and then Lemma 2.4 is not true for any value $R > R_p$.

In order to prove that the set $\mathcal{D}_p$ is a subset of $\mathcal{A}_p(1)$, we split $\mathcal{D}_p$ into two subsets. The former is sent by N_p into the disk of convergence found above, and the latter is sent into the former after some iterations. First, we give a technical lemma that states that any point of a blunt wedge is transformed by N_p into a point of the wedge. This will be used to show that a point of modulus greater than 1 gets closer to 1, after some iterations, but still remains in the sector.

LEMMA 2.5. [illegible] $|z| > 1$ [illegible] $z \in \mathcal{S}_{2p}$ [illegible] $|N_p(z)| < |z|$ [illegible] $|\arg(N_p(z))| \leq |\arg(z)|$

[illegible]. For the first statement, it is easy to show that $|z| > 1$ yields

$$|N_p(z)| = \left|\frac{(p-1)}{p}z + \frac{1}{pz^{p-1}}\right| \leq \frac{(p-1)}{p}|z| + \frac{1}{p|z|^{p-1}} < \frac{(p-1)}{p}|z| + \frac{1}{p} < |z|.$$

For the second statement, let $z = re^{i\theta}$ with $r > 1$ and $0 < |\theta| < \pi/2p$; moreover, let $N(z) = r_1 e^{i\theta_1}$. Our goal is to prove that $|\theta_1| \leq |\theta|$, which is equivalent to $|\tan\theta_1| \leq |\tan\theta|$. From the definition of N_p it can be shown that

$$\tan\theta_1 = \frac{r^p(p-1)\sin\theta - \sin((p-1)\theta)}{r^p(p-1)\cos\theta + \cos((p-1)\theta)}$$

so that inequality $|\tan\theta_1| \leq |\tan\theta|$, for $\theta > 0$, becomes

$$-\frac{\sin\theta}{\cos\theta} \leq \frac{r^p(p-1)\sin\theta - \sin((p-1)\theta)}{r^p(p-1)\cos\theta + \cos((p-1)\theta)} \leq \frac{\sin\theta}{\cos\theta}. \tag{2.11}$$

By means of trigonometric identities, the second inequality of (2.11) is equivalent to $\sin(p\theta) \geq 0$, which is true because $0 < \theta < \pi/2p$.

The first inequality (2.11) is equivalent to

$$r \geq \sqrt[p]{\frac{\sin((p-2)\theta)}{(p-1)\sin(2\theta)}},$$

which is true in the region we have considered, where $r > 1$, and

$$\frac{\sin((p-2)\theta)}{(p-1)\sin(2\theta)} < 1.$$

The case $\theta < 0$ is analogous by symmetry, and the case $\theta = 0$ is trivial. □

Even if a point of the sector having modulus greater than a real number $R > 1$ is transformed by the Newton step N_p into another point in the sector, we need to cut the wedge enough so that each point of the blunt wedge is transformed by N_p into a point of modulus greater than 1. In the next lemma, we find a real value ξ_p that satisfies this condition and is the least in modulus.

LEMMA 2.6. [illegible]

$$(p-1)^2 s^{2p} - p^2 s^{2p-2} + 1 = 0 \tag{2.12}$$

[illegible] $1 < \xi_p < 2$ [illegible] $z \in \mathcal{S}_{2p}$ [illegible] $|z| > \xi_p$ [illegible] $|N(z)| > 1$

[illegible]. Let $R \geq 1$ and let us consider the set $K = \{z \in \mathbb{C}, |z| \geq R, |\arg(z)| \leq \pi/2p\}$. The minimum of $|N_p(z)|$ on the set K is attained at the point $z_0 = R\mathrm{e}^{i\pi/2p}$. In order to prove this, let $z = r\mathrm{e}^{i\theta}$ and consider $|N_p(z)|$ as a function of θ. We have

$$f_r(\theta) = |N_p(z)| = \left| \frac{(p-1)z^p + 1}{pz^{p-1}} \right| = \frac{1}{pr^{p-1}} |(p-1)r^p \mathrm{e}^{ip\theta} + 1|.$$

Observe that $f_r(\theta)$ is minimum for $\theta = \pi/2p$. Moreover, since

$$g(r) = |N_p(r\mathrm{e}^{i\pi/2p})|^2 = \frac{(p-1)^2 r^{2p} + 1}{(pr^{p-1})^2}$$

is increasing for $r > 1$, we deduce that the minimum of $|N_p(z)|$ is attained at the corners of K and in particular at the point z_0. Now, in order to prove that $|N_p(z)| \geq 1$, we solve the equation $|N_p(s\mathrm{e}^{i\pi/2p})| = 1$, that is,

$$\frac{\sqrt{(p-1)^2 s^{2p} + 1}}{ps^{p-1}} = 1,$$

which yields (2.12). Now, it is not difficult to show that the function $g_p(s) = (p-1)^2 s^{2p} - p^2 s^{2p-2} + 1$ in (2.12) has the following properties: $g_p(1) < 0$, $g_p'(1) < 0$, $g_p(2) > 0$, and g_p has only a critical point (a minimum) in the interval $(1, 2)$. All these facts guarantee that (2.12) has a unique solution ξ_p in the interval $(1, 2)$. Therefore, if $|z| \geq \xi_p$, it holds that $|N_p(z)| \geq |N_p(\xi_p \mathrm{e}^{i\pi/2p})| = 1$ and this completes the proof. □

From Lemmas 2.5 and 2.6 we can conclude that a point of the set $\mathcal{D}_p$ having modulus greater than ξ_p is sent, after some iterations, into a point of $\mathcal{D}_p$ having modulus less than ξ_p.

Now, if the mincing knife

$$F_{p,R} = \{z \in \mathbb{C}, 1 \leq |z| \leq R, |\arg z| \leq \pi/2p\}, \tag{2.13}$$

with $R = \xi_p$, is sent into the ball $|N_p(z) - 1| < R_p$, then the theorem is true.

In the next lemma, we show that the maximum of the function $|N_p(z) - 1|$ on the mincing knife is attained at one of the corners and so it is enough to check if these four points are sent into the ball of convergence (for the symmetry of the problem we need to check only two of them).

LEMMA 2.7. [illegible] $R > 1$ [illegible] $f(z) = |N_p(z) - 1|$ [illegible] $F = F_{p,R}$ [illegible] (2.13) [illegible] F

. Let $z = re^{i\theta}$ be a point of F. Observing that $N(\overline{z}) = \overline{N(z)}$, it is enough to consider the case $\theta \geq 0$.

We show that, restricted to the circle of radius $r \geq 1$, the function f is nondecreasing with respect to θ (nonnegative), and hence the maximum lies on the segment corresponding to $\theta = \pi/2p$, $1 \leq r \leq R$. Then, we show that the function is convex in this segment and then takes its maximum at one of the two vertices, which are the top corners.

To simplify the problem, consider the function

$$\begin{aligned}\hat{f}(r,\theta) &= p^2|N_p(z) - 1|^2 - p^2 \\ &= (p-1)^2 r^2 + \frac{1}{r^{2p-2}} + \frac{2(p-1)}{r^{p-2}}\cos(p\theta) - 2(p^2-p)r\cos(\theta) - \frac{2p}{r^{p-1}}\cos((p-1)\theta),\end{aligned}$$

which has the same point of maximum of $|N_p(z) - 1|$ and is simpler.

First, consider the restriction of $\hat{f}$ to an arc relative to a fixed value of r and study the behavior with respect to θ.

Define $g_r(\theta) = \hat{f}(r,\theta)$. We prove that $g_r(\theta)$ is nondecreasing by showing that its derivative,

$$g_r'(\theta) = \frac{2p(p-1)}{r^{p-1}}\left(\sin((p-1)\theta) - r\sin(p\theta) + r^p\sin(\theta)\right),$$

is nonnegative for $0 \leq \theta \leq \pi/2p$. From the sine addition formula, one has

$$\begin{aligned}&\sin((p-1)\theta) - r\sin(p\theta) + r^p\sin(\theta) \\ &\qquad = \sin((p-1)\theta)(1 - r\cos(\theta)) + r\sin(\theta)(-\cos((p-1)\theta) + r^{p-1}) \geq 0,\end{aligned} \tag{2.14}$$

and the last inequality follows from

$$\frac{r\cos(\theta) - 1}{r(r^{p-1} - \cos((p-1)\theta))} \leq \frac{r-1}{r(r^{p-1}-1)} \leq \frac{1}{r\sum_{k=0}^{p-2} r^k} \leq \frac{1}{p-1} \leq \frac{\sin(\theta)}{\sin((p-1)\theta))},$$

where we used the fact that $r \geq 1$ and that the inequality $\sin(n\theta) \leq n\sin(\theta)$ holds for any positive integer n and $0 \leq \theta \leq \pi/2p$.

The inequality (2.14) implies that $g_r(\theta)$ is nondecreasing for any $r \geq 1$, and then the maximum of $\hat{f}$ (and of f) is assumed on the segment of F corresponding to $\theta = \pi/2p$.

Consider the function $\varphi(r) = f(r, \pi/2p)$ on the interval $[1, R]$. We claim that $\varphi(r)$ is a convex function, namely $\varphi''(r) \geq 0$. Since $\cos(p\frac{\pi}{2p}) = 0$ and $\cos((p-1)\frac{\pi}{2p}) = \sin(\pi/2p)$, it holds that

$$\varphi(r) = (p-1)^2 r^2 + \frac{1}{r^{2p-2}} - \frac{2p}{r^{p-1}}\sin(\pi/2p) - 2p(p-1)r\cos(\pi/2p).$$

For its second derivative it holds that

$$\begin{aligned}\varphi''(r) &= 2(p-1)^2 + \frac{2(p-1)(2p-1)}{r^{2p}} - \frac{2p^2(p-1)}{r^{p+1}}\sin(\pi/2p) \\ &\geq 2(p-1)^2 + \frac{2(p-1)(2p-1)}{r^{2p}} - \frac{2(p-1)(2p-1)}{r^{p+1}} = \widetilde{h}(r).\end{aligned}$$

The inequality follows from $p^2\sin(\pi/2p) \leq (2p-1)$ for $p \geq 2$.

Now, $\widetilde{h}(r)$ is positive, and in fact can be rewritten as

$$\frac{2(p-1)}{r^{2p}}\left(r^{p-1}(r^{p+1}(p-1)-2p+1)+2p-1\right) \geq \frac{2(p-1)}{r^{2p}}\left(r^{p+1}(p-1)\right) \geq 0,$$

since $r^{p-1} \geq 1$. The positivity of $\widetilde{h}$ implies that the second derivative of $\varphi(r)$ is positive as well; then the function $\varphi(r)$ is convex so that, restricted to any $[a,b] \subset [1,+\infty)$, it takes its maximum at one of the edges a or b, and the proof is completed. □

Finally we have a procedure to prove that for a value $p > 1$, Theorem 2.3 is true.

- Compute an approximation of R_p and ξ_p by means of some zero-finder method.
- Check if $|N_p(\xi_p e^{i\pi/2p}) - 1| < R_p$ and $|N_p(e^{i\pi/2p}) - 1| < R_p$.

To conclude, it is enough to prove that the two inequalities are true for every $p \geq 3$ (the case $p = 2$ is relatively easy and was treated in [8]). We find an explicit expression for a sequence $b_p \leq R_p$ and a sequence $a_p \geq \xi_p$, and then we prove that

$$|N_p(e^{i\pi/2p}) - 1| < b_p, \quad |N_p(a_p e^{i\pi/2p}) - 1| < b_p.$$

This is enough; in fact by Lemma 2.7 applied to the set F_{p,a_p}, it holds that

$$|N_p(\xi_p e^{i\pi/2p}) - 1| \leq |N_p(a_p e^{i\pi/2p}) - 1| < b_p \leq R_p.$$

We start with a lemma that gives explicitly values for a_p and b_p.

LEMMA 2.8. [illegible] $e^{-\alpha}(1+2\alpha)-1=0$ [illegible] α_0 [illegible]

$$\xi_p \leq a_p = \frac{p}{p-1}, \qquad R_p \geq b_p = \frac{\alpha}{p}$$

[illegible] $0 < \alpha \leq \alpha_0$

[illegible] ξ_p is the solution greater than 1 of $g_p(s) = 0$, where $g_p = (p-1)^2 s^{2p} - p^2 s^{2p-2} + 1 = s^{2p-2}((p-1)^2 s^2 - p^2) + 1$. Now, $g_p(\frac{p}{p-1}) = 1 > 0$ and from the arguments in the proof of Lemma 2.6, it follows that $a_p > \xi_p$.

Concerning R_p, let us consider the polynomial $f_p = (2p-1)s^p - 2ps^{p-1} + 1$. The number $s_p = 1 - R_p$ is the solution of the equation $f_p = 0$ and $0 < s_p < 1$. From the proof of Lemma 2.4, proving that $f_p(1-b_p) < 0$ means that $1 - b_p \geq s_p$ and then $b_p \leq R_p$.

To find a lower bound to R_p of the type α/p, let us consider a generic $0 < \alpha < 3$ that yields

$$f_p\left(1-\frac{\alpha}{p}\right) = \left(\frac{p-\alpha}{p}\right)^{p-1}\left(\frac{\alpha-(2\alpha+1)p}{p}\right)+1.$$

In order to have $f_p(1-\alpha/p) < 0$, it is enough to prove that for every $p > 2$ the sequence

$$d_p = \left(\frac{p-\alpha}{p}\right)^{p-1}\left(\frac{(2\alpha+1)p-\alpha}{p}\right)$$

is greater than 1. This sequence is decreasing for $\alpha > 0$, as we will prove in Lemma 2.9, and its limit is $e^{-\alpha}(1+2\alpha)$. Therefore, $f_p(1-\alpha/p) > 1$ if $e^{-\alpha}(1+2\alpha) > 1$ and this holds for each $0 < \alpha \leq \alpha_0$, where α_0 is the solution in $(0,3)$ of the equation $e^{-\alpha}(1+2\alpha) = 1$. It is easy to prove that this solution exists and is unique and that $\alpha_0 > 1.256$. □

LEMMA 2.9. [illegible] d_p [illegible] 2.8 [illegible]. It is sufficient to prove that the function

$$f(x) = \left(\frac{x-\alpha}{x}\right)^{x-1} \left(\frac{(2\alpha+1)x-\alpha}{x}\right)$$

is decreasing for $x \geq 3$. For this purpose we prove that $f'(x)$ is negative. We have $f'(x) = g(x)h(x)$ with $h(x)$ trivially positive and

$$g(x) = \log\left(\frac{x-\alpha}{x}\right) + \frac{(x-1)\alpha}{x(x-\alpha)} + \frac{\alpha}{x((2\alpha+1)x-\alpha)}$$

is negative; in fact it is increasing and its limit to infinity is 0. To prove that $g(x)$ is increasing, it is enough to show that its derivative is positive, which holds by a direct inspection. ∎

Now we can finally complete the proof of our main theorem by means of the following lemma.

LEMMA 2.10. [illegible] F_{p,a_p} [illegible] N_p [illegible] 1 [illegible] b_p

$$\left|N_p(\mathrm{e}^{i\pi/2p}) - 1\right| < \frac{\alpha_0}{p}, \qquad \left|N_p\left(\frac{p}{p-1}\mathrm{e}^{i\pi/2p}\right) - 1\right| < \frac{\alpha_0}{p}.$$

[illegible]. For the point $z = \mathrm{e}^{i\pi/2p}$ we have

$$|N_p(z) - 1|^2 = \frac{1}{p^2}\left(2p^2 - 2p - 2 - 2p(p-1)\cos\left(\frac{\pi}{2p}\right) - 2p\sin\left(\frac{\pi}{2p}\right)\right).$$

Since $p > 2$ and $\cos(x) \geq 1 - x^2/2$ and $\sin(x) \geq x - x^3/6$ for $0 < x < \pi/2$,

$$p^2|N_p(z) - 1|^2 \leq 2p^2 - 2p + 2 - (2p^2 - 2p)\left(1 - \frac{\pi^2}{8p^2}\right) - 2p\left(\frac{\pi}{2p} - \frac{\pi^3}{48p^3}\right)$$
$$= 2 + \frac{\pi^2}{4} - \pi + \left(\frac{\pi^3}{24p^2} - \frac{\pi^2}{4p}\right) \leq 2 + \frac{\pi^2}{4} - \pi + \frac{\pi^3}{24 \cdot 9} < 1.47 < 1.57 < \alpha_0^2 = p^2 b_p,$$

which is what we wanted to prove.

For the point $z = a_p \mathrm{e}^{i\pi/2p}$, setting $\gamma_p = (\frac{p-1}{p})^{p-1} = a_p^{1-p}$, one has

$$|N_p(z) - 1|^2 = \frac{1}{p^2}\left(2p^2 + \gamma_p^2 - 2p^2\cos\left(\frac{\pi}{2p}\right) - 2p\gamma_p\sin\left(\frac{\pi}{2p}\right)\right).$$

It is possible to prove as in Lemma 2.9 that γ_p is a decreasing sequence that tends to $1/\mathrm{e}$; thus it holds that $1/\mathrm{e} = \gamma_\infty \leq \gamma_p \leq \gamma_3 = 4/9$.

Finally we have

$$p^2|N_p(z) - 1|^2 \leq 2p^2 + \gamma_3^2 - 2p^2\left(1 - \frac{\pi^2}{8p^2}\right) - 2p\gamma_\infty\left(\frac{\pi}{2p} - \frac{\pi^3}{48p^3}\right)$$
$$= \left(\frac{4}{9}\right)^2 + \frac{\pi^2}{4} - \frac{\pi}{\mathrm{e}} + \frac{\pi^3}{24 \cdot 9\mathrm{e}} < 1.563 < 1.57 < \alpha_0^2 = p^2 b_p.$$

This completes the proof. ∎

A consequence of this proof is the applicability of the scalar Newton method because the sequence z_k of (2.4) never reaches zero in $\mathcal{D}_p$, and so the sequence (2.2) never reaches zero in $\mathcal{D}$.

2.2. Matrix convergence. We have shown that if the matrix A is diagonalizable, then the iteration can be reduced to uncoupled scalar iterations, one for each of the eigenvalues. In the general case, by means of the Jordan canonical form of A, we may restrict our attention to the case where $A \in \mathbb{C}^{n\times n}$ is a Jordan block, $J(\lambda, n)$, and λ belongs to the region $\mathcal{D}$ defined in (2.3).

In this case, define the functions $g_k(\lambda)$ as the kth iterate x_k of the sequence (2.2) and $f_k(z_0)$ as the kth iterate z_k of (2.4) and let $\phi(\lambda) = \lambda^{-1/p}$ be defined on the set $\mathbb{C}\backslash(-\infty, 0]$. From Proposition 2.2 it follows that for any $z \in \mathbb{C}\backslash(-\infty, 0]$, $g_k(z) = (f_k \circ \phi)(z)z^{1/p}$. Observe that for the matrix iteration (1.2) with initial value $X_0 = I$, it holds that $X_k = g_k(J)$. We aim to prove that $g_k(J)$ converges to $J^{1/p}$ and that the convergence is quadratic.

Let us recall that a function applied to a Jordan block is defined as [16, p. 311]

$$f(J) = \begin{bmatrix} f(\lambda) & f'(\lambda) & \dots & \frac{f^{(n-1)}(\lambda)}{(n-1)!} \\ & \ddots & \ddots & \vdots \\ & & f(\lambda) & f'(\lambda) \\ 0 & & & f(\lambda) \end{bmatrix}.$$

Then to prove Jordan block convergence from scalar convergence it is sufficient to prove that

$$\frac{g_k^{(n)}(\lambda)}{n!} \longrightarrow \frac{1}{n!}\frac{d^n}{dz^n} z^{1/p}\Big|_{z=\lambda}, \qquad n = 1, 2, 3, \dots.$$

We prove this fact in two steps. First, we show that the sequence $g_k(z)$ converges uniformly on any compact subset of an open neighborhood of any point z belonging to the set $\mathcal{D}$ of (2.3). Then, we show that the derivatives of g_k evaluated at λ converge to the derivative of the pth root function evaluated at λ and that the convergence of $g_k(J)$ to $J^{1/p}$ is dominated by a quadratically convergent sequence.

We use the notation $\|f(x)\|_K = \sup_K |f(x)|$.

LEMMA 2.11. [illegible] $g_k(z)$ [illegible] $z^{1/p}$ [illegible] $\mathcal{G} = \{z \in \mathbb{C}, \operatorname{Re} z > 0, |z| < 1 + \varepsilon\}$ [illegible] $\varepsilon > 0$ [illegible]. By the proof of Theorem 2.3, the set $\{z \in \mathbb{C},\ |z| \geq 1,\ |\arg z| \leq \pi/2p\}$ is a subset of the immediate basin of attraction $\mathcal{F}$ for the fixed point 1 of the rational iteration f_k, which is open; thus the compact arc $\{|z| = 1,\ |\arg z| \leq \pi/2\}$ admits a finite open covering belonging to $\mathcal{F}$ and then there exists δ such that $\mathcal{G}_p = \{z \in \mathbb{C},\ |z| > 1 - \delta,\ |\arg z| < \pi/2\}$ is a subset of $\mathcal{F}$ and, from the properties of the Fatou set [14, 1], the set $\{f_k\}$ is a normal family on $\mathcal{G}_p$, and, by an easy argument it can be shown that the sequence f_k converges uniformly to 1 for any compact subset of $\mathcal{G}_p$ (see [1, Thm. 6.3.1]).

Now, consider a compact set $\widetilde{K} \subset \mathcal{G} = \{z \in \mathbb{C}, \operatorname{Re} z > 0, |z| < (1-\delta)^{-p}\}$, since $\phi(z) = z^{-1/p}$ is a continuous map from the set $\mathcal{G}$ to the set $\mathcal{G}_p$, $\phi(\widetilde{K}) = K$ is a compact subset of $\mathcal{G}_p$, and, from what we said above, $\|f_k(z) - 1\|_K \to 0$. If we set $\|z^{1/p}\|_{\widetilde{K}} = M$, then

$$\|g_k(z) - z^{1/p}\|_{\widetilde{K}} = \|z^{1/p}((f_k \circ \phi)(z) - 1)\|_{\widetilde{K}} \leq M\|(f_k \circ \phi)(z) - 1\|_{\widetilde{K}} = M\|f_k(z) - 1\|_K,$$

and, since the last term tends to zero, the proof is thus achieved by choosing $\varepsilon = (1-\delta)^{-p} - 1$. ∎

To conclude, consider a compact neighborhood $K \subset \mathcal{D}$ of λ and a circle γ of radius R, centered in λ and fully contained in K. The Cauchy formula yields

$$\left| \frac{g_n^{(k)}(\lambda)}{k!} - \frac{1}{k!}\frac{d^k}{dz^k} z^{1/p} \Big|_{z=\lambda} \right| = \left| \frac{1}{2\pi \mathrm{i}} \oint_\gamma \frac{g_n(z) - z^{1/p}}{(z-\lambda)^{k+1}} dz \right| \le \frac{1}{R^k} \|g_n(z) - z^{1/p}\|_K \to 0,$$

and then $g_n(J)$ converges to $J^{1/p}$. Moreover, $\|g_n(J) - J^{1/p}\|_\infty \le \alpha \|f_n(z) - 1\|_{\phi(K)}$ for some constant α, and the sequence $f_n(z)$ converges to 1 in any compact subset of $\mathcal{D}_p$ and the convergence is quadratic (since it converges uniformly and in a neighborhood of 1, it converges quadratically).

This approach can be generalized without any effort to any rational iteration applied to a matrix.

3. Stable variants of the Newton method. Two stable iterations for the matrix square root, that is, the Denman and Beavers iteration [6, 8]

$$\begin{cases} X_0 = A, \quad Y_0 = I, \\ X_{k+1} = \dfrac{1}{2}\,(X_k + Y_k^{-1}), \quad Y_{k+1} = \dfrac{1}{2}\,(Y_k + X_k^{-1}), \quad k = 0, 1, \ldots, \end{cases} \tag{3.1}$$

and the Meini iteration [17]

$$\begin{cases} Y_0 = I - A, \quad Z_0 = 2(I + A), \\ Y_{k+1} = -Y_k Z_k^{-1} Y_k, \quad Z_{k+1} = Z_k - 2Y_k Z_k^{-1} Y_k, \quad k = 0, 1, \ldots, \end{cases} \tag{3.2}$$

are variants of the Newton iteration. In particular the latter can be rewritten as an iteration for the increment [13]

$$\begin{cases} X_0 = A, \quad H_0 = \dfrac{1}{2}(I - A), \\ X_{k+1} = X_k + H_k, \quad H_{k+1} = -\dfrac{1}{2} H_k X_{k+1}^{-1} H_k. \end{cases} \tag{3.3}$$

In fact, the instability of the simplified Newton iterations $X_{k+1} = (X_k + A X_k^{-1})/2$ and $X_{k+1} = (X_k + X_k^{-1} A)/2$, shown by Higham [8], is mainly due to the pre- or post-multiplication of X_k^{-1} by A. On the other hand, since X_k commutes with A (see [13]), the iteration can be rewritten as

$$X_{k+1} = \frac{X_k + A^{1/2} X_k^{-1} A^{1/2}}{2} \tag{3.4}$$

and also is stable in this new form, as one can see by a particular case of the analysis made in section 3.1. Obviously (3.4) is useless since it involves the square root of A, but it helps us to stabilize the iteration by introducing the variable $Y_k = A^{-1/2} X_k A^{-1/2} = A^{-1} X_k = X_k A^{-1}$. The resulting iteration is that of Denman and Beavers; we refer the reader to [13] for more details on this subject.

Repeating these arguments for the pth root, one has the simplified Newton iteration $X_{k+1} = \frac{1}{p}\big((p-1)X_k + A X_k^{1-p}\big)$ or $X_{k+1} = \frac{1}{p}\big((p-1)X_k + X_k^{1-p} A\big)$, which are unstable as shown in [18]. We show in section 3.1 that, since the instability is due to the one-sided multiplication by A, the modified equation

$$X_{k+1} = \frac{(p-1)X_k + (A^{1/p} X^{-1})^{p-1} A^{1/p}}{p} \tag{3.5}$$

provides in principle an iteration with optimal stability.

Now, with the square root in mind, we introduce the auxiliary variable $N_k = AX_k^{-p}$. It can be shown by induction that with the initial values $X_0 = I$ and $N_0 = A$, each of X_k, N_k, and A commutes with the others. This provides the following variant of the simplified Newton iteration:

$$\begin{cases} X_0 = I, \quad N_0 = A, \\ X_{k+1} = X_k \left(\dfrac{(p-1)I + N_k}{p} \right), \\ N_{k+1} = \left(\dfrac{(p-1)I + N_k}{p} \right)^{-p} N_k. \end{cases} \tag{3.6}$$

Observe that the matrix A does not explicitly appear in the iteration. We denote with the acronym `HWA` (handled without A) iterations having this feature. Observe that, while X_k converges to $A^{1/p}$, the sequence N_k converges to the identity matrix.

On the other hand, one can introduce the increment

$$H_k = \frac{AX_k^{1-p} - X_k}{p} = -\frac{X_k^{1-p}}{p} \left(X_k^p - A \right), \tag{3.7}$$

where $X_k^p - A$ is the [illegible] at the step k. Note that H_k commutes with A and X_k. From (3.7) we obtain $A = (X_k + pH_k)X_k^{p-1}$, which allows us to write

$$H_{k+1} = -\frac{X_{k+1}^{1-p}}{p} \left(X_{k+1}^p - A \right) = -\frac{X_{k+1}^{1-p}}{p} \left(X_{k+1}^p - (X_k + pH_k)X_k^{p-1} \right).$$

Now, because $X_{k+1} = X_k + H_k$ we obtain

$$\begin{aligned} H_{k+1} &= -\frac{X_{k+1}^{1-p}}{p} \left(X_{k+1}^p - (pX_{k+1} - (p-1)X_k)X_k^{p-1} \right) \\ &= -\frac{X_{k+1}X_{k+1}^{-p}}{p} \left(X_{k+1}^p - pX_{k+1}X_k^{p-1} + (p-1)X_k^p \right) \\ &= -\frac{X_{k+1}}{p} \left(I - pX_{k+1}^{1-p}X_k^{p-1} + (p-1)X_k^p X_{k+1}^{-p} \right). \end{aligned} \tag{3.8}$$

Setting $F_k = X_k X_{k+1}^{-1}$ we can write an iteration for the increment of the Newton iteration

$$\begin{cases} X_0 = I, \quad H_0 = \dfrac{(A-I)}{p}, \\ X_{k+1} = X_k + H_k, \quad F_k = X_k X_{k+1}^{-1}, \\ H_{k+1} = -X_{k+1} \left(\dfrac{I - F_k^p}{p} + F_k^{p-1}(F_k - I) \right), \end{cases} \tag{3.9}$$

where the expression for H_{k+1} has been written in a form that reduces the phenomenon of numerical cancellation.

Unfortunately, the iteration (3.9) does not reduce to (3.3) in the case of the square root. A nicer form that generalizes (3.3) is

(3.10)
$$\begin{cases} X_0 = I, \quad H_0 = \dfrac{(A-I)}{p}, \\ X_{k+1} = X_k + H_k, \quad F_k = X_k X_{k+1}^{-1} \\ H_{k+1} = -\dfrac{1}{p} H_k (X_{k+1}^{-1} I + 2X_{k+1}^{-1} F_k + 3X_{k+1}^{-1} F_k^2 + \cdots + (p-1) X_{k+1}^{-1} F_k^{p-2}) H_k. \end{cases}$$

We call it incremental Newton (IN). Even if the form (3.10) is more symmetric than (3.9), its computational cost is higher; in fact the computation of H_{k+1} in the iteration (3.10) can be performed in $O(n^3 p)$ ops, and in the iteration (3.9), it can be performed in $O(n^3 \log p)$ ops.

3.1. Stability analysis. In accordance with [5] we define an iteration $X_{k+1} = f(X_k)$ to be [illegible] $X = f(X)$ if the error matrices $E_k = X_k - X$ satisfy

$$E_{k+1} = L(E_k) + O(\|E_k\|^2),$$

where L is a linear operator that has bounded powers; that is, there exists a constant $c > 0$ such that for all $n > 0$ and arbitrary E of unit norm, $L^n(E) < c$. This means that a small perturbation introduced in a certain step will not be amplified in the subsequent iterations.

Note that this definition of stability is an asymptotic property and is different from the usual concept of numerical stability, which concerns the global error propagation, aiming to bound the minimum relative error over the computed iterates.

First, we show that the iteration (3.5) has [illegible]; i.e., the operator L coincides with the null operator. Then we show that the iterations (3.6) and (3.9) are stable.

With $E_k = X_k - A^{1/p}$, we have

$$E_{k+1} = X_{k+1} - A^{1/p} = \frac{p-1}{p} X_k + \frac{A^{1/p} X_k^{-1} \cdots A^{1/p} X_k^{-1} A^{1/p}}{p} - A^{1/p}. \tag{3.11}$$

Now

$$X_k^{-1} = (A^{1/p} + E_k)^{-1} = A^{-1/p} - A^{-1/p} E_k A^{-1/p} + O(\|E_k\|^2).$$

From this relation we obtain that

$$A^{1/p} X_k^{-1} \cdots A^{1/p} X_k^{-1} A^{1/p} = A^{1/p} - (p-1) E_k + O(\|E_k\|^2). \tag{3.12}$$

Finally combining (3.11) and (3.12) yields

$$E_{k+1} = \frac{p-1}{p}(A^{1/p} + E_k) + \frac{A^{1/p} - (p-1)E_k}{p} - A^{1/p} + O(\|E_k\|^2) = O(\|E_k\|^2), \tag{3.13}$$

which means that this iteration is stable, and the most stable possible according to our definition because $L = 0$.

Now we consider the iteration (3.6) and introduce the error matrices $E_k = X_k - A^{1/p}$ and $F_k = N_k - I$. For the sake of simplicity, we perform a [illegible]; that is, we omit all the terms that are quadratic in the errors. Equality up to second order terms is denoted with the symbol $\doteq$.

From $N_k = I + F_k$, one has

$$\left(\frac{(p-1)I + N_k}{p}\right)^{-p} \doteq \left(I + \frac{F_k}{p}\right)^{-p} \doteq I - F_k, \tag{3.14}$$

and the relation for the errors becomes

$$\begin{bmatrix} E_{k+1} \\ F_{k+1} \end{bmatrix} \doteq \begin{bmatrix} I & \frac{1}{p}A^{1/p} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} E_k \\ F_k \end{bmatrix} = L \begin{bmatrix} E_k \\ F_k \end{bmatrix}. \tag{3.15}$$

The coefficient matrix L is idempotent ($L^2 = L$) and hence has bounded powers. Thus the iteration is stable.

For the iteration (3.9) define the error matrices $M_k = X_k - A^{1/p}$ and H_k; then

$$M_{k+1} = X_{k+1} - A^{1/p} = X_k - A^{1/p} + H_k = M_k + H_k. \tag{3.16}$$

For H_{k+1} the relation is a bit more complicated.

Using (3.16) we can write

$$X_{k+1}^{-1} = (A^{1/p} + M_k + H_k)^{-1} \doteq A^{-1/p} - A^{-1/p} M_k A^{-1/p} - A^{-1/p} H_k A^{-1/p}$$

and

$$F_k = X_k X_{k+1}^{-1} = (A^{1/p} + M_k) X_{k+1}^{-1} \doteq I - H_k A^{-1/p}.$$

The latter equation enables us to write

$$(X_k X_{k+1}^{-1})^q \doteq (I - H_k A^{-1/p})^q \doteq I - q H_k A^{-1/p}. \tag{3.17}$$

Finally we have

$$\begin{aligned} H_{k+1} &= -X_{k+1}\left(\frac{I - F_k^p}{p} + F_k^{p-1}(F_k - I)\right) \\ &\doteq -X_{k+1}\left(\frac{I - I + pH_kA^{-1/p}}{p} + (I - (p-1)H_kA^{-1/p})H_kA^{-1/p}\right) \doteq 0. \end{aligned}$$

In conclusion it holds that

$$\begin{bmatrix} M_{k+1} \\ H_{k+1} \end{bmatrix} \doteq \begin{bmatrix} I & I \\ 0 & 0 \end{bmatrix} \begin{bmatrix} M_k \\ H_k \end{bmatrix} = L \begin{bmatrix} M_k \\ H_k \end{bmatrix}. \tag{3.18}$$

Since the matrix L is idempotent, then also this iteration is stable. A similar result holds for the iteration (3.10). Observe also that, unlike in the iteration (3.6), the norm of L is independent of A.

2. The iteration analyzed in [2],

$$X_{k+1} = \frac{1}{p}\left((p+1)X_k - X_k^{p+1}A\right), \quad X_0 = I, \tag{3.19}$$

is obtained by applying Newton's iteration to the equation $X^{-p} - A = 0$, which has the same convergence as the scalar iteration $x_{k+1} = \left((p+1)x_k - x_k^{p+1}\lambda_i\right)/p$ for $x_0 = 1$, applied to any eigenvalue λ_i of the matrix A.

Like any polynomial iteration of degree greater than 2, this one has the point $x = \infty$ as (super)attractive fixed point [1], and so the basins for the roots are considerably

smaller than the ones for the Newton iteration for $x^p - a = 0$. However, in [2] it is proved that the basin of attraction to 1 contains a disk of center 1 and radius 1. In the same paper, it is shown that the iteration (3.19) is unstable for general matrices. The instability of this iteration can be easily removed by applying the arguments of this section. In fact after simple manipulations we deduce the mathematically equivalent iteration

$$\text{(3.20)} \qquad \begin{cases} X_0 = I, \quad N_0 = A, \\ X_{k+1} = X_k \left(\dfrac{(p+1)I - N_k}{p} \right), \\ N_{k+1} = \left(\dfrac{(p+1)I - N_k}{p} \right)^p N_k, \end{cases}$$

which is proved to be stable near the solution. One can see the similarity to the `HWA` method.

The iteration (3.20) was already found by Lakić [15] as the first case of a family of stable iterative methods for computing the inverse pth root.

4. The algorithm. Here we present our algorithm for computing the principal pth root of a matrix having no nonpositive real eigenvalues. For $p = 2$ one can use the existing algorithms [6, 17, 13], so we assume that we can perform the square root.

ALGORITHM 1 (iteration for the principal pth root of a matrix A).

- Input: a matrix A, an integer $p > 2$, and an algorithm for computing the square root.
- Compute B, the principal square root of A.
- Set $C = B/\|B\|$ for a suitable norm. The eigenvalues of C belong to the set $\mathcal{D}$ of (2.3)
- By means of iteration (3.6) or (3.9)
  - If p is even, compute $S = C^{2/p}$, the $(p/2)$th root of C, and set $X = S\|B\|^{2/p}$.
  - If p is odd, compute $S = C^{1/p}$, the pth root of C, and set $X = \left(S\|B\|^{1/p}\right)^2$.

Observe that both iterations (3.6) and (3.9) can be performed in $O(n^3 \log p)$ ops per step, by means of the binary powering technique, much less than the cost of Schur method which is $O(n^3 p)$ ops. However, for small values of p, the total number of ops needed by Algorithm 1 might be larger than the number of ops needed by the Schur method.

For computing the square root, one can use the algorithm (3.3), possibly with a suitable scaling if needed, or the Schur method; this does not affect the asymptotic order of complexity with respect to p. In our numerical experiments, we have observed that the choice of the square root algorithm used in preprocessing the matrix is crucial for the accuracy of the computed solution. Using the Schur method for computing the preliminary square root and then the iteration (3.6) gives good results comparable to the ones obtained with the algorithm proposed by Smith [18]. In certain cases, it is more convenient to use an iterative method such as (3.3), to compute the preliminary square root.

More details can be given about the operation count and the number of steps needed for the numerical convergence; in fact, two matrix multiplications, one inversion, and a matrix exponentiation to the power p are sufficient to carry out one step of the `HWA` iteration. For computing the power X^p, with X being a matrix, one can use

the binary powering technique with a cost varying from $\lfloor \log_2 p \rfloor$ to $2\lfloor \log_2 p \rfloor$ matrix multiplications. The total cost of one step is then bounded by $(3 + 2\lfloor \log_2 p \rfloor)n^3$ ops. For the `IN` iteration, the cost is $(p+2)n^3$ ops per step, and for the iteration (3.9) the cost is $(5 + 2\lfloor \log_2(p-1) \rfloor)n^3$ ops per step.

As shown in section 2, the numerical convergence depends only on the localization of the eigenvalues. The closer they are to the boundary of the basin of convergence, the greater is the number of steps needed. For matrices of the form $C = A^{1/2}/\|A^{1/2}\|$, having eigenvalues in the set $\mathcal{D}$ of (2.3), the slow convergence occurs when some eigenvalue is near 0, namely, when the matrix A is ill-conditioned. For instance, if A is a symmetric positive definite matrix and we use the 2-norm, it is easy to show that the smallest eigenvalue of C is $\sqrt{1/\mu_2(A)}$, where $\mu_2(A) = \|A\|_2\|A^{-1}\|_2$ is the condition number of A. Being C diagonalizable by a unitary transform, the convergence of the matrix iteration is the same as the convergence of its smallest eigenvalue. To get an estimation of the number of steps needed by the Newton method applied to a symmetric matrix, it is enough to compute the number of steps needed by the sequence (1.1) with $a = \sqrt{1/\mu_2(A)}$ to converge.

Even though the number of steps is a growing function of p, it seems bounded from above by a constant.

Finally, it is important to point out that the algorithm we proposed works only to find the principal pth root. It is not clear if it can be used to compute any primary pth root, in particular, roots having eigenvalues in different sectors. One important advantage of the Schur method is that it can be used to compute any primary pth root, not just $A^{1/p}$.

5. Numerical experiments. We have performed several experiments in MATLAB 7. We have compared our algorithms with the simplified Newton (`SN`) method (1.2), with the Schur method implemented in the function `rootm` of the Matrix Computation Toolbox [10], and with the method based on the formula $A^{1/p} = \exp(\frac{1}{p}\log(A))$, using the functions `logm` and `expm` of MATLAB (this method was suggested by an anonymous referee).

For computing the square root of a matrix, we used the function `sqrtm` of MATLAB, which is based on the Schur form of A, or the iteration (3.3), and if a scaling is needed in (3.3) we used the one proposed in [13]. These algorithms have the same asymptotic cost of $O(n^3)$.

To compute the power to $-p$ in the iteration (3.6), first we compute the pth power of the matrix with the binary powering technique and then we invert the matrix. We stop the iterations when the residuals begin to grow or become `NaN`.

Test 1. To illustrate the instability near the solution of the `SN` method (1.2) and the stability of the proposed variants, we consider the simple 3×3 matrix

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

and compute the fourth root of the matrix A^4. In Figure 5.1 we have compared the relative residual defined as $\mathcal{R}(X) = \|X^p - A\|_F/\|A\|_F$ for the three methods: `SN` of equation (1.2), Newton in the version (`HWA`) provided by equation (3.6), and `IN` of equation (3.10). We denote by $\|A\|_F$ the Frobenius norm of the matrix A, i.e., $\|A\|_F = (\sum_{i,j=1}^n a_{ij}^2)^{1/2}$. As one can see, for some steps the three methods give the same residual; in fact they are analytically equivalent, but the `SN` method has some instability problems even after a few steps. Our methods show good stability.

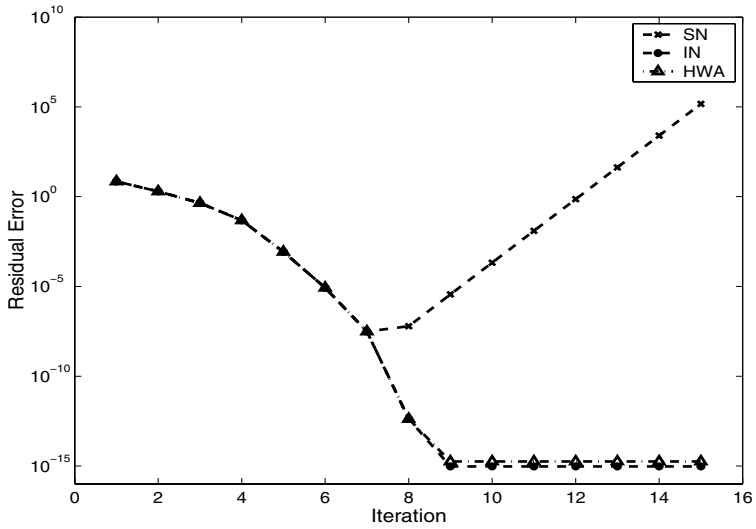


FIG. 5.1. *Comparison of the simplified Newton (*`SN`*) method, the iteration handled without A (*`HWA`*), and incremental Newton (*`IN`*).*

TEST 2. We consider some ill-conditioned matrices to compare the behavior of Algorithm 1 with the one based on the Schur form of A and with the formula $A^{1/p} = \exp(\frac{1}{p}\log(A))$. We also illustrate that the choice of the algorithm used for computing the preliminary square root on Algorithm 1 is very important for the numerical accuracy of the computed solution.

We compute pth root with four methods:

- Compute the square root with the function `sqrtm` in MATLAB and then the iteration (3.6) to compute the pth root (`sqrtm+HWA`).
- Compute the square root by means of the `IN` iteration (3.3) and then the iteration (3.6) to compute the pth root (`IN+HWA`).
- Compute the pth root with the algorithm based on the Schur form (`Smith`).
- Compute $A^{1/p} = \exp(\frac{1}{p}\log(A))$ (`explog`).

The first class of matrices we considered is the class of Hilbert matrices $H_{ij} = 1/(i+j)$ that is a traditional example of an ill-conditioned matrix. We denote by `hilb(n)` the n-dimensional Hilbert matrix. The second class is the prolate matrix, which is a symmetric ill-conditioned Toeplitz matrix whose entries are defined by the formula $A_{ii} = 1/2, A_{ij} = \sin(\pi(j-i)/2)/(\pi(j-i))$. We denote by `prolate(n)` the n-dimensional prolate matrix. The third class is the Frank matrix, an upper Hessenberg matrix with real, positive eigenvalues occurring in reciprocal pairs, half of which are ill-conditioned. We denote by `frank(n)` the n-dimensional Frank matrix. The fourth class is the companion matrix of the polynomial $x^n - 10^{-12}$, whose roots are the nth root of 10^{-12}. We denote by `compan(n)` the n-dimensional companion matrix.

In Table 5.1 we report the relative residuals and the number of iterations (for `HWA` iteration) in computing the 59th root for some of these matrices. As one can see, our algorithm, if provided with a Schur implementation for the preliminary square root, is competitive with the Smith method and provides the same results, in terms of accuracy, if tested with these ill-conditioned matrices. The `explog` algorithm gives good results, but a bit worse than our algorithm in the hardest examples.

A purely iterative algorithm (the second column) suffers from very bad condi-

TABLE 5.1
Comparison of methods for computing the 59*th root of some test matrix.*

Example	`sqrtm+HWA`		`IN+HWA`		`Smith`	`explog`
`hilb(5)`	$6.6 \cdot 10^{-15}$	11	$4.4 \cdot 10^{-15}$	11	$3.1 \cdot 10^{-14}$	$8.5 \cdot 10^{-15}$
`hilb(10)`	$1.7 \cdot 10^{-14}$	20	$1.6 \cdot 10^{-14}$	21	$2.2 \cdot 10^{-14}$	$2.7 \cdot 10^{-14}$
`prolate(10)`	$1.6 \cdot 10^{-14}$	14	$2.1 \cdot 10^{-14}$	12	$3.3 \cdot 10^{-14}$	$2.2 \cdot 10^{-14}$
`prolate(20)`	$3.1 \cdot 10^{-14}$	20	$4.3 \cdot 10^{-14}$	22	$3.4 \cdot 10^{-14}$	$4.8 \cdot 10^{-14}$
`frank(10)`	$2.0 \cdot 10^{-11}$	15	$7.4 \cdot 10^{-10}$	15	$3.5 \cdot 10^{-10}$	$4.5 \cdot 10^{-9}$
`frank(14)`	$3.5 \cdot 10^{-5}$	22	$2.6 \cdot 10^{-2}$	24	$9.8 \cdot 10^{-4}$	$8.4 \cdot 10^{-2}$
`compan(5)`	$1.7 \cdot 10^{-3}$	26	$8.3 \cdot 10^{-8}$	27	$5.0 \cdot 10^{-2}$	$1.5 \cdot 10^{-1}$
`compan(15)`	$1.4 \cdot 10^{0}$	31	$8.8 \cdot 10^{-6}$	30	$4.2 \cdot 10^{1}$	$6.0 \cdot 10^{0}$

tioning of the matrix, but it is faster and in certain cases, like the examples of the companion matrix, gives better results.

Note that when using the procedure described in section 4 for the Hilbert and prolate matrices, which are symmetric, one has a predicted number of steps that almost coincides with that of the examples.

Scaling the iteration for the preliminary square root was not necessary in these examples, but it is worth remarking that sometimes it is important to use a scaling technique in order to avoid poor results.

Acknowledgments. I would like to thank Prof. D. A. Bini for many useful discussions about almost every topic on numerical linear algebra and polynomial computation and Prof. N. J. Higham for providing very useful suggestions for improving the paper. I wish to thank two anonymous referees for their helpful comments which improved the presentation.

REFERENCES

[1] A. F. BEARDON, *Iteration of Rational Functions*, Grad. Texts in Math. 132, Springer-Verlag, New York, 1991.
[2] D. A. BINI, N. J. HIGHAM, AND B. MEINI, *Algorithms for the matrix pth root*, Numer. Algorithms, 39 (2005), pp. 349–378.
[3] Å. BJÖRCK AND S. HAMMARLING, *A Schur method for the square root of a matrix*, Linear Algebra Appl., 52/53 (1983), pp. 127–140.
[4] A. CAYLEY, *The Newton-Fourier imaginary problem*, Amer. J. Math., 2 (1879), p. 97.
[5] S. H. CHENG, N. J. HIGHAM, C. S. KENNEY, AND A. J. LAUB, *Approximating the logarithm of a matrix to specified accuracy*, SIAM J. Matrix Anal. Appl., 22 (2001), pp. 1112–1125.
[6] E. DENMAN AND A. BEAVERS, *The matrix sign function and computations in systems*, Appl. Math. Comput., 2 (1976), pp. 63–94.
[7] M. A. HASAN, A. A. HASAN, AND K. B. EJAZ, *Computation of matrix nth roots and the matrix sector function*, in Proceedings of the 40th Annual IEEE Conference on Decision and Control, 2001, pp. 4057–4062.
[8] N. J. HIGHAM, *Newton's method for the matrix square root*, Math. Comp., 46 (1986), pp. 537–549.
[9] N. J. HIGHAM, *Computing real square roots of a real matrix*, Linear Algebra Appl., 88 (1987), pp. 405–430.
[10] N. J. HIGHAM, *The Matrix Computation Toolbox*, http://www.ma.man.ac.uk/∼higham/mctoolbox.
[11] N. J. HIGHAM, *Functions of a Matrix: Theory and Computation*, in preparation.
[12] W. D. HOSKINS AND D. J. WALTON, *A faster, more stable method for computing the pth root of positive definite matrices*, Linear Algebra Appl., 26 (1979), pp. 139–163.
[13] B. IANNAZZO, *A note on computing the matrix square root*, Calcolo, 40 (2003), pp. 273–283.
[14] G. JULIA, *Mémoire sur l'itération des fonctions rationelles*, J. Math. Pures Appl. (9), 8 (1918), pp. 47–245.

[15] S. Lakić, *On the computation of the matrix kth root*, ZAMM Z. Angew. Math. Mech., 78 (1998), pp. 167–172.
[16] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, 2nd ed., Academic Press, Orlando, FL, 1985.
[17] B. Meini, *The matrix square root from a new functional perspective: Theoretical results and computational issues*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 362–376.
[18] M. I. Smith, *A Schur algorithm for computing matrix pth roots*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 971–989.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 2, pp. 524–550

MIQR: A MULTILEVEL INCOMPLETE QR PRECONDITIONER FOR LARGE SPARSE LEAST-SQUARES PROBLEMS*

NA LI† AND YOUSEF SAAD†

Abstract. This paper describes a multilevel incomplete QR factorization for solving large sparse least-squares problems. The algorithm builds the factorization by exploiting structural orthogonality in general sparse matrices. At any given step, the algorithm finds an independent set of columns, i.e., a set of columns that have orthogonal patterns. The other columns are then block orthogonalized against columns of the independent set, and the process is repeated recursively for a certain number of levels on these remaining columns. The final level matrix is processed with a standard QR or incomplete QR factorization. Dropping strategies are employed throughout the levels in order to maintain a good level of sparsity. A few improvements to this basic scheme are explored. Among these is the relaxation of the requirement of independent sets of columns. Numerical tests are proposed which compare this scheme with the standard incomplete QR preconditioner, the robust incomplete factorization preconditioner, and the algebraic recursive multilevel solver (on normal equations).

Key words. multilevel incomplete QR factorization, CGLS, QR factorization, orthogonal factorization, incomplete QR, preconditioning, iterative methods, large least-squares problems, normal equations

AMS subject classifications. 65F10, 65F20, 65F50

DOI. 10.1137/050633032

1. Introduction.

This paper considers iterative solution methods for linear least-squares problems of the form

$$\min_x \|b - Ax\|_2, \tag{1.1}$$

where $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$ is a large sparse matrix with full rank. Problems of this type arise in many scientific and engineering applications including data analysis, computational fluid dynamics, simulation, signal processing, and control problems, to name just a few. As engineers and scientists are benefiting from increased availability of data as well as computational resources, these problems are inevitably becoming harder to solve due to their size as well as their ill-conditioning. For example, the papers [4, 34] mention a problem of this type which arises from an animal breeding study with 60 million unknowns. In the very different area of three-dimensional computer graphics, one encounters certain least-squares problems which have complexity proportional to the number of geometry primitives [23], which, in desirable models, should include millions of polygons. Problems from such applications are usually very sparse and can be solved iteratively or by sparse orthogonal factorizations. Iterative solution methods may have an advantage over direct methods, depending on the underlying sparsity pattern.

However, if iterative methods are to be used, then preconditioning is essential. Although it is known that iterative solution algorithms are not effective without pre-

*Received by the editors June 3, 2005; accepted for publication (in revised form) by D. B. Szyld February 2, 2006; published electronically July 31, 2006. This work was supported by NSF grants ACI-0305120 and INT-0003274 and by the Minnesota Supercomputing Institute.
http://www.siam.org/journals/simax/28-2/63303.html
†Department of Computer Science and Engineering, University of Minnesota, 200 Union Street S.E., Minneapolis, MN 55455 (nli@cs.umn.edu, saad@cs.umn.edu).

conditioning, there has been little effort made in developing preconditioners for least-squares problems in recent years. This is in contrast with the solution of standard (square) linear systems, where enormous progress has been made in designing both general purpose preconditioners and specialized preconditioners that are tailored to specific applications. Part of the difficulty stems from the fact that many methods solve the system (1.1) by implicitly solving the normal equations

$$A^T Ax = A^T b, \tag{1.2}$$

whose solution is the same as that of (1.1). The condition number of the coefficient matrix of the normal equations system (1.2) is the square of that of the original matrix A. As a result, the normal equations will tend to be very ill-conditioned. In this situation preconditioning is critical for robustness. However, severe ill-conditioning of the matrix will also tend to make it difficult to obtain a good preconditioner.

Though it is possible to solve the least-squares problem (1.1) by solving normal equations (1.2), forming the system of normal equations explicitly and then solving it is not a recommended approach in general as this suffers from various numerical difficulties; see [6, 15] for details. For small dense problems, the best overall solution method is to use a good orthogonal factorization algorithm such as the Householder QR; see, e.g., [15]. If $A = QR$ is the "thin" QR factorization of A [15], then the solution of (1.1) can be obtained by solving $Rx = Q^T b$ for x. For a comprehensive survey of direct methods, see [6].

Alternatively, iterative methods such as LSQR [28] and SOR [27] have been advocated for solving least-squares problems of the type (1.1) when A is large. A well-known approach is one that is based on solving the normal equations by the conjugate gradient (CG) method. The resulting algorithm is sometimes termed CGNR [31] and sometimes CGLS [6]. The latter acronym is adopted here. This paper only considers CGLS as the accelerator and focuses on developing effective preconditioners. Since we refer to preconditioned CGLS throughout the paper, we now give a brief description of the algorithm, assuming that a preconditioner M for $A^T A$ is available. Recall that a preconditioner is a certain matrix M which approximates the original coefficient matrix (in this case $A^T A$) such that it is inexpensive to solve an arbitrary linear system $Mx = b$.

ALGORITHM 1.1. Left-preconditioned CGLS

1. Compute $r_0 := b - Ax_0$ $\widetilde{r}_0 := A^T r_0$ $z_0 := M^{-1}\widetilde{r}_0$ $p_0 := z_0$
2. For $i = 0, \ldots,$ until convergence
3. $\quad w_i := Ap_i$
4. $\quad \alpha_i := (z_i, \widetilde{r}_i)/\|w_i\|_2^2$
5. $\quad x_{i+1} := x_i + \alpha_i p_i$
6. $\quad r_{i+1} := r_i - \alpha_i w_i$
7. $\quad \widetilde{r}_{i+1} := A^T r_{i+1}$
8. $\quad z_{i+1} := M^{-1}\widetilde{r}_{i+1}$
9. $\quad \beta_i := (z_{i+1}, \widetilde{r}_{i+1})/(z_i, \widetilde{r}_i)$
10. $\quad p_{i+1} := z_{i+1} + \beta_i p_i$
11.

Many variants of the above algorithm exist. In particular, when M is available in the form of a product $M = LL^T$, where L is lower triangular, then the preconditioning operation can be split into two parts and a split-preconditioned CGLS option can be derived. A right-preconditioned option can be developed as well. We consider only the left-preconditioned variant in this paper.

Developing preconditioners for the normal equations, or for problem (1.1), can be approached in a number of ways. A naive approach would be to form the squared matrix A^TA and try to find an incomplete Cholesky factorization of this matrix. The fact that this matrix is symmetric positive definite does not make it easy to find a preconditioner for it. Indeed, most of the theory for preconditioning techniques relies on some form of diagonal dominance. In addition, forming the normal equations suffers from other disadvantages, some of which are the same as those mentioned above for the dense case; in particular there is some loss of information when forming A^TA [6]. Moreover, A^TA can be much denser than the original matrix. In fact one dense row of A will make the entire A^TA matrix dense.

Another approach, one that is taken here, is to try to compute an approximate orthogonal factorization of A. This approach is not new, as will be seen in section 2 which discusses related work. If $A \approx QR$, then $A^TA \approx R^TR$ and this matrix can be used as a preconditioner M. Notice that this approach ignores the factor Q which is not used. In this paper we exploit multilevel ideas similar to those defined for the algebraic recursive multilevel solver (ARMS) in [33, 25]. The idea of multilevel incomplete QR (MIQR) factorization can be easily described with the help of recursion. It is important to observe at the outset that when A is sparse, then many of its columns will be orthogonal because of their structure. These are called structurally orthogonal columns. It is therefore possible to find a large set S of [illegible] orthogonal columns. This set is called an [illegible] of columns. Independent sets are the main ingredient used in ARMS [33, 25]. Once the first independent set S is obtained, we can block orthogonalize the remaining columns against the columns in S. Since the matrix of the remaining columns will still be sparse in general, it is natural to think of recursively repeating the process until a small number of columns are left which can be orthogonalized with standard methods. The end result is a QR factorization of a column-permuted A. With this simple strategy MIQR gradually reduces a large sparse least-squares system into one with a significantly smaller size. It is worth pointing out that although we focus on overdetermined systems ($m > n$), the techniques described are applicable to square matrices ($m = n$) and underdetermined matrices ($m < n$) as well.

Recent developments in the solution of standard linear systems have shown that multilevel preconditioners have excellent scalability and robustness properties; see, e.g., [33, 9, 31, 1, 2, 3]. However, it appears that when it comes to the solution of large general sparse least-squares problems, similar multilevel methods have not been considered so far, in spite of an increasing demand for solving such problems.

The remainder of this paper is organized as follows. After a short section on related work (section 2), we discuss in section 3 the issue of finding independent sets of columns, as this is an important ingredient used in MIQR. Then, a detailed description of MIQR is presented in section 4 followed by strategies to improve the performance of MIQR as well as other implementation details. Numerical results are reported in section 5, and the paper ends with concluding remarks in section 6.

2. Related work. Several general-purpose preconditioners based on techniques such as SSOR, incomplete orthogonal factorization, and incomplete Cholesky factorization have been proposed and analyzed in the literature.

In 1979, Björck introduced a preconditioner based on the SSOR method [5]. In the proposed method, A^TA is written as $A^TA = L + D + L^T$, where L is lower triangular. The normal equations are then preconditioned by

$$M = \omega(2-\omega)(D+\omega L)D(D+\omega L^T).$$

To avoid forming A^TA explicitly, row (or column) projection methods have also been exploited and applied to normal equations [10]. In these methods, only a row or a column of A is needed at any given relaxation step. Block versions of these methods have also been studied [7, 20].

In a 1984 pioneering article, Jennings and Ajiz proposed preconditioners based on incomplete versions of Givens rotations and the Gram–Schmidt process [18]. Since then, several other preconditioners based on incomplete orthogonal factorizations have been studied [30, 35, 29]. If $A = QR$ is the exact thin QR factorization of A, where R is an $n \times n$ upper triangular matrix and Q is an $m \times n$ orthogonal matrix, then $A^TA = R^TR$, and it is usually inexpensive to solve the equation $R^TRx = y$. The incomplete version of the QR factorization (IQR) can be used as a preconditioner for (1.2). Unlike the matrix Q produced by incomplete Givens rotations, which is always orthogonal, the factor Q produced by the incomplete Gram–Schmidt process is not necessarily orthogonal. Nonetheless, the incomplete Gram–Schmidt-based preconditioners are robust and can avoid breakdown when A has full rank. In this approach, dropping strategies can be employed in Q as well as R to reduce intermediate storage requirements. Let P_Q and P_R be zero patterns chosen for matrices Q and R, respectively. The incomplete QR factorization based on the Gram–Schmidt process can be described by the following modification of the incomplete LQ (ILQ) algorithm given in [30]. Note that in practice P_Q and P_R are normally determined dynamically based on the magnitude of the elements generated.

ALGORITHM 2.1. Incomplete QR factorization (IQR)

1. For $j = 1, \ldots, n$
2. Compute $r_{ij} := (a_j, q_i)$ for $i = 1, 2, \ldots, j-1$
3. Replace r_{ij} by zero if $(i, j) \in P_R$
4. Compute $q_j := a_j - \sum_{i=1}^{j-1} r_{ij} q_i$
5. Replace q_{ij} by zero if $(i, j) \in P_Q \quad i = 1, 2, \ldots, m$
6. Compute $r_{jj} := \|q_j\|_2$
7. If $r_{jj} == 0$, then stop; Else compute $q_j := q_j / r_{jj}$
8.

In the above algorithm, the step represented by line 2 computes the inner products of the jth column of A with all previous columns of Q. Most of these inner products are equal to zero because of sparsity. Therefore, it is important to ensure that only the nonzero inner products are calculated for efficiency. The strategy proposed in [30] calculates these inner products as a linear combination of sparse vectors. Specifically, let $r_j = [r_{1j}, r_{2j}, \ldots, r_{j-1,j}]^T$ and $Q_{j-1} = [q_1, q_2, \ldots, q_{j-1}]$; then $r_j = Q_{j-1}^T a_j$ is a sparse matrix by sparse vector product. This product can be computed as a linear combination of the rows in Q_{j-1}; i.e., only the rows corresponding to the nonzero elements in a_j are linearly combined. Since the matrix Q is normally stored column-wise, a linked list pointing to the elements in each row of Q needs to be dynamically maintained. This strategy is also utilized in the implementation of the proposed MIQR algorithm.

Preconditioners based on the incomplete modified Gram–Schmidt process have also been developed. The Cholesky incomplete modified Gram–Schmidt (CIMGS) algorithm of Wang, Gallivan, and Bramley is an incomplete orthogonal factorization preconditioner based on the modified Gram–Schmidt process [35]. The paper explores rigorous strategies for defining incomplete Cholesky factorizations, based on the relation between the Cholesky factorization of A^TA and the QR factorization of A. Other authors studied direct ways to obtain the Cholesky factorization [19, 6]. This type

of approach obtains the incomplete Cholesky factorization of $C = A^T A$, where C may or may not be formed explicitly. As an alternative, Benzi and Tůma proposed a robust incomplete factorization (RIF) preconditioner which computes an incomplete LDL^T factorization of $A^T A$ without explicitly forming it [4]. Their approach utilizes a conjugate Gram–Schmidt process to calculate the factorization $Z^T C Z = D$, where Z is unit upper triangular and D is diagonal. Using the fact that $Z^T = L^{-1}$, they showed that $L_{ji} = (z_j^T C z_i)/(z_j^T C z_j)$. Therefore, the L factor of C can be obtained as a side product of the conjugate orthogonalization process without any extra cost. In section 5 a few comparisons are made between this approach and the MIQR technique proposed in this paper.

There were also a number of attempts to precondition positive definite matrices which may be far from diagonally dominant. In a 1980 paper, Manteuffel [26] suggested shifting a positive definite matrix to get an incomplete Cholesky factorization. This work was pursued more recently in [32], where other diagonal shifting techniques were studied for both incomplete orthogonal factorizations and incomplete Cholesky factorizations.

The idea of utilizing independent sets of columns (rows) in the context of least-squares, or more precisely for normal equations, is not new; see, e.g., [20, 21]. The main goal of these two papers was to exploit independent sets to improve parallelism. Independent sets of columns will be the main ingredient in obtaining an MIQR factorization. In terms of parallel algorithms, Elmroth and Gustavson developed recursive parallel QR factorizations that can be used in direct solvers for dense normal equations [12, 13].

3. Independent sets of columns. The MIQR algorithm proposed in this paper exploits successive independent sets of columns. This section discusses column independent set orderings.

Given a matrix $A = [a_1, a_2, \ldots, a_n]$, where $a_1, a_2, \ldots, a_n$ are column vectors, a subset $\{a_{j_1}, a_{j_2}, \ldots, a_{j_s}\}$ is called [illegible] of A if columns l and k of A are structurally orthogonal for any $l, k \in \{j_1, j_2, \ldots, j_s\}$ and $l \neq k$. Figure 3.1(a) shows an example of such an independent set of five columns (marked as open circles). The issue of finding independent sets of columns is not new and has been discussed in depth in the literature in different contexts; see, e.g., [11, 22, 24, 31] or [14] for a more comprehensive review. Here, we formalize the problem into that of finding an independent set in a graph.

3.1. Finding independent sets of columns. Two columns a_i and a_j of A will be said to be [illegible] if their patterns overlap. This means that if $\hat{a}_k$ is the column vector obtained from a_k by replacing all its nonzero entries by ones, then a_i and a_j are adjacent if and only if $\hat{a}_i^T \hat{a}_j \neq 0$. The opposite of adjacent is [illegible]: two columns a_i and a_j are structurally orthogonal if $\hat{a}_i^T \hat{a}_j = 0$.

Let $\hat{A}$ be the pattern matrix obtained from A by replacing all its nonzero entries by ones. Then, the [illegible] (CIG) (see [11]) of A is the graph with n vertices representing the n columns of A, and with edges defined by the nonzero pattern of $\hat{A}^T \hat{A}$. This means that there is an edge between vertex i and j if and only if $\hat{a}_i$ and $\hat{a}_j$ are adjacent.

Note that an edge from vertex i to vertex j is defined if $\cos\theta_{ij} \neq 0$, where θ_{ij} is the angle between vectors $\hat{a}_i$ and $\hat{a}_j$. Define the following matrices:

$$B = \left[\frac{\hat{a}_1}{\|\hat{a}_1\|}, \frac{\hat{a}_2}{\|\hat{a}_2\|}, \ldots, \frac{\hat{a}_n}{\|\hat{a}_n\|} \right] \quad \text{and} \quad C = B^T B. \tag{3.1}$$

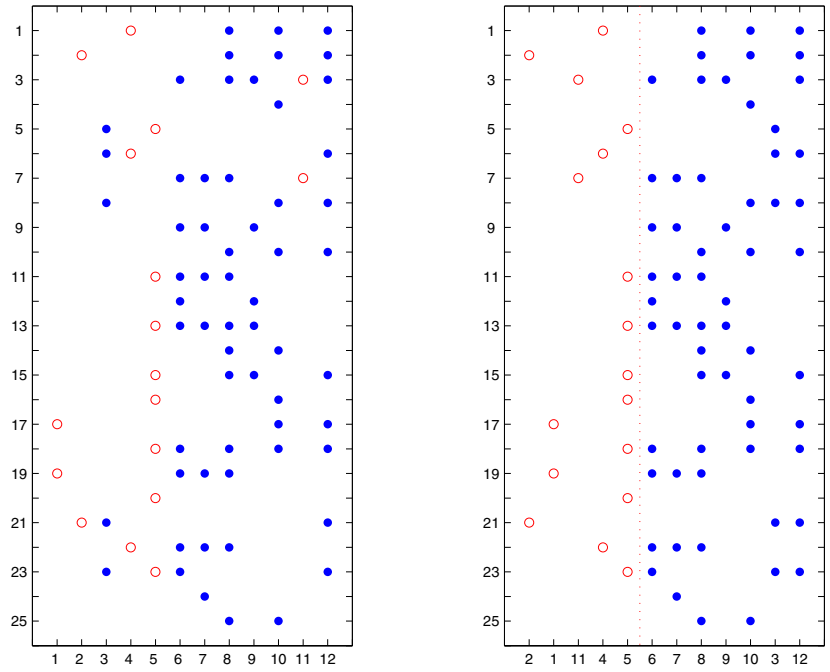


FIG. 3.1. (a) *An independent set of five columns (open circles) in a* 25×12 *matrix.* (b) *The independent set of columns are permuted to the first five columns of the matrix.*

Since the generic entry c_{ij} of C can be written as $c_{ij} \equiv \cos\theta_{ij}$, the graph $CIG(A)$ is nothing but the adjacency graph of C. Therefore, the problem is to find a maximal independent set of a graph. Let E be the set of all edges in $CIG(A)$ and let V be the set of its vertices. Recall that an independent set S is a subset of the vertex set V such that

$$\text{if } x \in S, \text{ then } [(x,y) \in E \text{ or } (y,x) \in E] \Rightarrow y \notin S;$$

i.e., any vertex in S is not allowed to be adjacent with any other vertex in S either by incoming or outgoing edges. An independent set S is maximal if

$$S' \supseteq S \text{ is an independent set} \Rightarrow S' = S.$$

Note that the maximal independent set is not necessarily the independent set with maximum cardinality. In fact, to find the latter is NP hard. In the following, the term independent set will always mean a maximal independent set. The following greedy algorithm (see, e.g., [31]) can be used to find an independent set S. In the algorithm, U is the set of all unmarked vertices, which initially includes all the vertices.

ALGORITHM 3.1. Independent set ordering

1. Let $S := \phi$ and $U := \{1, 2, \ldots, n\}$ $j := 1$
2. While $U \neq \phi$ and $j \leq maxSteps$
3. Let $k :=$ next unmarked vertex in U
4. $S := S \cup \{k\}$ Mark and remove k from U
5. Mark all vertices adjacent to k and remove them from U
6. $j := j + 1$
7.

Let $|S|$ be the size of S. Assume that the maximum degree of all vertices in S is d_S. According to Algorithm 3.1, the total number of marked vertices is $n - |S|$. At the same time, whenever a vertex is added to S, at most d_S vertices will be marked, which means the total number of vertices marked is at most $d_S|S|$. Therefore, we have $n - |S| \leq d_S|S|$ and as a result

$$|S| \geq \frac{n}{1 + d_S}. \tag{3.2}$$

This suggests that we may obtain S with a larger number of vertices by first visiting the vertices with smaller degrees [31].

ALGORITHM 3.2. Independent set ordering with increasing degree traversal

1. Find an ordering $i_1, i_2, \ldots, i_n$ of the vertices by increasing degree.
2. Let $S := \phi$ and $U := \{i_1, i_2, \ldots, i_n\}$. $j := 1$.
3. While $U \neq \phi$ and $j \leq maxSteps$:
4. Let $i_k :=$ next unmarked vertex in U.
5. $S := S \cup \{i_k\}$ Mark and remove i_k from U.
6. Mark all i_k's adjacent vertices and remove them from U.
7. $j := j + 1$
8.

Algorithm 3.2 first sorts the vertices in increasing degree order and then applies the greedy algorithm. In general, Algorithm 3.2 will find a larger independent set at the cost of an initial sorting of the vertices. Algorithm 3.2 is used in the implementation of MIQR.

3.2. Estimates for the size of the independent set. The lower bound of the independent set size given by (3.2) is a rough one. The goal of this section is to find a more accurate estimate of the size of the independent set using a simple probabilistic model.

Consider an $m \times n$ sparse matrix A with N_{nz} nonzero entries and assume that these nonzero entries are randomly distributed. In particular each column will have on average the same number of nonzero entries, which is $\nu \equiv N_{nz}/n$. Under this assumption, Algorithms 3.1 and 3.2 would be equivalent and therefore, we can restrict our study to Algorithm 3.1. We denote by μ the average number of nonzero entries per row; thus $\mu \equiv N_{nz}/m$.

For any column vector a of A, we first calculate the expected number of column vectors that are structurally orthogonal to a. If a has only one nonzero element, then there are on average $n - 1 - (\mu - 1) = n - \mu$ possible columns among $n - 1$ that will be orthogonal to a; thus the probability that any given column is orthogonal to a is $(n - \mu)/(n - 1)$. Since a has ν nonzero elements on average, the probability that any given column is orthogonal to a is

$$p = \left(\frac{n - \mu}{n - 1}\right)^{\nu}. \tag{3.3}$$

As a result the probability that any given column is not orthogonal to a is $1 - p$. Thus, the expected number of column vectors that are structurally orthogonal to a is

$$\eta = (n - 1) \times \left(1 - \left(\frac{n - \mu}{n - 1}\right)^{\nu}\right). \tag{3.4}$$

Note that η is simply the average degree of a node in $CIG(A)$ in the very first step of Algorithm 3.1, since it represents the average number of columns that are not orthogonal to a given column of A.

Consider now an arbitrary step j of Algorithm 3.1. We will call N_j the number of columns left to be considered, i.e., the number of unmarked columns in U at the end of step j of Algorithm 3.1.

LEMMA 3.1. [illegible] N_j [illegible] j [illegible] 3.1 [illegible] $N_0 = n$ [illegible] N_j [illegible]

$$N_j = \left(\left(1 - \frac{\nu}{m}\right) \frac{N_{j-1}}{N_{j-1} - 1}\right)^{\nu} (N_{j-1} - 1). \tag{3.5}$$

[illegible]. We begin by observing that if we consider the matrix consisting of the unmarked columns of A at any given step, then its average number of nonzero entries per column remains unchanged and equal to ν. In contrast, the removal of one column will change μ. If μ_j is the average number of nonzero entries per row for the matrix of unmarked columns, then

$$\mu_j = \frac{\nu}{m} \times N_j. \tag{3.6}$$

Assume that the independent set obtained is $S = \{i_1, i_2, \ldots, i_s\}$, where i_1 is the first vertex added into S, i_2 is the second vertex added into S, and so on. When i_j is added, the estimated number of vertices that are [illegible] marked in line 5 in Algorithm 3.1 is simply the expected number of columns that are not orthogonal to a given column for the matrix of unmarked columns. This is simply the expression (3.4) in the very first step, i.e., when $j = 1$. For a general step it will be the same expression with n replaced by N_j and μ by μ_j. Note that i_j itself is also marked. The new number of unmarked columns is therefore

$$\begin{aligned} N_j &= N_{j-1} - 1 - (N_{j-1} - 1) \times \left(1 - \left(\frac{N_{j-1} - \mu_{j-1}}{N_{j-1} - 1}\right)^{\nu}\right) \\ &= (N_{j-1} - 1) \left(\frac{N_{j-1} - \mu_{j-1}}{N_{j-1} - 1}\right)^{\nu}. \end{aligned}$$

We now introduce $n_j \equiv N_j - 1$ to simplify notation. The above equalities become

$$n_j = \left(\frac{n_{j-1} + 1 - \mu_{j-1}}{n_{j-1}}\right)^{\nu} n_{j-1} - 1 = \left(1 - \frac{\mu_{j-1} - 1}{n_{j-1}}\right)^{\nu} n_{j-1} - 1.$$

Substituting μ_{j-1} given by (3.6) gives

$$\begin{aligned} n_j &= \left(1 - \frac{\frac{(n_{j-1}+1)\nu}{m} - 1}{n_{j-1}}\right)^{\nu} n_{j-1} - 1 = \left(1 - \frac{\nu}{m} + \frac{1 - \frac{\nu}{m}}{n_{j-1}}\right)^{\nu} n_{j-1} - 1 \\ &= \left(\left(1 - \frac{\nu}{m}\right)\left(1 + \frac{1}{n_{j-1}}\right)\right)^{\nu} n_{j-1} - 1, \end{aligned}$$

which is the expression to be proved when $N_j = n_j + 1$ is substituted. □

The lemma should be interpreted in a probabilistic sense: Given a large number of matrices, with the same size (n and m) and the same ν, on average the number

of unmarked columns at the jth step is given by the number resulting from solving the recurrence equation (3.5). The actual number may, of course, be larger or smaller than the N_j given by the lemma.

It is important to note that (3.5) is an exact equality. However, it does not seem possible to obtain a simple closed form expression for N_j. One would be tempted to make the approximation $1/N_j \approx 0$ but this is not valid since toward the end N_j will become small. On the other hand, one can find rough bounds for N_j and substitute them above.

Thus, since $N_j \leq N$ for all j we have

$$N_j \geq \left[\left(1 - \frac{\nu}{m}\right)\frac{n}{n-1}\right]^{\nu}(N_{j-1} - 1).$$

Define

$$\alpha \equiv \left[\left(1 - \frac{\nu}{m}\right)\frac{n}{n-1}\right]^{\nu} \quad \text{and} \quad \gamma \equiv \frac{\alpha}{1-\alpha}.$$

Clearly, we have $\alpha \leq (n/(n-1))^{\nu}$. For large n and m, α will typically be smaller than 1 for a sparse matrix. For example, noting that $\alpha = [(1-\nu/m)(1+1/(n-1)]^{\nu}$, we have

$$\frac{1}{n-1} \leq \frac{\nu}{m} \quad \rightarrow \quad \alpha \leq \left(1 - \left(\frac{\nu}{m}\right)^2\right)^{\nu} < 1.$$

As can be easily seen, the assumption $1/(n-1) \leq \nu/m$ is equivalent to $\mu \geq 1 + \nu/m$, which is generally verified because $\nu \ll m$. We will make the assumption that $\alpha < 1$. This condition is equivalent to

$$\left(1 - \frac{\nu}{m}\right)\frac{n}{n-1} < 1 \quad \leftrightarrow \quad \left(1 - \frac{\nu}{m}\right) < 1 - \frac{1}{n} \quad \leftrightarrow \quad \frac{\nu n}{m} > 1 \quad \leftrightarrow \quad \mu > 1.$$

Then, we have $N_j \geq \alpha N_{j-1} - \alpha$ and since $-\alpha = \alpha\gamma - \gamma$ this becomes $(N_j + \gamma) \geq \alpha(N_{j-1} + \gamma)$. Using this it is possible to estimate the total number of steps which will result in the algorithm, which will be the size of S, since each step will add one more member to S. Let s be the last step of the algorithm and note that we have $N_s \leq N_{s-1} \leq \cdots \leq N_0 = n$. Then the above inequality will yield

$$(N_s + \gamma) \geq \alpha(N_{s-1} + \gamma) \geq \cdots \geq \alpha^s(N_0 + \gamma) \rightarrow (N_s + \gamma) \geq \alpha^s(N_0 + \gamma).$$

The step s at which the algorithm is stopped corresponds to a final size of $N_s = 1$. This means that

$$\alpha^s \leq \frac{1+\gamma}{n+\gamma} = \frac{1}{(1-\alpha)(n + \alpha/(1-\alpha))} = \frac{1}{(1-\alpha)n + \alpha}.$$

Taking logarithms and recalling that $\alpha < 1$ yields

$$s \geq s_{min} \equiv \frac{\log\left[\alpha + (1-\alpha)n\right]}{-\log\alpha}. \tag{3.7}$$

The accuracy of the estimates derived above will be tested in section 5. It will be verified in the experiments that the lower bound (3.7) is not sharp. In fact the experiments indicate, with good consistency, that it is better to use $2s_{min}$ as an estimate of the actual size of S. On the other hand, the estimate given by the direct application of the formula (3.5) can be quite accurate in spite of the simplicity of the underlying model.

4. Multilevel incomplete QR (MIQR) factorizations. This section presents the MIQR preconditioning method for solving sparse least-squares systems. It begins with a discussion of the complete version of the multilevel QR factorization (section 4.1). Then, strategies are proposed to approximate the factorization for preconditioning purposes (section 4.2).

4.1. Multilevel QR factorization (MQR). When the matrix A in (1.1) is sparse, it will most likely have an independent set of columns $a_{j_1}, a_{j_2}, \ldots, a_{j_s}$. Let P_1^T be the permutation matrix which permutes $a_{j_1}, a_{j_2}, \ldots, a_{j_s}$ into the first s columns. Then we have

$$AP_1^T = [A^{(1)}, A^{(2)}], \tag{4.1}$$

where $A^{(1)} = [a_{j_1}, a_{j_2}, \ldots, a_{j_s}]$ is an $m \times s$ matrix and $A^{(2)}$ is an $m \times (n-s)$ matrix. Figure 3.1(b) shows an example of such an ordering. Without loss of generality and for simplicity, we still use $[a_1, a_2, \ldots, a_s]$ and $[a_{s+1}, a_{s+2}, \ldots, a_n]$ to denote the columns of $A^{(1)}$ and $A^{(2)}$, respectively.

Since the columns in $A^{(1)}$ are orthogonal to each other, $(A^{(1)})^T A^{(1)}$ is a diagonal matrix. Then $A^{(1)}$ can be trivially factored as $A^{(1)} = Q_1 D_1$ with

$$Q_1 = \left[\frac{a_1}{\|a_1\|_2}, \frac{a_2}{\|a_2\|_2}, \ldots, \frac{a_s}{\|a_s\|_2} \right] \quad \text{and} \quad D_1 = \operatorname{diag}(\|a_1\|_2, \|a_2\|_2, \ldots, \|a_s\|_2).$$

Now let

$$\begin{aligned} F_1 &= Q_1^T A^{(2)}, \\ A_1 &= A^{(2)} - Q_1 F_1. \end{aligned}$$

Then (4.1) can be rewritten as

$$AP_1^T = [A^{(1)}, A^{(2)}] = [Q_1, A_1] \begin{bmatrix} D_1 & F_1 \\ 0 & I \end{bmatrix}. \tag{4.2}$$

This is a block version of the Gram–Schmidt process, and we have

$$Q_1^T A_1 = 0, \tag{4.3}$$

because $Q_1^T A_1 = Q_1^T (A^{(2)} - Q_1 F_1) = Q_1^T A^{(2)} - F_1 = 0$.

In the simplest one-level method, we apply a standard QR factorization to the reduced $m \times (n-s)$ system A_1:

$$A_1 \widetilde{P}_2^T = Q_2 \widetilde{R}_2,$$

where $\widetilde{P}_2^T$ is an $(n-s) \times (n-s)$ permutation matrix ($\widetilde{P}_2^T$ is the identity matrix when pivoting is not used), Q_2 is an $m \times (n-s)$ orthogonal matrix, and $\widetilde{R}_2$ is an $(n-s) \times (n-s)$ upper triangular matrix. Equation (4.2) can then be rewritten as

$$A = [Q_1, Q_2] \begin{bmatrix} I & 0 \\ 0 & \widetilde{R}_2 \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \widetilde{P}_2 \end{bmatrix} \begin{bmatrix} D_1 & F_1 \\ 0 & I \end{bmatrix} P_1 \tag{4.4}$$

or

$$A = QR_2 P_2 R_1 P_1 = Q\widehat{R} \tag{4.5}$$

if we use the following notations:

$$Q = [Q_1, Q_2], \quad R_1 = \begin{bmatrix} D_1 & F_1 \\ 0 & I \end{bmatrix}, \quad R_2 = \begin{bmatrix} I & 0 \\ 0 & \widetilde{R}_2 \end{bmatrix}, \quad P_2 = \begin{bmatrix} I & 0 \\ 0 & \widetilde{P}_2 \end{bmatrix},$$

$$\text{and } \widehat{R} = R_2 P_2 R_1 P_1.$$

If A has full rank, then $\widetilde{R}_2$ is nonsingular. It is easy to show that Q is orthogonal because

$$Q_1^T Q_2 = Q_1^T (A_1 \widetilde{P}_2^T \widetilde{R}_2^{-1}) = (Q_1^T A_1) \widetilde{P}_2^T \widetilde{R}_2^{-1} = 0.$$

As is the case in similar situations related to Gram–Schmidt with pivoting, the final result is equivalent to applying the standard Gram–Schmidt process to a matrix obtained from A by permuting its columns. Indeed, starting with (4.4), we have

$$\begin{aligned} AP_1^T &= [Q_1, Q_2] \begin{bmatrix} I & 0 \\ 0 & \widetilde{R}_2 \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \widetilde{P}_2 \end{bmatrix} \begin{bmatrix} D_1 & F_1 \\ 0 & I \end{bmatrix} \\ &= [Q_1, Q_2] \begin{bmatrix} I & 0 \\ 0 & \widetilde{R}_2 \end{bmatrix} \begin{bmatrix} D_1 & F_1 \\ 0 & \widetilde{P}_2 \end{bmatrix} \\ &= [Q_1, Q_2] \begin{bmatrix} I & 0 \\ 0 & \widetilde{R}_2 \end{bmatrix} \begin{bmatrix} D_1 & F_1 \widetilde{P}_2^T \\ 0 & I \end{bmatrix} P_2, \end{aligned}$$

which yields the following QR factorization of a column-permuted A:

$$AP_1^T P_2^T = [Q_1, Q_2] \begin{bmatrix} D_1 & F_1 \widetilde{P}_2^T \\ 0 & \widetilde{R}_2 \end{bmatrix}. \tag{4.6}$$

Since A is sparse, F_1 and A_1 are usually sparse as well. Sparsity can also be improved by relaxing the orthogonality and applying dropping strategies in the incomplete version, as will be discussed in section 4.2. Moreover, because A_1 is likely to still be large, the above reduction process can be applied to A_1 recursively instead of obtaining its QR factorization with a standard algorithm. The recursion continues until the reduced matrix is small enough or the matrix cannot be further reduced.

Let $A_0 \equiv A$. Then, generally, the factorization at levels $i = 1, 2, \ldots, p$ can be recursively defined as follows:

$$A_{i-1} \widetilde{P}_i^T = [A_{i-1}^{(1)}, A_{i-1}^{(2)}] = [Q_i, A_i] \begin{bmatrix} D_i & F_i \\ 0 & I \end{bmatrix}, \tag{4.7}$$

where $A_{i-1}^{(1)}$ has s_i columns and, similarly to the one-level case, $\widetilde{P}_i^T$ is the column permutation which orders the set of independent columns first. Let

$$D_i = \operatorname{diag}\left(\|A_{i-1}^{(1)} e_j\|_2 \right)_{j=1,\ldots,s_i}, \tag{4.8}$$

$$Q_i = A_{i-1}^{(1)} D_i^{-1}, \tag{4.9}$$

$$F_i = Q_i^T A_{i-1}^{(2)}, \tag{4.10}$$

$$A_i = A_{i-1}^{(2)} - Q_i F_i. \tag{4.11}$$

We will also define as before

$$P_i = \begin{bmatrix} I & 0 \\ 0 & \widetilde{P}_i \end{bmatrix},$$

where the identity block completes the matrix $\widetilde{P}_i$ into an $n \times n$ matrix.

The MQR algorithm can be simply defined as follows.

ALGORITHM 4.1. MQR

0. $A_0 \equiv A$
1. For $i = 1, \ldots, p$
2. Compute permutation $\widetilde{P}_i$ and apply it to A_{i-1} $A_{i-1}\widetilde{P}_i^T = [A_{i-1}^{(1)}, A_{i-1}^{(2)}]$
3. Compute Q_i, D_i, $F_i = Q_i^T A_{i-1}^{(2)}$, and $A_i = A_{i-1}^{(2)} - Q_i F_i$.
4.
5. $A_p\widetilde{P}_{p+1}^T = Q_{p+1}\widetilde{R}_{p+1}$ (standard QR with/without pivoting).

We can now establish a result which generalizes the relation (4.6).

LEMMA 4.1. [illegible]

$$AP_1^T \ldots P_i^T = [Q_1, \ldots, Q_i \mid A_i] \begin{bmatrix} R_{11} & R_{12} \\ 0 & I \end{bmatrix}, \tag{4.12}$$

[illegible] R_{11} [illegible] $[Q_1, \ldots, Q_i]$

[illegible]. The proof is by induction on i. We begin by pointing out that $\widetilde{P}_1 \equiv P_1$. Since $A_0 \equiv A$, (4.7) shows that the result is trivially true for $i = 1$. We now assume that (4.12) is true for i and will show that it is true for $i+1$. From (4.7) we can write

$$A_i = [Q_{i+1}, A_{i+1}] \begin{bmatrix} D_{i+1} & F_{i+1} \\ 0 & I \end{bmatrix} \widetilde{P}_{i+1}$$

which, when substituted in (4.12) yields

$$\begin{aligned}
&[Q_1, \ldots, Q_i, A_i] \begin{bmatrix} R_{11} & R_{12} \\ 0 & I \end{bmatrix} \\
&\quad = \left[Q_1, \ldots, Q_i, [Q_{i+1}, A_{i+1}] \begin{bmatrix} D_{i+1} & F_{i+1} \\ 0 & I \end{bmatrix} \widetilde{P}_{i+1}\right] \begin{bmatrix} R_{11} & R_{12} \\ 0 & I \end{bmatrix} \\
&\quad = [Q_1, \ldots, Q_i, [Q_{i+1}, A_{i+1}]] \begin{bmatrix} R_{11} & & R_{12} \\ 0 & \begin{bmatrix} D_{i+1} & F_{i+1} \\ 0 & I \end{bmatrix} \widetilde{P}_{i+1} \end{bmatrix} \\
&\quad = [Q_1, \ldots, Q_i, Q_{i+1}, A_{i+1}] \begin{bmatrix} R_{11} & R_{12}\widetilde{P}_{i+1}^T \\ 0 & \begin{bmatrix} D_{i+1} & F_{i+1} \\ 0 & I \end{bmatrix} \end{bmatrix} \begin{bmatrix} I & 0 \\ 0\widetilde{P}_{i+1} & \end{bmatrix}.
\end{aligned}$$

This shows that

$$AP_1^T \cdots P_i^T P_{i+1}^T = [Q_1, \ldots, Q_i, Q_{i+1} \mid A_{i+1}] \begin{bmatrix} R_{11} & R_{12}\widetilde{P}_{i+1}^T \\ 0 & \begin{bmatrix} D_{i+1} & F_{i+1} \\ 0 & I \end{bmatrix} \end{bmatrix},$$

which is the desired result for level $i+1$. □

If the procedure stops at the pth level, then A_p is the final reduced system and we factor it as

$$A_p\widetilde{P}_{p+1}^T = Q_{p+1}\widetilde{R}_{p+1}.$$

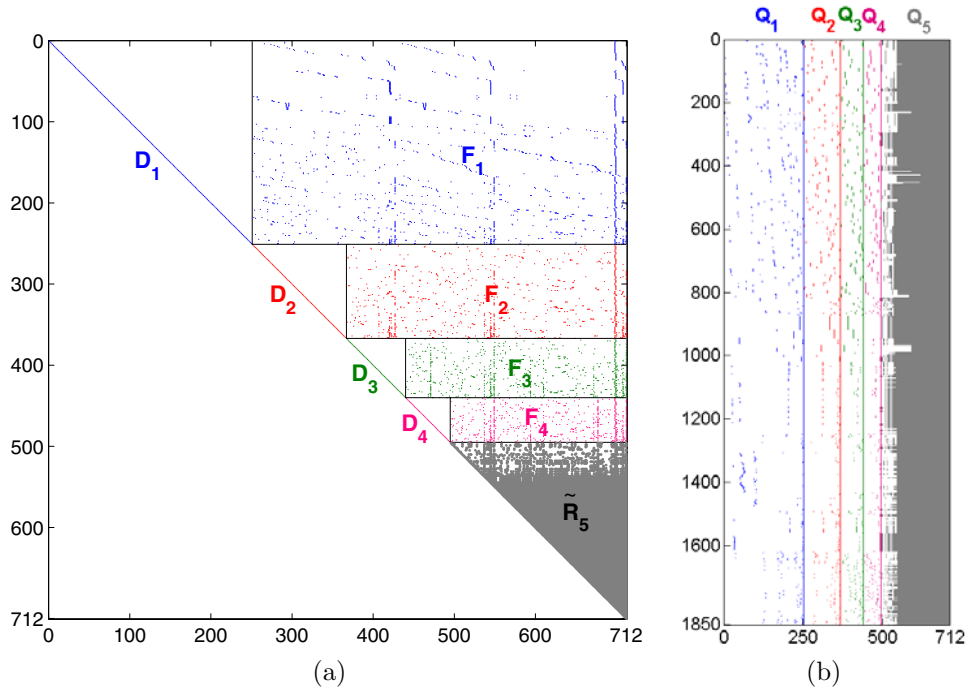


FIG. 4.1. *The MQR structure for matrix* WELL1850 (1,850 × 712, $nnz = 8{,}758$).

Note that $\widetilde{P}_{p+1}^T$ is the identity matrix when pivoting is not used. Then, the above lemma shows that

$$AP_1^T \cdots P_p^T P_{p+1}^T = [Q_1, \ldots, Q_p, Q_{p+1}] \begin{bmatrix} R_{11} & R_{12}\widetilde{P}_{p+1}^T \\ 0 & \widetilde{R}_{p+1} \end{bmatrix}. \tag{4.13}$$

This yields a permuted QR factorization, since it is easily shown that the columns of $[Q_1, Q_2, \ldots, Q_{p+1}]$ are orthonormal.

LEMMA 4.2. [illegible] $\mathcal{P}^T = P_1^T P_2^T \cdots P_{p+1}^T$ [illegible] $\mathcal{Q} = [Q_1, \ldots, Q_p, Q_{p+1}]$ [illegible] $\mathcal{Q}$ [illegible] A.

$$A\mathcal{P}^T = \mathcal{Q}\mathcal{R}, \tag{4.14}$$

[illegible] $\mathcal{R}$ [illegible]. Part of the result is established in (4.13). The only situation when the algorithm will break down is when a diagonal entry in D_i is zero or when the last factorization fails. This is impossible when A is of full rank. It remains only to show that the matrix $[Q_1, \ldots, Q_p, Q_{p+1}]$ in (4.13) is indeed unitary. Within the same block Q_i the columns are orthogonal structurally and are normalized. So $Q_i^T Q_i = I$. For $j > i$ we have $Q_i^T Q_j = 0$. Indeed, the columns of Q_j are linear combinations of columns of A_i because $j > i$. However, by construction A_i is orthogonal to Q_i so we have $Q_i^T Q_j = 0$. □

An illustration of the sizes and the positions of $D_1, F_1, \ldots, D_p, F_p$, and $\widetilde{R}_{p+1}$ can be visualized in Figure 4.1(a), where a four-level QR factorization process ($p = 4$) has been applied to matrix WELL1850. (Some information on this matrix can be found

in section 5.) Figure 4.1(b) shows the corresponding matrix $Q = [Q_1, \ldots, Q_p, Q_{p+1}]$. Note that in order to obtain a better quality picture, the Q and R factors are scaled differently in the figure (the column/row size of R is the same as the column size of Q).

Another formulation of the factorization, which will be used later, is to write

$$A = Q\widehat{R}, \tag{4.15}$$

where $\widehat{R} = R_{p+1}P_{p+1}R_pP_p \cdots R_1P_1 \in \mathbb{R}^{n\times n}$ and $Q = [Q_1, \ldots, Q_p, Q_{p+1}] \in \mathbb{R}^{m\times n}$. Under similar notations as the one-level process, R_i has the form

$$R_i = \begin{bmatrix} I & 0 & 0 \\ 0 & D_i & F_i \\ 0 & 0 & I \end{bmatrix}, \quad i = 1, 2, \ldots, p, \quad \text{and} \quad R_{p+1} = \begin{bmatrix} I & 0 \\ 0 & \widetilde{R}_{p+1} \end{bmatrix}.$$

4.2. Multilevel incomplete QR factorization. If the MQR factorization is complete, we have

$$A^TA = \widehat{R}^T\widehat{R}.$$

Since $\widehat{R}$ is a product of permutation matrices and upper triangular matrices, it is normally inexpensive to solve the equation $(\widehat{R}^T\widehat{R})x = y$. Therefore, an approximation $M \approx \widehat{R}^T\widehat{R}$, obtained through an incomplete multilevel QR factorization process, can be used as a preconditioner for solving (1.2). In the following, a few strategies are considered for developing practical variants of the exact MQR algorithm just described. These will lead to the MIQR preconditioner.

4.2.1. Relaxed independent set of columns. At each level of MQR, we would like to find a larger independent set of columns so that the reduced matrix is smaller. However, there are cases when an independent set with a large size does not even exist. For example, in an extreme case where all entries in one row of a matrix are nonzero, any two column vectors of the matrix are adjacent to each other; i.e., an edge exists between any two vertices in $CIG(A)$. In this case, the largest independent set will consist of only one vertex.

For the purpose of preconditioning, the orthogonality requirement can be somewhat relaxed since only an approximation of the factorization is needed. Therefore, in order to obtain a larger independent set, as well as to reduce fill-in, we will treat two column vectors as being "orthogonal" whenever the acute angle between them is "close" to a right angle. Specifically, for a given small value $\tau_\theta > 0$, an edge from vertex i to vertex j is considered to belong to $CIG(A)$ if $|\cos\theta_{ij}| \geq \tau_\theta$. This replaces the original condition that $\cos\theta_{ij} \neq 0$. The scalar τ_θ is termed the [illegible]. We denote the CIG obtained under the angle threshold τ_θ by $CIG(A, \tau_\theta)$.

Let $C(\tau_\theta)$ be the matrix obtained by replacing all elements less than τ_θ in absolute value in the matrix C defined by (3.1) with 0. Clearly, the graph $CIG(A, \tau_\theta)$ is the adjacency graph of the matrix $C(\tau_\theta)$. Recall that the entries in C (and $C(\tau_\theta)$) are cosines of columns of B, which represent the patterns of the columns of A. Alternatively, the cosines can be calculated using the real values in A instead. To do so, $B = [\frac{a_1}{\|a_1\|}, \frac{a_2}{\|a_2\|}, \ldots, \frac{a_n}{\|a_n\|}]$ is used to calculate C instead of B given by (3.1). In our implementation the cosines were evaluated using the real values in A. Once $CIG(A, \tau_\theta)$ is obtained, Algorithm 3.1 or 3.2 can be applied to $CIG(A, \tau_\theta)$ to find an independent set. The independent set found in this way is in general significantly larger than that found by applying the same algorithm on $CIG(A)$. The effectiveness of this relaxed independent set ordering strategy is illustrated in section 5.

4.2.2. Dropping strategies. The multilevel process yields denser and denser intermediate matrices F_i in general. To ensure a moderate memory usage, we usually drop small terms from F_i. Since a relaxed orthogonality strategy is employed (see previous section), this same strategy is applied when computing the matrix F_i. Recall that $F_i = Q_i^T A_{i-1}^{(2)}$, where Q_i includes normalized columns in the independent set S found at level i, and $A_{i-1}^{(2)}$ includes all remaining columns which are not in S. Therefore, any element in F_i is an inner product between a column vector in S and another column vector not in S. For a given angle threshold τ_θ, the element is replaced by 0 if the cosine of the angle between these two column vectors is less than τ_θ in absolute value. Assume that $F_i = \{f_{uv}\}$, $Q_i = [q_1, q_2, \ldots, q_s]$, and $A_{i-1}^{(2)} = [a_1, a_2, \ldots, a_t]$. Then $f_{uv} = q_u^T a_v = \|a_v\|_2 \cos\theta_{uv}$, where θ_{uv} is the angle between q_u and a_v. Thus, f_{uv} is dropped if $|f_{uv}| < \tau_\theta \|a_v\|_2$.

The final reduced matrix is normally much denser than the original matrix A. If it is small enough (e.g., around 100 columns), a standard QR factorization can be applied. Note that this matrix can now be treated as dense in order to take advantage of effective block computations. Otherwise, an incomplete QR factorization is applied. In this paper, we aim to compare MIQR with the IQR preconditioner outlined in Algorithm 2.1. Therefore, we use the same IQR algorithm on the reduced matrix to ensure that the implementations are as close as possible. In the implementation of IQR, fill-ins are dropped dynamically when the columns of Q and R are being formed based on the magnitude of the columns generated.

4.2.3. MIQR. With the relaxed independent sets of columns and dropping strategies described above, the MIQR algorithm can be described as follows.

ALGORITHM 4.2. Multilevel incomplete QR factorization with angle threshold (MIQR(τ_θ))

1. $k := 0$. $A^{(0)} = A$. $p = maxlev$
2. While $k < p$
3. Construct column intersection graph under angle threshold τ_θ: $CIG(A_k, \tau_\theta)$.
4. Find an independent set permutation $\widetilde{P}_{k+1}$ for $CIG(A_k, \tau_\theta)$.
5. Apply permutation $[A_k^{(1)}, A_k^{(2)}] = A_k \widetilde{P}_{k+1}^T$.
6. Let $D_{k+1} := \mathrm{diag}(||a_1^{(k)}||, \ldots, ||a_s^{(k)}||)$, where $A_k^{(1)} = [a_1^{(k)}, \ldots, a_s^{(k)}]$.
7. Let $Q_{k+1} := A_k^{(1)} D_{k+1}^{-1}$ and $F_{k+1} := Q_{k+1}^T A_k^{(2)}$.
8. Apply a dropping strategy to F_{k+1}.
9. $A_{k+1} := A_k^{(2)} - Q_{k+1} F_{k+1}$
10. Apply a dropping strategy to A_{k+1}.
11. $k := k + 1$
12.
13. Apply IQR (or QR) on A_p: $A_p \widetilde{P}_{p+1}^T \approx Q_{p+1} \widetilde{R}_{p+1}$.

Some implementation details are now discussed. Theoretically, we can continue the multilevel incomplete QR factorization process until the reduced matrix is very small or the system cannot be further reduced. Practically, the overhead of the multilevel process increases substantially as more levels are taken. At the same time, since the multilevel process yields denser and denser matrices, the number of independent columns available becomes much smaller. Therefore, it is best to stop the multilevel process when a certain number of levels (*maxlev* in Algorithm 4.2) is reached or the size of the reduced problem is not significant (e.g., less than 30% of the previous problem size).

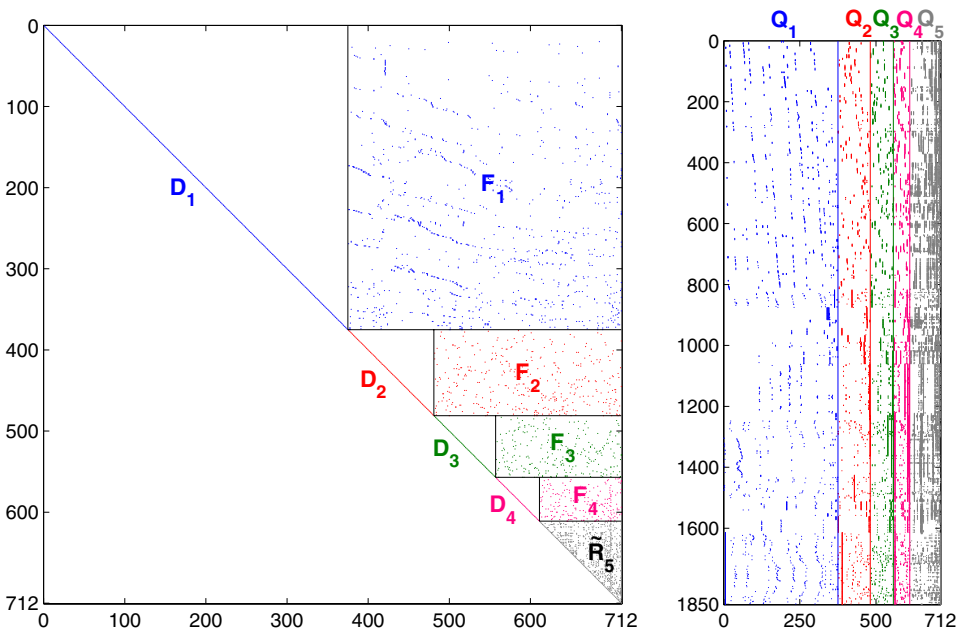


FIG. 4.2. *The MIQR structure for matrix* WELL1850 (1,850 × 712, $nnz = 8{,}758$).

Recall that at each level $CIG(A, \tau_\theta)$ is the adjacency graph of $C(\tau_\theta)$ which is available from the matrix $C = B^T B$; see (3.1). However, the matrix C need not be calculated explicitly. Since we determine the degrees of the vertices one by one, only a single row of C is needed at any given time. In other words, to determine the degree of vertex i, only the ith row of C needs to be calculated. This row contains the inner products between the ith column of B and all other columns. As indicated in [30] (see also section 2), these inner products can be efficiently calculated as a linear combination of the rows of B. For this reason, although a reduced matrix is naturally formed and stored columnwise during the MQR factorization process, we maintain an index array for easily accessing its elements rowwise. Furthermore, since C is symmetric, only its upper part (i.e., the inner products between the ith column of B and columns from $i+1$ to n) needs to be calculated.

As is standard practice, the permutation matrices $\widetilde{P}_k$ are not formed explicitly. Instead, a permutation array $perm^{(k)}$ is employed to hold the new ordering of the columns at each level, along with the inverse permutation array $iperm^{(k)}$. With this strategy, the columns of matrix A_k are kept in their original ordering and the permutation step (line 5 in Algorithm 4.2) can be avoided. To construct an MIQR preconditioner, the matrix array $\{D_1, F_1, D_2, F_2, \ldots, D_p, F_p, \widetilde{R}_{p+1}\}$ and the permutation arrays $perm^{(k)}$ and $iperm^{(k)}$ are stored. Similarly to Figure 4.1, this matrix array can be organized in an $n \times n$ matrix as illustrated in Figure 4.2, where $\tau_\theta = 0.1$ is used and IQR is applied to the final reduced matrix. Moreover, since the matrices Q_k are not needed for the preconditioning purpose, they are discarded at the end of the kth level recursion, respectively.

Other than the dropping strategies discussed in section 4.2.2, an optional dropping rule may be applied to A_{k+1} as well (see line 10 of Algorithm 4.2). This is used mainly

as a means of preventing A_{k+1} from becoming too dense. For example, when needed, any nonzero entry in A_{k+1} whose absolute value is less than a threshold τ times a certain norm of the column under consideration will be dropped. However, dropping in A_{k+1} should be applied sparingly as we observed in our experiments that this can negatively impact the robustness of the resulting preconditioner. Using a small value of the threshold value τ is recommended if dropping is to be done on A_{k+1}, as dropping very small terms is less harmful. For the test problems used in this paper, all the intermediate matrices generated at line 10 in Algorithm 4.2 were reasonably sparse. Therefore, for the sake of maintaining robustness, no dropping was applied to A_{k+1} in the tests reported in section 5, (i.e., we use $\tau = 0$).

As mentioned before, the matrix $\widehat{R}$ defined as in (4.15) is a product of permutation matrices and upper triangular matrices. To precondition the normal equations, we need to solve the systems $\widehat{R}x = y$ and $\widehat{R}^T x = y$. Let $s_k (k = 1, 2, \ldots, p)$ be the size of the independent set of columns at level k. Define $r_1 = 1$ and $r_k = r_{k-1} + s_{k-1}$ for $k = 2, \ldots, p+1$. Algorithms 4.3 and 4.4 are used to solve the systems $\widehat{R}x = y$ and $\widehat{R}^T x = y$, respectively.

ALGORITHM 4.3. Solving $\widehat{R}x = y$

1. $x(1:n) := y(1:n)$
2. For $k = 1 : p$
3. Apply permutation $perm^{(k)}$ to $x(r_k : n)$
4. $x(r_k : r_{k+1} - 1) := D_k^{-1} x(r_k : r_{k+1} - 1)$
5. $x(r_{k+1} : n) := x(r_{k+1} : n) - F_k^T x(r_k : r_{k+1} - 1)$
6.
7. Solve $\widetilde{R}_{p+1} z = x(r_{p+1} : n)$ for z.
8. $x(r_{p+1} : n) := z$

ALGORITHM 4.4. Solving $\widehat{R}^T x = y$

1. $x(1:n) := y(1:n)$
2. Solve $\widetilde{R}_{p+1} z = y(r_{p+1} : n)$ for z.
3. $x(r_{p+1} : n) := z$
4. For $k = p : -1 : 1$
5. $x(r_k : r_{k+1} - 1) := D_k^{-1}[x(r_k : r_{k+1} - 1) - F_k x(r_{k+1} : n)]$
6. Apply permutation $iperm^{(k)}$ to $x(r_k : n)$.
7.

4.3. Analysis. In this section we will analyze the errors generated by the MIQR factorization that are due to dropping small terms. Two questions are important to consider. The first is, How far does the result deviate from satisfying the relation $A\mathcal{P}^T = \mathcal{Q}\mathcal{R}$? The second is, How much does $\mathcal{Q}$ deviate from orthogonality in the presence of dropping?

The basic step of MIQR can be described by approximate versions of the relations (4.7)–(4.11). Specifically, the equations that define F_i and A_i will change while Q_i, D_i can be assumed to be exact:

$$D_i = \text{diag}\left(\|A_{i-1}^{(1)} e_j\|_2\right)_{j=1,\ldots,s_i}, \tag{4.16}$$

$$Q_i = A_{i-1}^{(1)} D_i^{-1}, \tag{4.17}$$

$$\widetilde{F}_i = Q_i^T A_{i-1}^{(2)} + E_{F,i}, \tag{4.18}$$

$$\widetilde{A}_i = A_{i-1}^{(2)} - Q_i \widetilde{F}_i + E_i. \tag{4.19}$$

The term E_i comes from dropping entries when forming the block A_i while the term $E_{F,i}$ comes from dropping when forming the matrix F_i; see Algorithm 4.2. Then assuming A_{i-1} is exact, the relation (4.7) becomes

$$A_{i-1}\widetilde{P}_i^T = [A_{i-1}^{(1)}, A_{i-1}^{(2)}] \ = [Q_i, \widetilde{A}_i] \begin{bmatrix} D_i & \widetilde{F}_i \\ 0 & I \end{bmatrix} + [0, E_i]. \tag{4.20}$$

Note the remarkable absence of $E_{F,i}$ from the error in (4.20). The error is imbedded in $\widetilde{F}_i$. Before continuing, we can further note that if there is no dropping when building A_i, then $E_i \equiv 0$ and the relation (4.7) becomes exact. This means that in the end we would expect to have the relation $A\mathcal{P}^T = \mathcal{Q}\mathcal{R}$ exactly satisfied, but $\mathcal{Q}$ is not necessarily unitary. Now we consider the general case and will attempt to analyze the difference between $A\mathcal{P}^T$ and $\mathcal{Q}\mathcal{R}$. This case is important to consider because Algorithm 4.14 may include dropping in A_{k+1} after it is constructed; see line 10 of the algorithm. (Although we did not apply this dropping in our tests reported in this paper, it helps to keep A_{k+1} sparse in general.)

Consider the result of Lemma (4.1) which would be desirable to generalize. In the following we attempt to extend the argument used in the proof of Lemma 4.1. We write the above relation for level $i+1$ as

$$\widetilde{A}_i = \Big[Q_{i+1}, \widetilde{A}_{i+1}\Big] \underbrace{\begin{bmatrix} D_{i+1} & \widetilde{F}_{i+1} \\ 0 & I \end{bmatrix}}_{Z_{i+1}} \widetilde{P}_{i+1} + \underbrace{[0, E_{i+1}]}_{G_{i+1}} \widetilde{P}_{i+1}.$$

Substituting in (4.12) (where A_i is replaced by the computed $\widetilde{A}_i$ at the ith step) yields

$$\begin{aligned}
&\Big[Q_1, \ldots, Q_i, \widetilde{A}_i\Big] \begin{bmatrix} R_{11} & R_{12} \\ 0 & I \end{bmatrix} \\
&\quad = \Big[Q_1, \ldots, Q_i \mid [Q_{i+1}, \widetilde{A}_{i+1}]Z_{i+1}\widetilde{P}_{i+1} + G_{i+1}\widetilde{P}_{i+1}\Big] \left[\begin{array}{c|c} R_{11} & R_{12} \\ \hline 0 & I \end{array}\right] \\
&\quad = \Big[Q_1, \ldots, Q_i \mid [Q_{i+1}, \widetilde{A}_{i+1}]\Big] \left[\begin{array}{c|c} R_{11} & R_{12} \\ \hline 0 & Z_{i+1}\widetilde{P}_{i+1} \end{array}\right] + [0, G_{i+1}\widetilde{P}_{i+1}] \\
&\quad = \Big[Q_1, \ldots, Q_i, Q_{i+1}, \widetilde{A}_{i+1}\Big] \begin{bmatrix} R_{11} & R_{12}\widetilde{P}_{i+1}^T \\ 0 & Z_{i+1} \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \widetilde{P}_{i+1} \end{bmatrix} \\
&\qquad + [0, G_{i+1}] \begin{bmatrix} I & 0 \\ 0 & \widetilde{P}_{i+1} \end{bmatrix}.
\end{aligned}$$

In what follows we will denote by $\mathcal{R}_i$ the matrix $\left[\begin{smallmatrix} R_{11} & R_{12} \\ 0 & I \end{smallmatrix}\right]$ at step i. So the above relation translates into

$$\Big[Q_1, \ldots, Q_i, \widetilde{A}_i\Big] \mathcal{R}_i = \Big[Q_1, \ldots, Q_i, Q_{i+1}, \widetilde{A}_{i+1}\Big] \mathcal{R}_{i+1}P_{i+1} + [0, G_{i+1}]P_{i+1}. \tag{4.21}$$

The left-hand side of the above relation is equal to $AP_1^T \cdots P_i^T + \mathcal{E}_i$, where $\mathcal{E}_i$ denotes the total error made at step i of the factorization. Using a similar notation for the first term of the right-hand side will transform (4.21) into

$$AP_1^T \cdots P_i^T + \mathcal{E}_i = (AP_1^T \cdots P_iP_{i+1}^T + \mathcal{E}_{i+1})P_{i+1} \ + \ [0, G_{i+1}]P_{i+1}. \tag{4.22}$$

This means that $\mathcal{E}_i = \mathcal{E}_{i+1}P_{i+1} + [0, G_{i+1}]P_{i+1}$, and it establishes the remarkably simple recurrence relation for the total error

$$\mathcal{E}_{i+1} = \mathcal{E}_i P_{i+1}^T - [0, E_{i+1}], \tag{4.23}$$

where the zero block in the right-hand side has the same number of columns as $[Q_1, Q_2, \ldots, Q_i]$. In particular we have

$$\|\mathcal{E}_{p+1}\| \leq \sum_{i=1}^{p+1} \|E_{i+1}\|.$$

However, this inequality does not say everything about the errors. For example, it is clear that the last columns (after permutation) will undergo more perturbations than the first ones and they will therefore be less accurate. This is understandable since, for example, the columns of the first level are not perturbed by the other columns.

Consider now the accuracy of the process with respect to orthogonality. For simplicity, we will consider only the situation where there is no dropping in forming $\widetilde{A}_i$; i.e., the case where $E_i = 0$. Furthermore, we assume that each Q_i, considered individually, is exactly orthonormal; i.e., $Q_i^T Q_i = I$. In particular this means that τ_θ should be equal to zero. In this case, it is easy to see from (4.23) that the relation $A\mathcal{P}^T = \mathcal{Q}\mathcal{R}$ is exactly satisfied. However, dropping entries in F_i will cause loss of orthogonality.

Consider only one step of the process. From (4.17), (4.18), and (4.19) (with $E_i = 0$) we obtain

$$Q_i^T \widetilde{A}_i = Q_i^T (A_{i-1}^{(2)} - Q_i \widetilde{F}_i) = Q_i^T A_{i-1}^{(2)} - (Q_i^T A_{i-1}^{(2)} + E_{F,i}) = -E_{F,i}.$$

Next we wish to establish a relation between this term and $Q_i^T \mathcal{Q}$. Specifically, because of the recursive nature of the algorithm, at step i we have a relation similar to that given by Lemma 4.2. A little additional notation is needed. Let $\mathcal{Q}_i = [Q_i, \ldots, Q_{p+1}]$ and $\mathcal{P}_i^T = P_i^T \cdots P_{p+1}^T$. Then,

$$\widetilde{A}_i \mathcal{P}_i = \mathcal{Q}_i \mathcal{R}_i,$$

where $\mathcal{R}_i$ is an upper triangular matrix. Multiplying to the left by Q_i^T yields

$$Q_i^T \widetilde{A}_i \mathcal{P}_i = Q_i^T \mathcal{Q}_i \mathcal{R}_i$$

so that

$$-E_{F,i} \mathcal{P}_i \mathcal{R}_i^{-1} = Q_i^T [Q_i, \ldots, Q_{p+1}].$$

The above relation shows in a simple way how the error made at step i will propagate to the matrices $Q_i^T Q_j$ for $j > i$. The result involves the inverse of an unknown triangular matrix. The matrix $\mathcal{R}_i$ establishes the relation between the Q_j's and the matrix $\widetilde{A}_i$. For example, for $i = 1$ we would get all the matrices $Q_1^T Q_j$ for $j > 1$ in terms of $E_{F,1}$, the error related to dropping in the F matrix in the first step.

An interesting and important question which we do not address in this paper is the issue of effective dropping. In the papers [8, 9], an idea was considered for dropping in such a way that the preconditioned matrix is close to the identity. This is in contrast with other methods which try to make the preconditioner close to A. Though this idea involves the inverse of the preconditioner, heuristics can be used to provide quite effective methods. In the context of IQR, this may be doable but it will undoubtedly be more complex.

TABLE 5.1
Information on the test problems: Size = $m \times n$; nnz = the number of nonzeros; μ = the average number of nonzeros per row; ν = the average number of nonzeros per column.

Matrix	m	n	nnz	μ	ν	Source
ILLC1850	1,850	712	8,758	4.73	12.30	Surveying
WELL1850	1,850	712	8,758	4.73	12.30	Surveying
MESHPAR1	31,258	15,994	187,498	6.00	11.72	3D mesh parameterization
MESHPAR2	75,650	38,384	453,846	6.00	11.82	3D mesh parameterization
SMALL2	6,280	3,976	25,530	4.07	6.42	Animal breeding
MEDIUM2	18,794	12,238	75,039	3.99	6.13	Animal breeding
LARGE	28,254	17,264	75,018	2.66	4.35	Animal breeding
LARGE2	56,508	34,528	225,054	3.98	6.52	Animal breeding
VERYL	174,193	105,882	463,303	2.66	4.38	Animal breeding
VERYL2	348,386	211,764	1,389,909	3.99	6.56	Animal breeding

TABLE 5.2
Independent set sizes: Comparison of the lower bounds (3.7)*, the estimated numbers* (3.5)*, the actual sizes, and the rough lower bounds* (3.2)*.*

Matrix	New low. s	Est. val. s	Real val. s	Old low. s
ILLC1850	59	196	220	16
WELL1850	59	196	237	16
MESHPAR1	1,116	3,594	2,191	269
MESHPAR2	2,660	8,613	5,090	639
SMALL2	612	1,438	1,171	193
MEDIUM2	1,976	4,544	3,297	633
LARGE	5,048	9,648	9,690	2108
LARGE2	5,359	12,681	9,957	1,690
VERYL	30,788	59,020	60,454	12,815
VERYL2	32,661	77,552	61,638	10,269

5. Numerical results. In this section, we test the performance of the MIQR method on ten least-squares problems from real applications. Table 5.1 provides some basic information about the test matrices. In the table, m is the number of rows, n is the number of columns, nnz is the total number of nonzeros, μ is the average number of nonzeros per row, and ν is the average number of nonzeros per column. Matrices ILLC1850 and WELL1850 are available from the Matrix Market.[1] The next two matrices are from a three-dimensional mesh parameterization problem.[2] These matrices are generated using the method of least-squares conformal maps as described in [23]. The last six matrices (SMALL2, MEDIUM2, LARGE, LARGE2, VERYL, VERYL2) arise in animal breeding studies [16, 17] and can be downloaded from ftp://ftp.cerfacs.fr/pub/algo/matrices/animal. The matrices SMALL2, MEDIUM2, LARGE2, and VERYL2 are generated from the code `conv2.f` while LARGE and VERYL are generated using `conv.f`. Both `conv.f` and `conv2.f` are included in the package available from the website.[3]

We first test the accuracy of the estimate on the number of independent sets of columns as described in section 3.2. Table 5.2 shows the lower bounds (in field "New low. s") estimated by (3.7) and the values estimated by solving (3.5) numerically (in

[1] http://math.nist.gov/MatrixMarket/.

[2] This problem was provided to us by Minh Nguyen from the Graphics group at the University of Minnesota, Department of Computer Science and Engineering.

[3] The difference between `conv.f` and `conv2.f` is explained in READ_ME as follows: "The code conv.f constructs problems where only the single trait 'weight increase' is considered. The code `conv2.f` deals with both traits. The matrices constructed using `conv.f` are smaller than those constructed using conv2.f." For more details see the references cited above.

TABLE 5.3
Numbers of independent columns found using different angle tolerances for the first two levels.

Matrix	Level	$\tau_\theta = 0.00$	$\tau_\theta = 0.05$	$\tau_\theta = 0.10$	$\tau_\theta = 0.15$	$\tau_\theta = 0.20$
ILLC1850	1	220	327	338	343	350
	2	130	146	152	159	170
WELL1850	1	237	346	372	395	432
	2	122	101	115	119	134
MESHPAR1	1	2,191	6,151	6,594	6,471	6,681
	2	2,191	2,483	3,397	4,213	5,411
MESHPAR2	1	5,090	15,685	16,829	15,818	15,655
	2	5,090	5,506	7,973	9,823	13,918
SMALL2	1	1,171	1,623	1,995	2,793	2,884
	2	1,147	1,022	1,087	683	684
MEDIUM2	1	3,297	4,836	6,154	8,092	8,308
	2	3,110	3,274	2,941	2,273	2,340
LARGE	1	9,690	10,280	10,222	10,545	12,554
	2	3,794	4,113	4,563	4,509	3,320
LARGE2	1	9,957	14,681	18,523	24,994	25,724
	2	9,312	9,891	8,552	6,038	6,131
VERYL	1	60,454	61,032	64,107	66,579	77,343
	2	22,095	23,495	25,781	25,000	20,203
VERYL2	1	61,638	89,500	114,752	153,992	157,109
	2	57,924	58,804	51,687	36,741	37,615

field "Est. val. s") for the ten matrices. These lower bounds and estimated values are compared with the real values calculated by Algorithm 3.2 (in field "Real val. s"). For reference, we also calculate the rough lower bounds estimated by (3.2) (in field "Old low. s"), where the average degree η (3.4) is used as the value of d_S. It is clear that the values calculated using (3.7) provide much closer lower bounds. Recall that (3.5) and (3.7) are derived under the assumption that the nonzero elements of a matrix are randomly distributed. In spite of this assumption, the estimated values can still provide good approximations for the matrices from real applications.

Table 5.3 presents the results of finding the independent columns using different angle tolerances τ_θ as described in section 4.2.1. In the table, $\tau_\theta = 0.00$, 0.05, 0.10, 0.15 and 0.20 (corresponding to angles 90°, 87.13°, 84.26°, 81.37°, and 78.46°, respectively) are tested. Algorithm 3.2 is used for all tests. For each τ_θ, we list the number of independent columns found in the first two levels. From the table, as expected, the number of independent columns found in each level is significantly increased as the angle tolerance increases. As an example, for matrix VERYL2, without relaxing the criterion of finding independent columns (i.e., $\tau_\theta = 0.00$), only 61,638 and 57,924 independent columns are found in the first two levels of reduction respectively; i.e., the problem size reduced after the first two levels is 119,562. With the angle tolerances, the problem sizes reduced after the first two levels increase to 148,304, 166,439, 190,733, and 194,724, respectively, for $\tau_\theta = 0.05$, 0.10, 0.15, and 0.20.

Next, we test MIQR on the ten least-squares problems and compare the results with IQR (Algorithm 2.1), RIF [4] preconditioners, and ARMS (on the normal equations). MIQR, IQR, and ARMS were coded in C. RIF provided by Benzi and Tůma was in FORTRAN90.[4] All codes were compiled in 64-bit mode with the -O2 opti-

[4]The RIF package (rifsrnri.tar.gz) is also available at http://www.cs.cas.cz/~tuma/sparslab.html. We only changed the two drop tolerances in the `rifsrnri` source code, one for the SAINV process (to approximate $A^{-1}b$) and the other for the postfiltration of RIF, to achieve the desired fill-in factors for comparison purposes in our tests. All other parameters in the code were left unchanged.

TABLE 5.4
Performance of MIQR under different angle tolerances ($\tau_\theta = 0.000$ and $\tau_\theta = 0.015$).

Matrix	τ_θ	Levels	Res.#	Fill-in	Pre.T	ITS	Its.T	Tot.T
ILLC1850	0.00	2	360	0.668	0.04	290	0.35	0.39
	0.10	4	77	0.328	0.04	172	0.18	0.22
	0.20	4	34	0.219	0.02	183	0.17	0.19
WELL1850	0.00	2	353	0.482	0.04	85	0.10	0.14
	0.10	5	60	0.322	0.06	68	0.06	0.12
	0.20	5	10	0.224	0.02	133	0.12	0.14
MESHPAR1	0.00	2	11,612	0.763	7.25	460	14.27	21.52
	0.04	2	7,536	0.567	5.52	405	10.81	16.33
	0.08	2	6,424	0.513	4.60	530	13.59	18.19
MESHPAR2	0.00	3	25,458	1.097	34.89	731	71.13	106.02
	0.05	3	12,672	0.650	22.02	800	60.84	82.86
	0.10	3	8,938	0.495	13.00	1,357	91.39	104.39
SMALL2	0.00	3	1,299	1.073	0.21	247	1.19	1.40
	0.10	3	506	0.530	0.15	241	0.93	1.08
	0.20	3	132	0.334	0.08	284	0.89	0.97
MEDIUM2	0.00	3	4,724	1.299	0.88	223	3.89	4.77
	0.10	3	1,756	0.628	0.55	235	3.03	3.58
	0.25	3	254	0.304	0.16	407	4.09	4.25
LARGE	0.00	3	2,134	1.247	0.72	44	0.86	1.58
	0.05	3	1,446	1.014	0.52	69	1.21	1.73
	0.10	4	594	0.867	2.22	128	2.12	4.34
LARGE2	0.00	3	12,087	1.234	3.13	361	20.62	23.75
	0.10	3	3,769	0.515	2.25	442	18.29	20.54
	0.20	3	292	0.251	0.37	461	15.47	15.84
VERYL	0.00	4	8,176	1.120	6.88	118	17.11	23.99
	0.18	4	613	0.512	2.19	221	25.59	27.78
	0.20	4	259	0.447	1.56	196	21.51	23.07
VERYL2	0.00	1	150,126	1.430	40.66	916	401.60	442.26
	0.10	4	11,586	0.502	26.50	497	157.71	184.21
	0.25	3	1,522	0.256	3.22	737	186.02	189.24

mization option. All experiments were performed on an IBM SP machine, which has four 222MHz processors sharing 16GB memory. Note that we did not take advantage of parallelism in our tests; i.e., only one of the four processors was used. The right-hand sides available from the original data were used. This is in contrast to [4] where artificial right-hand sides were employed. Algorithm 1.1 with a zero initial guess was used to solve all the problems. The iterations were stopped when

$$\|A^T b - A^T A x^{(k)}\|_2 < 10^{-8} \|A^T b - A^T A x^{(0)}\|_2$$

or the maximum iteration count of 2,000 was reached. To better compare the preconditioners, we use an indicator called a fill-in factor to indicate the memory usage for each method. The fill-in factor is defined as the ratio between the memory used in a preconditioner and the memory used in the original matrix. The memory used in MIQR is represented by the total number of nonzero entries in matrices D_1, $F_1, \ldots, D_p$, F_p, $\tilde{R}_{p+1}$ as shown in Figure 4.2. We wish to compare the preconditioners under similar fill-in factors.

In Table 5.4, we test MIQR on the ten matrices under different angle tolerances τ_θ. In the table, "Levels" is the number of levels used, "Res.#" is the number of columns of the final reduced matrix, "Fill-in" is the fill-in factor, "Pre.T" is the preconditioning time in seconds, "ITS" is the number of iterations for CGLS to reach convergence, "Its.T" is the iteration time in seconds, and "Tot.T" is the total time in

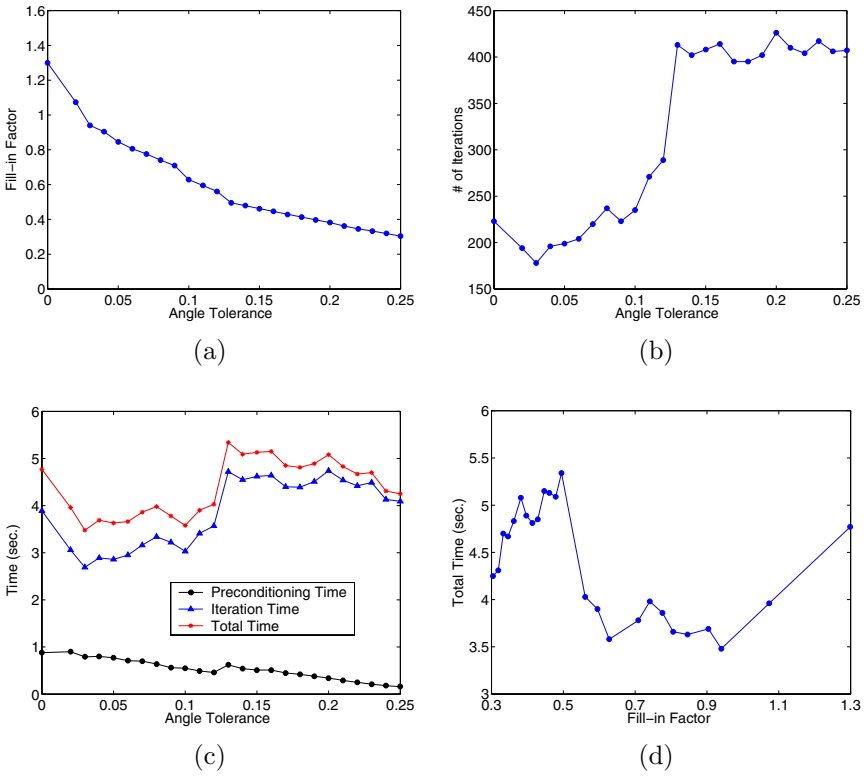


FIG. 5.1. *MIQR performance on matrix* MEDIUM2*:* (a) *Fill-in factor versus angle tolerance.* (b) *Total number of iterations versus angle tolerance.* (c) *Preconditioning time, iteration time, and total time versus angle tolerance.* (d) *Total Time versus fill-in factor.*

seconds. According to the table, and as expected, the size of the final reduced system decreases as the angle tolerance increases. For example, after the same number of reduction levels, the reduced system sizes for matrix MEDIUM2 are 4,724, 1,756, and 254 under angle tolerances 0.00, 0.10, and 0.25, respectively. As a result, the memory usage of MIQR decreases correspondingly. It is also apparent that allowing more fill-ins for the MIQR preconditioner does not necessarily provide faster convergence rates. For matrix ILLC1850, as an example, it takes CGLS 290 iterations to converge under a fill-in factor of 0.668 when $\tau_\theta = 0.00$. However, it takes only 172 iterations under a smaller fill-in factor of 0.328 when $\tau_\theta = 0.10$. This is because the accuracy of the MIQR preconditioner is determined not only by the fill-ins allowed but by many other factors. For example, it is not known how dropping affects the orthogonality of the underlying $\mathcal{Q}$ factor. Recall that $A^TA - M$ is not as important as $I - M^{-1}A^TA$ when it comes to analyzing convergence. From the table, we also observe that only one level is applied to VERYL2 when $\tau_\theta = 0.0$. In this case, it is due to the fact that the reduced size in the first level (61,638 according to Table 5.3) is only a small portion of the original problem size (211,764), which is less than the threshold we set to stop the multilevel process (see section 4.2.3).

We further examine the relationships among the angle tolerance, the fill-in factor, the number of iterations, and the execution times for matrix MEDIUM2 in Figure 5.1. Figure 5.1(a) shows the fill-in factors of the MIQR preconditioner as a function of the angle tolerance τ_θ. Figure 5.1(b) shows the number of iterations required

TABLE 5.5
A comparison of MIQR, IQR, and RIF.

Matrix	Method	Fill-in	Pre.T	ITS	Its.T	Tot.T
ILLC1850	MIQR	0.328	0.04	172	0.18	0.22
	IQR	0.332	0.02	1,512	1.45	1.47
	RIF	0.332	0.19	406	0.28	0.47
WELL1850	MIQR	0.322	0.06	68	0.06	0.12
	IQR	0.333	0.01	439	0.43	0.44
	RIF	0.324	0.21	90	0.07	0.28
MESHPAR1	MIQR	0.567	5.52	405	10.81	16.33
	IQR	0.582	2.06	666	17.31	19.37
	RIF	0.586	3.23	700	12.44	15.67
MESHPAR2	MIQR	0.650	22.02	800	60.84	82.86
	IQR	0.650	5.36	1,462	96.27	101.63
	RIF	0.642	12.30	1,567	74.14	86.44
SMALL2	MIQR	0.334	0.08	284	0.89	0.97
	IQR	0.330	0.03	621	2.30	2.33
	RIF	0.321	6.13	285	0.75	6.88
MEDIUM2	MIQR	0.304	0.16	407	4.09	4.25
	IQR	0.334	0.08	1,315	15.93	16.01
	RIF	0.326	22.49	526	4.50	26.99
LARGE	MIQR	1.014	0.53	69	1.23	1.76
	IQR	1.013	0.43	157	3.02	3.45
	RIF	1.014	2.44	59	0.76	3.20
LARGE2	MIQR	0.251	0.37	461	15.47	15.84
	IQR	0.319	0.22	-	-	-
	RIF	0.266	88.86	634	17.27	106.13
VERYL	MIQR	0.512	2.19	221	25.59	27.78
	IQR	0.515	0.70	997	117.28	117.98
	RIF	0.538	4.44	188	15.56	20.00
VERYL2	MIQR	0.256	3.22	737	186.02	189.24
	IQR	1.090	7.38	-	-	-
	RIF	0.288	1,306.88	1,220	226.70	1,533.58

for CGLS to converge as a function of the angle tolerance τ_θ. Figure 5.1(c) shows the preconditioning time, the iteration time, and the total time as a function of the angle tolerance τ_θ. Finally, Figure 5.1(d) shows the total time used to solve problem MEDIUM2 as a function of the fill-in factor of the MIQR preconditioner.

Table 5.5 compares MIQR with IQR and RIF. The symbol "-" in the table indicates that convergence was not obtained in 2,000 iterations. Note that matrices ILLC1850, WELL1850, LARGE, and VERYL have also been tested in [4] but with using artificial right-hand sides. In order to obtain comparable runs, we have put much effort into adjusting parameters in order to obtain similar fill factors for the same matrix. One should view the comparisons in this section with the following interpretation: How do the techniques compare for the same matrix if roughly the same fill-in is allowed for each method? Recall that MIQR implicitly uses a reordering of the columns and this is in fact the essence of the method. The other methods tested do not use any reordering of the columns. The superior performance of MIQR clearly shows that reordering can be vital. This observation is in agreement with [29], where it is shown that an IQR preconditioner can be made much more effective by applying a permutation prior to computing it. The table indicates that the set-up times for MIQR are slightly higher than those of IQR under similar memory usage, but that both its robustness and total execution times are significantly better. For matrices LARGE2 and VERYL2, IQR failed to converge in 2,000 iterations, even when much more fill-in was allowed. We also observe that MIQR had better overall performances

TABLE 5.6
A comparison of $nnz(A)$ and $nnz(A^TA)$.

Matrix	$nnz(A)$	$nnz(A^TA)$
ILLC1850	8,758	9,126
WELL1850	8,758	9,126
MESHPAR1	187,498	220,916
MESHPAR2	453,846	532,816
SMALL2	25,530	58,848
MEDIUM2	75,039	177,066
LARGE	75,018	128,798
LARGE2	225,054	524,036
VERYL	463,303	782,772
VERYL2	1,389,909	3,184,684

TABLE 5.7
A comparison of MIQR and ARMS (on the normal equations).

Matrix	Method	Levels	Fill-in	Pre.T	ITS	Its.T	Tot.T
ILLC1850	MIQR	4	0.328	0.04	172	0.18	0.22
	ARMS	3	1.329	0.05	94	0.20	0.25
WELL1850	MIQR	5	0.322	0.06	68	0.06	0.12
	ARMS	3	1.120	0.05	70	0.14	0.19
MESHPAR1	MIQR	2	0.567	5.52	405	10.81	16.33
	ARMS	2	1.401	2.78	-	-	-
MESHPAR2	MIQR	3	0.650	22.02	800	60.84	82.86
	ARMS	3	1.262	6.97	-	-	-
SMALL2	MIQR	3	0.334	0.08	284	0.89	0.97
	ARMS	3	1.216	0.32	261	3.92	4.24
MEDIUM2	MIQR	3	0.304	0.16	407	4.09	4.25
	ARMS	4	1.281	1.20	182	9.20	10.40
LARGE	MIQR	3	1.014	0.53	69	1.23	1.76
	ARMS	3	1.010	1.66	427	24.90	26.56
LARGE2	MIQR	3	0.251	0.37	461	15.47	15.84
	ARMS	3	1.356	5.30	-	-	-
VERYL	MIQR	4	0.512	2.19	221	25.59	27.78
	ARMS	4	1.025	37.10	55	21.80	58.90
VERYL2	MIQR	3	0.256	3.22	737	186.02	189.24
	ARMS	4	1.200	73.20	-	-	-

than RIF in general. For most matrices, MIQR required fewer iterations to converge than RIF. This is especially true for matrices MESHPAR1, MESHPAR2, and VERYL2, for which MIQR required significantly fewer iterations under similar memory costs. We caution, however, that comparisons are always difficult with iterative methods because of the large number of parameters that can be adjusted.

Finally, we compared MIQR with ARMS applied to the corresponding normal equations. The results are shown in Table 5.7. The symbol "-" in the table indicates failure to converge in 2,000 iterations or less. GMRES(100) was used as the accelerator for ARMS. In contrast to MIQR, the fill-in factors for ARMS are calculated based on the number of nonzeros of the normal equations (see Table 5.6). This makes it less relevant to compare the MIQR and ARMS-GMRES techniques from the point of view of memory usage (fill factors have a different meaning). Therefore, we choose the best results (in terms of overall time and memory usage) for ARMS in our tests. The results in Table 5.7 clearly show that solving the least-squares problems with MIQR is more effective than solving the corresponding normal equations by ARMS.

6. Conclusion. We have presented a preconditioning technique for solving large sparse least-squares systems that is based on a MIQR factorization. The algorithm exploits a divide-and-conquer strategy which takes advantage of structurally orthogonal columns. This allows us to gradually reduce a large problem to a significantly smaller one with little computational effort. The algorithm first finds an independent set of columns, which are structurally orthogonal. The remaining columns are then orthogonalized against this first set of columns, and the resulting set is orthogonalized recursively. In order to increase the size of the independent sets of columns, we proposed a strategy which consists of relaxing the orthogonality requirement. Numerical results have shown that this strategy is quite effective in finding independent column sets with large cardinality. The MIQR preconditioner has been tested and compared with a standard IQR factorization and with the RIF. The numerical tests show that MIQR is robust and efficient. We have not implemented a parallel version of the algorithm. However, the method has been designed with parallelism in mind and a parallel implementation should scale well.

In section 5, we have observed that the performances of MIQR may be very different when the angle tolerance varies. It remains to investigate a systematic way of selecting a good angle tolerance for a given problem. As mentioned in section 4.2.2, the current dropping strategies at all levels use simple techniques which tend to yield a factorization that is accurate, i.e., such that $A\mathcal{P}^T - \mathcal{Q}\mathcal{R}$ is small. As is the case for ILU factorizations, this is not necessarily a good strategy [8]. It may be possible to adapt Bollhöfer's work [8] to this context and develop more sophisticated dropping strategies which will in all likelihood improve the robustness of the scheme. Finally, as a further improvement, we would like to investigate more sophisticated IQR methods for the final reduced matrix, including using some effective reordering techniques as discussed in [29].

Acknowledgments. We would like to thank Michele Benzi and Miroslav Tůma, who provided the RIF source code used in section 5. We also thank Minh Nguyen for supplying the test matrices MESHPAR1 and MESHPAR2, and Sui Ruan for the valuable discussion on calculating the expected size of independent sets. We are grateful to three anonymous referees whose numerous suggestions helped improve the quality of this paper.

REFERENCES

[1] O. Axelsson and P. Vassilevski, *Algebraic multilevel preconditioning methods.* I. Numer. Math., 56 (1989), pp. 157–177.

[2] O. Axelsson and P. S. Vassilevski, *Algebraic multilevel preconditioning methods,* II, SIAM J. Numer. Anal., 27 (1990), pp. 1569–1590.

[3] R. Bank and C. Wagner, *Multilevel ILU decomposition*, Numer. Math., 82 (1999), pp. 543–576.

[4] M. Benzi and M. Tůma, *A robust preconditioner with low memory requirements for large sparse least squares problems*, SIAM J. Sci. Comput., 25 (2003), pp. 499–512.

[5] Å. Björck, *SSOR preconditioning methods for sparse least squares problems*, in Proceedings of the Computer Science and Statistics 12th Annual Symposium on the Interface, University of Waterloo, Waterloo, ON, Canada, 1979, pp. 21–25.

[6] Å. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996.

[7] Å. Björck and T. Elfving, *Accelerated projection methods for computing pseudoinverse solutions of systems of linear equations*, BIT, 19 (1979), pp. 145–163.

[8] M. Bollhöfer, *A robust ILU with pivoting based on monitoring the growth of the inverse factors*, Linear Algebra Appl., 338 (2001), pp. 201–213.

[9] M. Bollhöfer and Y. Saad, *Multilevel preconditioners constructed from inverse-based ILUs*, SIAM J. Sci. Comput., 27 (2006), pp. 1627–1650.

[10] R. Bramley and A. Sameh, *Row projection methods for large nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 168–193.
[11] T. F. Coleman and J. J. Moré, *Estimation of sparse Jacobian matrices and graph coloring problems*, SIAM J. Numer. Anal., 20 (1983), pp. 187–209.
[12] E. Elmroth and F. Gustavson, *New serial and parallel recursive QR factorization algorithms for SMP systems*, in Applied Parallel Computing, Lecture Notes in Comput. Sci. 1541, Springer-Verlag, Berlin, 1998, pp. 120–128.
[13] E. Elmroth and F. Gustavson, *Applying recursion to serial and parallel QR factorization leads to better performance*, IBM J. Res. Develop., 44 (2000), p. 605–624.
[14] A. H. Gebremedhin, F. Manne, and A. Pothen, *What color is your Jacobian? Graph coloring for computing derivatives*, SIAM Rev., 47 (2005), pp. 629–705.
[15] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.
[16] M. Hegland, *On the computation of breeding values*, in Proceedings of CONPAR 90 - VAPP IV, Joint International Conference on Vector and Parallel Processing, Lecture Notes in Comput. Sci. 457, Springer-Verlag, New York, 1990, pp. 232–242.
[17] M. Hegland, *Description and Use of Animal Breeding Data for Large Least Squares Problems*, Tech. Report TR/PA/93/50, CERFACS, Toulouse, France, 1993.
[18] A. Jennings and M. A. Ajiz, *Incomplete methods for solving $A^T Ax = b$*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 978–987.
[19] M. Jones and P. Plassmann, *An improved incomplete Cholesky factorization*, ACM Trans. Math. Software, 21 (1995), pp. 5–17.
[20] C. Kamath, *Solution of Nonsymmetric Systems of Equations on a Multiprocessor*, Ph.D. thesis, Center for Supercomputing Research and Development, University of Illinois at Urbana-Champaign, Urbana, IL, 1986.
[21] C. Kamath and A. H. Sameh, *A projection method for solving nonsymmetric linear systems on multiprocessors*, Parallel Comput., 9 (1989), pp. 291–312.
[22] R. Leuze, *Independent set orderings for parallel matrix factorizations by Gaussian elimination*, Parallel Comput., 10 (1989), pp. 177–191.
[23] B. Lévy, S. Petitjean, N. Ray, and J. Maillot, *Least squares conformal maps for automatic texture atlas generation*, in ACM Trans. Graphics, 21 (2002), pp. 362–371.
[24] J. G. Lewis, B. W. Peyton, and A. Pothen, *A fast algorithm for reordering sparse matrices for parallel factorization*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 1146–1173.
[25] Z. Li, Y. Saad, and M. Sosonkina, *pARMS: A parallel version of the algebraic recursive multilevel solver*, Numer. Linear Algebra Appl., 10 (2003), pp. 485–509.
[26] T. A. Manteuffel, *An incomplete factorization technique for positive definite linear systems*, Math. Comp., 34 (1980), pp. 473–497.
[27] W. Niethammer, J. de Pillis, and R. Varga, *Convergence of block iterative methods applied to sparse least-squares problems*, Linear Algebra Appl., 58 (1984), pp. 327–341.
[28] C. C. Paige and M. A. Saunders, *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Software, 8 (1982), pp. 43–71.
[29] A. T. Papadopoulos, I. S. Duff, and A. J. Wathen, *A class of incomplete orthogonal factorization methods.* II: *Implementation and results*, BIT, 45 (2005), pp. 159–179.
[30] Y. Saad, *Preconditioning techniques for indefinite and nonsymmetric linear systems*, J. Comput. Appl. Math., 24 (1988), pp. 89–105.
[31] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelphia, 2003.
[32] Y. Saad and M. Sosonkina, *Enhanced Preconditioners for Large Sparse Least Squares Problems*, Tech. Report umsi-2001-1, Minnesota Supercomputer Institute, University of Minnesota, Minneapolis, MN, 2001.
[33] Y. Saad and B. Suchomel, *ARMS: An algebraic recursive multilevel solver for general sparse linear systems*, Numer. Linear Algebra Appl., 9 (2002), pp. 359–378.
[34] I. Strandén, S. Tsuruta, and I. Misztal, *Simple preconditioners for the conjugate gradient method: Experience with test day models*, J. Anim. Breed. Genet., 119 (2002), pp. 166–174.
[35] X. Wang, K. A. Gallivan, and R. Bramley, *CIMGS: An incomplete orthogonal factorization preconditioner*, SIAM J. Sci. Comput., 18 (1997), pp. 516–536.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 2, pp. 551–575

THE PAGERANK VECTOR: PROPERTIES, COMPUTATION, APPROXIMATION, AND ACCELERATION*

CLAUDE BREZINSKI† AND MICHELA REDIVO-ZAGLIA‡

This work is dedicated to the memory of Prof. Germund Dahlquist

Abstract. An important problem in Web search is determining the importance of each page. After introducing the main characteristics of this problem, we will see that, from the mathematical point of view, it could be solved by computing the left principal eigenvector (the PageRank vector) of a matrix related to the structure of the Web by using the power method. We will give expressions of the PageRank vector and study the mathematical properties of the power method. Various Padé-style approximations of the PageRank vector will be given. Since the convergence of the power method is slow, it has to be accelerated. This problem will be discussed. Recently, several acceleration methods were proposed. We will give a theoretical justification for these methods. In particular, we will generalize the recently proposed Quadratic Extrapolation, and we interpret it on the basis of the method of moments of Vorobyev, and as a Krylov subspace method. Acceleration results are given for the various ϵ-algorithms, and for the E-algorithm. Other algorithms for this problem are also discussed.

Key words. PageRank, Web matrix, eigenvector computation, power method, Padé approximation, convergence acceleration

AMS subject classifications. 65F15, 65B99, 68U35

DOI. 10.1137/050626612

1. Introduction.

An important problem in Web search is classifying the pages according to their importance. In section 2, we formulate and discuss this problem in mathematical terms and explain how a rank is assigned to each page for creating the so-called PageRank vector. Various expressions of this vector are given in section 3. Since the PageRank vector is the dominant eigenvector of a stochastic and irreducible matrix, it can be computed by the power method, whose iterates are analyzed in section 4. The results of these two sections will justify the choices made for approximating of the PageRank vector and for accelerating the power method. Section 5 is devoted to the construction of Padé-style rational approximations of the PageRank vector. In section 6, we first present some general ideas on the acceleration of vector sequences by extrapolation. Procedures based on vector least squares extrapolation are discussed. Then, using this framework, we consider several algorithms which were recently proposed for accelerating the convergence of the power method [31]. Their effectiveness is theoretically justified. One of them is connected to Krylov subspace methods, and to the method of moments of Vorobyev [50, 8]. The application of the various ϵ-algorithms and of the E-algorithm to the PageRank problem is studied, and convergence acceleration results are proved. Finally, other possible acceleration techniques are considered.

*Received by the editors March 11, 2005; accepted for publication (in revised form) by V. Simoncini February 7, 2006; published electronically July 31, 2006.
http://www.siam.org/journals/simax/28-2/62661.html

†Laboratoire Paul Painlevé, UMR CNRS 8524, UFR de Mathématiques Pures et Appliquées, Université des Sciences et Technologies de Lille, France (Claude.Brezinski@univ-lille1.fr). The work of this author was partially supported by the project INTAS 03-51-6637.

‡Dipartimento di Matematica Pura ed Applicata, Università degli Studi di Padova, Italy (Michela.RedivoZaglia@unipd.it). The work of this author was supported by MIUR grant 2004015437.

Let us also mention that other classes of acceleration techniques, such as aggregation/disaggregation [30, 35, 25], lumping [37], adaptive methods [29], and the parallel computation of the PageRank vector [22, 32] are not discussed herein.

2. The problem. A query to a Web search engine often produces a very long list of answers because of the enormous number of pages (over 8 billion in Google's database). To help the surfer, these pages have to be listed starting from the most relevant ones. Google uses several metrics and strategies for solving this problem.

The importance of a page is called its, and the [18, 41] is reportedly one of the main ingredients of Google's link analysis. A page is considered to be important if many other important pages are pointing to it. So, the importance of a page is determined by the importance of the other pages. This means that the row vector $\mathbf{r}^T$ of all PageRanks is only defined implicitly as the solution of a fixed-point problem, as we will see now.

Let $\deg(i) \geq 1$ be the outdegree (that is, the number of pages it points to) of the page i. Let $P = (p_{ij})$ be the matrix which describes the transitions between the pages i and j, where $p_{ij} = 1/\deg(i)$, $p_{ij} = 0$ if there is no outlink from page i to page j, and $p_{ii} = 0$.

The PageRank vector $\mathbf{r}^T$ satisfies $\mathbf{r}^T = \mathbf{r}^T P$, that is, $\mathbf{r} = P^T\mathbf{r}$, and it can be computed recursively by the standard power method

$$\mathbf{r}^{(n+1)} = P^T\mathbf{r}^{(n)}, \quad n = 0, 1, \ldots,$$

assuming that $\mathbf{r}$ is present in the spectral decomposition of $\mathbf{r}^{(0)}$. Unfortunately, this iterative procedure has convergence problems. It can cycle, or the limit can depend on the starting vector $\mathbf{r}^{(0)}$ [33].

To avoid these drawbacks, the original PageRank algorithm was revised.

First, since some pages have no outlink (pages), P is not stochastic (some of its rows are zero). Different strategies were proposed to remedy this problem, but the most used one is to replace P by another matrix $\widetilde{P}$ as follows. Let $\mathbf{w} = (w_1, \ldots, w_p)^T \in \mathbb{R}^p$ be a probability vector, that is, such that $\mathbf{w} \geq 0$ and $\mathbf{e}^T\mathbf{w} = 1$ with $\mathbf{e} = (1, \ldots, 1)^T$, and p is the total number of pages. Let $\mathbf{d} = (d_i) \in \mathbb{R}^p$ be the vector with $d_i = 1$ if $\deg(i) = 0$, and 0 otherwise. We set

$$\widetilde{P} = P + \mathbf{d}\mathbf{w}^T.$$

The effect of the additional matrix $\mathbf{d}\mathbf{w}^T$ is to modify the probabilities so that a surfer visiting a page without outlinks jumps to another page with the probability distribution defined by $\mathbf{w}$. This matrix $\widetilde{P}$ is stochastic, and thus it has 1 as its dominant eigenvalue, with $\mathbf{e}$ as its corresponding right eigenvector. So $I - \widetilde{P}$ is singular.

Another problem arises since $\widetilde{P}$ is reducible. In that case, $\widetilde{P}$ can have several eigenvalues on the unit circle, thus causing convergence problems to the power method. Moreover, $\widetilde{P}$ can have several left eigenvectors corresponding to its dominant eigenvalue 1 (see [3, 47, 49] for a general discussion, and [17] or [44] for the particular case of the PageRank problem). Then $\widetilde{P}$ itself is finally replaced by the matrix

$$P_c = c\widetilde{P} + (1 - c)E, \qquad E = \mathbf{e}\mathbf{v}^T,$$

with $c \in [0, 1]$ and $\mathbf{v}$ a probability vector. It corresponds to adding to all pages a new set of outgoing transitions with small probabilities. The probability distribution

given by the vector $\mathbf{v}$ can differ from a uniformly distributed vector, and the resultant PageRank can be biased to give preference to certain kinds of pages. For that reason, $\mathbf{v}^T$ is called the [illegible] vector. The matrix P_c is nonnegative, stochastic, and now irreducible since $\mathbf{v}$ is a positive vector. It has only one eigenvalue on the unit circle. This eigenvalue is equal to 1, and $\mathbf{e}$ is its corresponding right eigenvector [3, 40, 47, 49]. Indeed

$$P_c\mathbf{e} = c\widetilde{P}\mathbf{e} + (1-c)\mathbf{e}\mathbf{v}^T\mathbf{e} = c\mathbf{e} + (1-c)\mathbf{e} = \mathbf{e}.$$

Thus, the matrix $I - P_c$ is singular. The power iterations for the matrix P_c^T now converge to a unique vector $\mathbf{r}_c$ (obviously, depending on c), which is chosen as the PageRank vector. Let us mention that P is extremely sparse, while P_c is completely dense. However, the power method could be implemented only with sparse matrix-vector multiplications, and without even storing P_c as described in section 4. As will be seen below, the vector $\mathbf{r}_c$ can also be computed as the solution of a system of linear equations.

The PageRank problem is closely related to Markov chains [34]. For properties of stochastic matrices, we refer the interested reader to [40] and [47]. For nonnegative matrices, see [3] and [49].

We are finally faced with the following mathematical problem. We set $A_c = P_c^T$. The $p \times p$ matrix A_c has eigenvalues $|c\widetilde{\lambda}_p| \leq \cdots \leq |c\widetilde{\lambda}_2| < \widetilde{\lambda}_1 = 1$, where the $\widetilde{\lambda}_i$ are the eigenvalues of $\widetilde{P}$, and we have to compute $\mathbf{r}_c$, its unique right eigenvector corresponding to the eigenvalue $\widetilde{\lambda}_1 = 1$ [20, 34]. For that purpose, we can use the power method, which consists in the iterations

$$\mathbf{r}_c^{(n+1)} = A_c\mathbf{r}_c^{(n)}, \qquad n = 0, 1, \ldots, \tag{1}$$

with $\mathbf{r}_c^{(0)}$ given.

The vector $\mathbf{r}_c^{(0)}$ is the probability distribution over the surfer's location at step time 0, and $\mathbf{r}_c^{(n)}$ is its probability distribution at time n. The unique stationary distribution vector of the Markov chain characterized by A_c is the limit of the sequence $(\mathbf{r}_c^{(n)})$, which always exists since A_c is primitive and irreducible, and it is independent of $\mathbf{r}_c^{(0)}$. This limit is the right eigenvector $\mathbf{r}_c$ of the matrix A_c corresponding to its dominant eigenvalue 1, and it is exactly the vector that we would like to compute [40, p. 691].

The sequence $(\mathbf{r}_c^{(n)})$ given by (1) always converges to $\mathbf{r}_c$, but if $c \simeq 1$, the convergence is slow since the power method converges as c^n (see [34], and Property 12 below). So, a balance has to be found between a small value of c, which insures a fast convergence of $(\mathbf{r}_c^{(n)})$, but to a vector $\mathbf{r}_c$ which is not close to the real PageRank vector $\widetilde{\mathbf{r}} = \lim_{c\to 1} \mathbf{r}_c$, and a value of c close to 1, which leads to a better approximation $\mathbf{r}_c$ of $\widetilde{\mathbf{r}}$, but with a slow convergence. Originally, Google chose $c = 0.85$, which insures a good rate of convergence [18].

However, computing a PageRank vector can take several days, and so convergence acceleration is essential, in particular, for providing continuous updates to ranking. Moreover, some recent approaches require the computation of several PageRank vectors corresponding to different personalization vectors. Recently, several methods for accelerating the computation of the PageRank vector by the power method were proposed [31, 29]. In this paper we will provide a theoretical justification of the methods of [31], and we will put them on a firm theoretical basis. Other convergence acceleration procedures will also be proposed and discussed. In order to be able to prove that

these algorithms accelerate the convergence of the power method, they have to be strongly supported by theoretical results. This is what will be achieved in this paper. It is not our purpose here to test these algorithms numerically, nor to compare them with other possible procedures for obtaining the PageRank vector.

For a detailed exposition of the PageRank problem, see [34] and [36]. Other reviews are [26] and [2].

3. The PageRank vector. Since P_c is stochastic and irreducible, $\mathbf{r}_c$ is the unique right eigenvector of $A_c = P_c^T$ corresponding to the simple eigenvalue 1, that is, $A_c\mathbf{r}_c = \mathbf{r}_c$. By the Perron–Frobenius theorem (see, for example, [49, p. 35]), $\mathbf{r}_c \geq 0$. It is normalized so that $\mathbf{e}^T\mathbf{r}_c = 1$, and, thus, it is a probability vector.

In this section, we will study the properties of this vector, and, in particular, we will give implicit and explicit expressions for it. Then we will discuss its computation by the power method. This discussion will lead us, in the next two sections, to various procedures for its approximation, and to processes for accelerating the convergence of the power method.

3.1. Implicit expressions for the PageRank vector. Let us give implicit expressions for $\mathbf{r}_c$.

Setting $\widetilde{A} = \widetilde{P}^T$, we have

$$\begin{aligned} A_c\mathbf{r}_c &= c\widetilde{A}\mathbf{r}_c + (1-c)\mathbf{v}\mathbf{e}^T\mathbf{r}_c \\ &= c\widetilde{A}\mathbf{r}_c + (1-c)\mathbf{v} \\ &= \mathbf{r}_c. \end{aligned}$$

Thus, $(I - c\widetilde{A})\mathbf{r}_c = (1-c)\mathbf{v}$, that is, we have the following.

PROPERTY 1.

$$\begin{aligned} \mathbf{r}_c &= (1-c)(I - c\widetilde{A})^{-1}\mathbf{v} \\ &= \mathbf{v} + c(\widetilde{A} - I)(I - c\widetilde{A})^{-1}\mathbf{v}. \end{aligned}$$

The second expression is deduced from the first one by noticing that $(I - c\widetilde{A})^{-1} = I + c\widetilde{A}(I - c\widetilde{A})^{-1}$.

Following Property 1, $\mathbf{r}_c$ can be obtained as the solution of the dense system of linear equations $(I - c\widetilde{A})\mathbf{r}_c = (1-c)\mathbf{v}$. Replacing A_c by its expression leads to $(I - cP^T - c\mathbf{w}\mathbf{d}^T)\mathbf{r}_c = (1-c)\mathbf{v}$. But $\mathbf{e}^T\mathbf{w}\mathbf{d}^T = \mathbf{e}^T\widetilde{A} - \mathbf{e}^TP^T$. Thus, since $\mathbf{e}^T\mathbf{w} = 1$ and $\mathbf{e}^T = \mathbf{e}^T\widetilde{A}$, we have $\mathbf{d}^T = \mathbf{e}^T - \mathbf{e}^TP^T$, and, when $\mathbf{w} = \mathbf{v}$, we finally obtain the sparse system $(I - cP^T)\mathbf{r}_c = \gamma\mathbf{v}$, where $\gamma = \|\mathbf{r}_c\|_1 - c\|P^T\mathbf{r}_c\|_1$ [22]. A particular choice of γ only results in a rescaling of the solution of this system, and it can always be chosen so that $\mathbf{r}_c$ is a probability vector. Various iterative methods for the solution of this system are discussed in many papers, including [1, 4, 19].

From Property 1, we immediately obtain the following.

PROPERTY 2.

$$\begin{aligned} \mathbf{r}_c &= (1-c)\sum_{i=0}^{\infty} c^i\widetilde{A}^i\mathbf{v} \\ &= \mathbf{v} + c(\widetilde{A} - I)\sum_{i=0}^{\infty} c^i\widetilde{A}^i\mathbf{v}. \end{aligned}$$

These results were proved in [5]. These series are convergent since $\rho(\widetilde{A}) = 1$ and $0 \leq c < 1$. Since $\mathbf{r}_c$ can be expressed as a power series, it will allow us to construct rational approximations of it; see section 5.

1. It is easy to check from the result of Property 2 that $\mathbf{e}^T\mathbf{r}_c = 1$. Indeed $\mathbf{e}^T\widetilde{A}^i = \mathbf{e}^T$, and thus

$$\mathbf{e}^T\mathbf{r}_c = (1-c)\sum_{i=0}^{\infty} c^i \mathbf{e}^T\widetilde{A}^i\mathbf{v} = (1-c)\sum_{i=0}^{\infty} c^i \mathbf{e}^T\mathbf{v} = (1-c)\sum_{i=0}^{\infty} c^i,$$

since $\mathbf{e}^T\mathbf{v} = 1$. But $\sum_{i=0}^{\infty} c^i = (1-c)^{-1}$, which shows the result.

3.2. Explicit expressions for the PageRank vector. Let us now give explicit forms for $\mathbf{r}_c$. We will first express it as a rational function, and then propose a polynomial form.

3.2.1. Rational expressions. We assume that $\widetilde{P}$ is diagonalizable. Thus, $\widetilde{P} = XDX^{-1}$, where $D = \mathrm{diag}(1, \widetilde{\lambda}_2, \ldots, \widetilde{\lambda}_p)$, and where $X = [\mathbf{e}, \mathbf{x}_2, \ldots, \mathbf{x}_p]$ is the matrix whose columns are the right eigenvectors of $\widetilde{P}$. Also, let $Y = [\widetilde{\mathbf{r}}, \mathbf{y}_2, \ldots, \mathbf{y}_p]$ be the matrix whose columns are the right eigenvectors of $\widetilde{P}^T$, that is, $\widetilde{A}$. We have $X^{-T} = Y$ and

$$(I - c\widetilde{A})^{-1} = X^{-T}(I - cD)^{-1}X^T.$$

But

$$(I-cD)^{-1} = \begin{pmatrix} (1-c)^{-1} & & & \\ & (1-c\widetilde{\lambda}_2)^{-1} & & \\ & & \ddots & \\ & & & (1-c\widetilde{\lambda}_p)^{-1} \end{pmatrix}, \quad X^T\mathbf{v} = \begin{pmatrix} 1 \\ \mathbf{x}_2^T\mathbf{v} \\ \vdots \\ \mathbf{x}_p^T\mathbf{v} \end{pmatrix},$$

and it follows that

$$\mathbf{u} = (I-cD)^{-1}X^T\mathbf{v} = \begin{pmatrix} 1/(1-c) \\ \mathbf{x}_2^T\mathbf{v}/(1-c\widetilde{\lambda}_2) \\ \vdots \\ \mathbf{x}_p^T\mathbf{v}/(1-c\widetilde{\lambda}_p) \end{pmatrix}.$$

So, we finally obtain $\mathbf{r}_c = (1-c)X^{-T}\mathbf{u} = (1-c)Y\mathbf{u}$. This result was given in [44], where a similar proof when $\widetilde{P}$ is not diagonalizable could also be found, thus leading to the following result.

PROPERTY 3. $\widetilde{P}$

$$\mathbf{r}_c = \widetilde{\mathbf{r}} + (1-c)\sum_{i=2}^{p} \frac{\alpha_i}{1-c\widetilde{\lambda}_i}\mathbf{y}_i,$$

$\alpha_i = \mathbf{x}_i^T\mathbf{v}$

$$\mathbf{r}_c = \widetilde{\mathbf{r}} + \sum_{i=2}^{p} w_i(c)\,\mathbf{y}_i,$$

$$\begin{aligned} w_2(c) &= (1-c)\alpha_2/(1-c\widetilde{\lambda}_2), \\ w_i(c) &= [(1-c)\alpha_i + c\beta_i w_{i-1}(c)]/(1-c\widetilde{\lambda}_i), \quad i = 3, \ldots, p, \end{aligned}$$

β_i 0 1

It follows from this result that since $\mathbf{r}_c$ is a rational function without poles at $c = 1$, there exists a unique vector which is the limit, when c tends to 1, of $\mathbf{r}_c$. This vector is only one of the nonnegative normalized dominant left eigenvectors of $\widetilde{P}$, and it will be chosen as the real PageRank vector. This vector depends on $\mathbf{v}$, a natural property since $\widetilde{P}$ depends on the personalization vector, and also on the multiplicity of the eigenvalue 1 as explained in [17]. Let us mention that, as proved in [20, 34], the eigenvalues of the matrix P_c are $c\widetilde{\lambda}_i$ for $i \geq 2$; see also [24], where it is stated that $\widetilde{\lambda}_2 = 1$ in the special case of the Google matrix.

Let us now give another expression for $\mathbf{r}_c$. It comes from the well-known expression for the resolvent of a matrix (see, for example, [9, pp. 19–20]). We have

$$(I - c\widetilde{A})^{-1} = \frac{1}{\det(I - c\widetilde{A})}(I + cB_1 + \cdots + c^{p-1}B_{p-1}),$$

where the matrices B_i are given by the Le Verrier–Faddeev–Souriau algorithm

$$\begin{aligned} B_1 &= \widetilde{A} + \gamma_1 I, \qquad \gamma_1 = -\mathrm{tr}(\widetilde{A}),\\ B_i &= \widetilde{A}B_{i-1} + \gamma_i I, \quad \gamma_i = -\frac{1}{i}\,\mathrm{tr}(\widetilde{A}B_{i-1}), \qquad i = 2, \ldots, p, \end{aligned}$$

where "tr" designates the trace of a matrix. Moreover

$$\begin{aligned} \det(I - c\widetilde{A}) &= 1 + \gamma_1 c + \cdots + \gamma_p c^p,\\ B_p &= 0,\\ \widetilde{A}^{-1} &= -(1/\gamma_p)B_{p-1}. \end{aligned}$$

Thus, it follows from Property 1 that

$$\mathbf{r}_c = \frac{1 - c}{1 + \gamma_1 c + \cdots + \gamma_p c^p}(I + cB_1 + \cdots + c^{p-1}B_{p-1})\mathbf{v}.$$

This result shows that $\mathbf{r}_c$ is a rational function with a vector numerator of degree p at most, and a scalar denominator of degree p, while in Property 3 both degrees were at most $p - 1$. Let us conciliate these two results.

Since $\widetilde{A}$ has an eigenvalue equal to 1, then $1 + \gamma_1 + \cdots + \gamma_p = 0$, and it follows that

$$\begin{aligned} 1 + \gamma_1 c + \cdots + \gamma_p c^p &= -\gamma_1 - \cdots - \gamma_p + \gamma_1 c + \cdots + \gamma_p c^p\\ &= \gamma_1(c - 1) + \gamma_2(c^2 - 1) + \cdots + \gamma_p(c^p - 1). \end{aligned}$$

Thus, after cancellation of $c - 1$ in the numerator and in the denominator, we obtain the following.

PROPERTY 4.

$$\mathbf{r}_c = -\frac{(I + cB_1 + \cdots + c^{p-1}B_{p-1})\mathbf{v}}{\gamma_1 + \gamma_2(1 + c) \cdots + \gamma_p(1 + \cdots + c^{p-1})}.$$

2. In this expression of $\mathbf{r}_c$, the denominator can also be written as $\beta_0 + \cdots + \beta_{p-1}c^{p-1}$ with $\beta_i = \gamma_{i+1} + \cdots + \gamma_p$ for $i = 0, \ldots, p - 1$.

Notice that if the minimal polynomial of A_c for the vector $\mathbf{v}$ has degree $m < p$, then cancellation occurs between the scalar denominator polynomial and the matrix numerator polynomial, thus reducing $\mathbf{r}_c$ to a rational function of type $(m - 1, m - 1)$ [21, pp. 87–94].

Since, by Properties 3 and 4, $\mathbf{r}_c$ is a rational function in the variable c, it is justifiable to approximate it by a rational function with lower degrees, as proposed in section 5.

3.2.2. Polynomial form. We will now give a polynomial expression for $\mathbf{r}_c$. Let $\Pi_m(\lambda) = a_0 + a_1\lambda + \cdots + a_m\lambda^m$ be the minimal polynomial of A_c for the vector $\mathbf{v}$ with $m \leq p$. Since A_c has a unique eigenvalue equal to 1, Π_m can be written as $\Pi_m(\lambda) = (\lambda - 1)Q_{m-1}(\lambda)$. So

$$\Pi_m(A_c)\mathbf{v} = (A_c - I)Q_{m-1}(A_c)\mathbf{v} = A_cQ_{m-1}(A_c)\mathbf{v} - Q_{m-1}(A_c)\mathbf{v} = 0.$$

Thus, $Q_{m-1}(A_c)\mathbf{v}$ is the eigenvector of A_c corresponding to the eigenvalue 1, that is, we have the following.

PROPERTY 5.

$$\mathbf{r}_c = Q_{m-1}(A_c)\mathbf{v}.$$

If we set $Q_{m-1}(\lambda) = b_0 + \cdots + b_{m-1}\lambda^{m-1}$, then $b_i = -(a_0 + \cdots + a_i) = a_{i+1} + \cdots + a_m$ for $i = 0, \ldots, m-1$ (compare with Remark 2).

Property 5, shows that approximating Q_{m-1} in some sense will lead to approximations of the vector $\mathbf{r}_c$. Such procedures will be described in section 6.

4. Computation of the PageRank vector. The PageRank vector $\mathbf{r}_c$ can be computed by the power method starting from any nonzero vector such that $\mathbf{e}^T\mathbf{r}_c^{(0)} = 1$. We will start it from $\mathbf{v}$, a choice justified by Property 5, and by Property 7 given below:

$$\begin{aligned} \mathbf{r}_c^{(0)} &= \mathbf{v}, \\ \mathbf{r}_c^{(n+1)} &= A_c\mathbf{r}_c^{(n)}, \quad n = 0, 1, \ldots. \end{aligned}$$

Obviously, for all $n, \mathbf{r}_c^{(n)} \geq 0$. Moreover, $\mathbf{e}^T\mathbf{r}_c^{(0)} = 1$. So, by induction, $\mathbf{e}^T\mathbf{r}_c^{(n+1)} = \mathbf{e}^TA_c\mathbf{r}_c^{(n)} = (P_c\mathbf{e})^T\mathbf{r}_c^{(n)} = \mathbf{e}^T\mathbf{r}_c^{(n)}$. Thus, we have the following.

PROPERTY 6.

$$\mathbf{r}_c^{(n)} = A_c^n\mathbf{v} \geq 0, \qquad \|\mathbf{r}_c^{(n)}\|_1 = \mathbf{e}^T\mathbf{r}_c^{(n)} = 1, \quad n = 0, 1, \ldots.$$

Substituting A_c by its expression, an iterate of the power method can be written as

$$\mathbf{r}_c^{(n+1)} = cP^T\mathbf{r}_c^{(n)} + c(\mathbf{d}^T\mathbf{r}_c^{(n)})\mathbf{w} + (1 - c)\mathbf{v}.$$

So, an iteration costs only one matrix–vector product by the very sparse matrix P^T. Moreover, neither A_c nor $\widetilde{A}$ has to be stored. In addition, the vector $\mathbf{d}$ can be eliminated, thus making the power method easy and cheap to implement. Since, as seen above, $\mathbf{d}^T = \mathbf{e}^T - \mathbf{e}^TP^T$, then, for any vector $\mathbf{x}$, it holds, after replacing A_c, $\widetilde{A}$, and $\mathbf{d}^T$ by their expressions, that

$$A_c\mathbf{x} = cP^T\mathbf{x} + (c\|\mathbf{x}\|_1 - \|cP^T\mathbf{x}\|_1)\mathbf{w} + (1 - c)\|\mathbf{x}\|_1\mathbf{v}.$$

If $\mathbf{w} = \mathbf{v}$, the formula given in [31, Alg. 1] for computing such matrix–vector products is recovered. In the particular case of the power method, $\mathbf{x} = \mathbf{r}_c^{(n)}$, $\|\mathbf{r}_c^{(n)}\|_1 = 1$, and the above formula simplifies to

$$\mathbf{r}_c^{(n+1)} = A_c\mathbf{r}_c^{(n)} = cP^T\mathbf{r}_c^{(n)} + (c - \|cP^T\mathbf{r}_c^{(n)}\|_1)\mathbf{w} + (1 - c)\mathbf{v}.$$

Only one vector has to be stored by iteration. See [34] for details about the operational count.

As will be seen in Property 14, it follows from Property 6 that the vectors $\mathbf{r}_c^{(n)} - \mathbf{r}_c$ satisfy a difference equation, a result that will be used for proving that the ϵ-algorithms accelerate the convergence of the power method (see section 6.3).

As proved in [5], an important property is that the vectors $\mathbf{r}_c^{(n)}$ computed by the power method are the partial sums of the second series for $\mathbf{r}_c$ given in Property 2. Let us give a simpler proof of this result.

PROPERTY 7.

$$\begin{aligned}\mathbf{r}_c^{(n)} &= (1-c)\sum_{i=0}^{n-1} c^i \widetilde{A}^i \mathbf{v} + c^n \widetilde{A}^n \mathbf{v}, \quad n \geq 0,\\ &= \mathbf{v} + c(\widetilde{A} - I)\sum_{i=0}^{n-1} c^i \widetilde{A}^i \mathbf{v}.\end{aligned}$$

PROOF. Let us prove the second identity. For $n = 0$, the sum is zero and the result is true. For $n = 1$, we have

$$\mathbf{r}_c^{(1)} = c\widetilde{A}\mathbf{r}_c^{(0)} + (1-c)\mathbf{v}\mathbf{e}^T\mathbf{r}_c^{(0)} = \mathbf{v} + (\widetilde{A} - I)c\mathbf{v}.$$

Assuming that the result holds for n, we have

$$\begin{aligned}\mathbf{r}_c^{(n+1)} &= [c\widetilde{A} + (1-c)\mathbf{v}\mathbf{e}^T]\mathbf{r}_c^{(n)}\\ &= c\widetilde{A}\mathbf{r}_c^{(n)} + (1-c)\mathbf{v} \quad \text{by Property 6}\\ &= c\widetilde{A}\mathbf{v} + c^2\widetilde{A}(\widetilde{A} - I)\sum_{i=0}^{n-1} c^i \widetilde{A}^i \mathbf{v} + (1-c)\mathbf{v}\\ &= c\widetilde{A}\mathbf{v} + c(\widetilde{A} - I)\sum_{i=1}^{n} c^i \widetilde{A}^i \mathbf{v} + (1-c)\mathbf{v}\\ &= \mathbf{v} + c(\widetilde{A} - I)\sum_{i=0}^{n} c^i \widetilde{A}^i \mathbf{v}.\end{aligned}$$

The first result can be easily obtained from the second one. ∎

Since the power method furnishes the partial sums of the power series for $\mathbf{r}_c$, its iterates will be directly used for constructing Padé-type approximants of this vector; see section 5.

Property 8 immediately follows from Property 7.

PROPERTY 8.

$$\begin{aligned}\mathbf{r}_c^{(0)} &= \mathbf{v},\\ \mathbf{r}_c^{(n+1)} &= \mathbf{r}_c^{(n)} + c^{n+1}(\widetilde{A} - I)\widetilde{A}^n \mathbf{v}, \quad n = 0, 1, \ldots.\end{aligned}$$

Moreover, the following holds.

PROPERTY 9.

$$(\widetilde{A} - I)\widetilde{A}^n \mathbf{v} = \frac{1}{c^{n+1}}(\mathbf{r}_c^{(n+1)} - \mathbf{r}_c^{(n)}), \quad n = 0, 1, \ldots.$$

This property, proved in [5], shows that it is possible to apply the power method simultaneously for several values of c with only a small additional cost. Indeed, by Property 9, one only has to compute the vectors $(\widetilde{A} - I)\widetilde{A}^n \mathbf{v}$ once, and then use

Property 8 for computing the partial sums $\mathbf{r}_{\widetilde{c}}^{(n)}$ of the series $\mathbf{r}_{\widetilde{c}}$ for a different value $\widetilde{c}$ of c. So, we have

$$\begin{aligned}\mathbf{r}_{\widetilde{c}}^{(0)} &= \mathbf{v},\\ \mathbf{r}_{\widetilde{c}}^{(n+1)} &= \mathbf{r}_{\widetilde{c}}^{(n)} + \widetilde{c}^{n+1}\frac{1}{c^{n+1}}(\mathbf{r}_c^{(n+1)} - \mathbf{r}_c^{(n)}), \quad n = 0, 1, \ldots.\end{aligned}$$

Since $I + c\widetilde{A} + \cdots + c^{n-1}\widetilde{A}^{n-1} = (I - c\widetilde{A})^{-1}(I - c^n\widetilde{A}^n)$, the results of Property 7 can also be written as follows for comparison with Property 1.

PROPERTY 10.

$$\begin{aligned}\mathbf{r}_c^{(n)} &= (1-c)(I - c\widetilde{A})^{-1}(I - c^n\widetilde{A}^n)\mathbf{v} + c^n\widetilde{A}^n\mathbf{v}\\ &= \mathbf{v} + c(\widetilde{A} - I)(I - c\widetilde{A})^{-1}(I - c^n\widetilde{A}^n)\mathbf{v}.\end{aligned}$$

Let us now give expressions for the error. From Properties 1, 6, and 10, it is easy to prove the following.

PROPERTY 11.

$$\begin{aligned}\mathbf{r}_c - \mathbf{r}_c^{(n)} &= A_c^n(\mathbf{r}_c - \mathbf{v})\\ &= c^n\widetilde{A}^n(\mathbf{r}_c - \mathbf{v})\\ &= (I - c\widetilde{A})^{-1}(\mathbf{r}_c^{(n+1)} - \mathbf{r}_c^{(n)}).\end{aligned}$$

Since $\widetilde{A}$ is a column stochastic matrix $\|\widetilde{A}\|_1 = 1$, and since, in our case, it is also reducible, then $|\widetilde{\lambda}_2| = 1$, and we obtain the following.

PROPERTY 12.

$$\begin{aligned}\|\mathbf{r}_c - \mathbf{r}_c^{(n)}\|_1 &\le c^n \ \|\mathbf{r}_c - \mathbf{v}\|_1\\ &\le \frac{1}{1-c}\ \|\mathbf{r}_c^{(n+1)} - \mathbf{r}_c^{(n)}\|_1.\end{aligned}$$

Let us note that $1/(1-c)$ is the 1-norm of the matrix $(I - c\widetilde{A})^{-1}$ and that the condition number of the PageRank problem is $(1+c)/(1-c)$ [28].

Let us now explain how rational and polynomial approximations of $\mathbf{r}_c$ could be obtained from the iterates of the power method. In both cases, increasing the degree of the approximation produces a sequence of approximations of $\mathbf{r}_c$ of increasing order which, under certain assumptions, converge to $\mathbf{r}_c$ faster than the iterates of the power method.

5. Padé approximation of the PageRank vector. As proved in Properties 3 and 4, $\mathbf{r}_c$ is a vector rational function of type $(p-1, p-1)$ (or $(m-1, m-1)$, where m is the degree of the minimal polynomial of A_c for the vector $\mathbf{v}$) in the variable c, that is, a rational function with a numerator of degree $p-1$ (or $m-1$) with vector coefficients, and a common scalar denominator of degree $p-1$ (or $m-1$). Moreover, by Property 2, the vector Taylor series expansion of $\mathbf{r}_c$ is known. So, the partial sums of this series could be used for constructing rational approximations of $\mathbf{r}_c$ of type $(k-1, k-1)$ with $k < p$ (or $k < m$). The coefficients of these rational functions will be chosen so that their power series expansion agrees with that of $\mathbf{r}_c$ as far as possible. Such types of rational functions are called Padé approximants.

The first possibility is to construct the scalar Padé approximants $[k-1/k-1]$ separately for each component of $\mathbf{r}_c$. In that case, each component will be matched up to the term of degree $2k-2$ inclusively. However, each scalar Padé approximant could

have a different denominator for each component. For more on Padé approximation, see [6, 9].

A solution that seems preferable is to use vector Padé approximants since the components of $\mathbf{r}_c$ are rational functions with a common denominator, which is exactly the characterizing property of vector Padé approximants. These approximants, introduced by Van Iseghem [48], are defined as follows.

Let $\mathbf{f}$ be a vector formal power series

$$\mathbf{f}(\xi) = \sum_{i=0}^{\infty} \boldsymbol{\sigma}_i \xi^i, \quad \boldsymbol{\sigma}_i \in \mathbb{R}^p. \tag{2}$$

We look for a vector rational function whose series expansion in ascending powers of ξ agrees with $\mathbf{f}$ as far as possible. By vector rational function, we mean a function with vector coefficients in the numerator and with a scalar denominator. More precisely, we look for $\mathbf{a}_0, \ldots, \mathbf{a}_{k-1} \in \mathbb{R}^p$, and $b_0, \ldots, b_{k-1} \in \mathbb{R}$, with $k \leq p$ (or $k \leq m$), such that

$$(b_0 + \cdots + b_{k-1}\xi^{k-1})(\boldsymbol{\sigma}_0 + \boldsymbol{\sigma}_1\xi + \cdots) - (\mathbf{a}_0 + \cdots + \mathbf{a}_{k-1}\xi^{k-1}) = \mathcal{O}(\xi^s), \tag{3}$$

with s, the order of approximation, as high as possible. If $s = k$, this vector rational function is called a [illegible] of $\mathbf{f}$, while it is called a [illegible] if $s = 2k - 1$.

Identifying to zero the vector coefficients of the terms of degree 0 to $k-1$ in the left-hand side of (3), we obtain

$$\begin{aligned} \mathbf{a}_0 &= b_0\boldsymbol{\sigma}_0, \\ \mathbf{a}_1 &= b_0\boldsymbol{\sigma}_1 + b_1\boldsymbol{\sigma}_0, \\ &\vdots \\ \mathbf{a}_{k-1} &= b_0\boldsymbol{\sigma}_{k-1} + \cdots + b_{k-1}\boldsymbol{\sigma}_0. \end{aligned} \tag{4}$$

For any choice of the coefficients b_i of the denominator with $b_0 \neq 0$, the rational function $(\mathbf{a}_0 + \cdots + \mathbf{a}_{k-1}\xi^{k-1})/(b_0 + \cdots + b_{k-1}\xi^{k-1})$ obtained by this procedure is a vector Padé-type approximant, and it is denoted by $(k-1/k-1)_{\mathbf{f}}(\xi)$. Its order of approximation is $s = k$, that is, $(k-1/k-1)_{\mathbf{f}}(\xi) - \mathbf{f}(\xi) = \mathcal{O}(\xi^k)$. The computation of the approximant $(k-1/k-1)_{\mathbf{f}}$ needs the knowledge of $\boldsymbol{\sigma}_0, \ldots, \boldsymbol{\sigma}_{k-1}$. Thus, in practice, only small values of k could be used depending on the number of vectors one could store.

Let us now try to improve the order of approximation, that is, to construct vector Padé approximants. However, as we will see now, this order could not be improved simultaneously for all components of the approximants since k has to be smaller than p. Indeed, for eliminating the term of degree k in (3), it is necessary and sufficient that $0 = b_0\boldsymbol{\sigma}_k + \cdots + b_{k-1}\boldsymbol{\sigma}_1$. Since a rational function is defined apart from a multiplying factor, we can set $b_0 = 1$, and we get

$$b_1\boldsymbol{\sigma}_{k-1} + \cdots + b_{k-1}\boldsymbol{\sigma}_1 = -\boldsymbol{\sigma}_k.$$

This is a system of p equations with $k - 1 \leq p$ unknowns. It has to be solved in the least squares sense. Setting $C = [\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_{k-1}] \in \mathbb{R}^{p\times(k-1)}$ and $\mathbf{b} = (b_{k-1}, \ldots, b_1)^T$, this system can be rewritten as $C\mathbf{b} = -\boldsymbol{\sigma}_k$. Let $C^\dagger$ be a left inverse of C, that is, a $(k-1) \times p$ matrix such that $C^\dagger C = I \in \mathbb{R}^{(k-1)\times(k-1)}$. Thus, $\mathbf{b} = -C^\dagger\boldsymbol{\sigma}_k$. Once the b_i's have been obtained, the $\mathbf{a}_i$'s can be directly computed by the relations (4). The matrix $C^\dagger$ has the form $C^\dagger = (Z^TC)^{-1}Z^T$, where $Z = [\mathbf{z}_1, \ldots, \mathbf{z}_{k-1}] \in \mathbb{R}^{p\times(k-1)}$ is

any matrix such that $Z^TC \in \mathbb{R}^{(k-1)\times(k-1)}$ is nonsingular. The rational approximant constructed in that way is called a vector Padé approximant of $\mathbf{f}$, although its order of approximation s is only equal to k, and it is denoted by $[k-1/k-1]_{\mathbf{f}}(\xi)$. The reason for this abuse of language is that we now have $\mathbf{z}_i^T[k-1/k-1]_{\mathbf{f}}(\xi) - \mathbf{z}_i^T\mathbf{f}(\xi) = \mathcal{O}(\xi^{2k-1})$ for $i = 1, \ldots, k-1$. Obviously, since a left inverse is not unique, different vector Padé approximants could be constructed. Of course, the simplest choice is $Z = C$. In that case, $C^\dagger$ is the pseudoinverse of C, and the corresponding Padé approximants can be computed by the Recursive Projection Algorithm (RPA) [13, sect. 4.4].

Another way of proceeding for obtaining the coefficients b_i is to consider the $k-1$ scalar equations

$$\begin{aligned} b_1(\mathbf{z}, \boldsymbol{\sigma}_{k-1}) + \cdots + b_{k-1}(\mathbf{z}, \boldsymbol{\sigma}_1) &= -(\mathbf{z}, \boldsymbol{\sigma}_k), \\ &\vdots \\ b_1(\mathbf{z}, \boldsymbol{\sigma}_{2k-3}) + \cdots + b_{k-1}(\mathbf{z}, \boldsymbol{\sigma}_{k-1}) &= -(\mathbf{z}, \boldsymbol{\sigma}_{2k-2}), \end{aligned}$$

where $\mathbf{z}$ is any vector such that the matrix of this system is nonsingular. For these approximants, we again have $[k-1/k-1]_{\mathbf{f}}(\xi) - \mathbf{f}(\xi) = \mathcal{O}(\xi^k)$, but now $\mathbf{z}^T[k-1/k-1]_{\mathbf{f}}(\xi) - \mathbf{z}^T\mathbf{f}(\xi) = \mathcal{O}(\xi^{2k-1})$. These approximants are more costly than the previous ones since more iterates of the power method are needed. They can be recursively computed by applying the topological ϵ-algorithm [6] to the iterates of the power method.

Intermediate strategies between using one vector equation and $k-1$ vectors $\mathbf{z}_i$ or using $k-1$ vector equations and only one vector $\mathbf{z}$ could also be employed as those described in [11]. Other sequence transformations of the same type are given in [15].

By Property 2, $\mathbf{r}_c$ is also the product of the vector $\mathbf{v}$ by a matrix power series. Thus, $\mathbf{r}_c$ can be approximated by matrix Padé approximants. However, this solution involves high-dimensional matrix inversion, which is not possible in our case.

Let us now apply these general procedures for constructing Padé-type and Padé approximants of $\mathbf{r}_c$. The second series of Property 2 for $\mathbf{r}_c$ corresponds to $\mathbf{f}$ with $\xi = c$, $\boldsymbol{\sigma}_0 = \mathbf{v}$, and $\boldsymbol{\sigma}_i = (\widetilde{A} - I)\widetilde{A}^{i-1}\mathbf{v}$ for $i \geq 1$. Following Property 9, these vector coefficients are obtained directly from the power method since $\boldsymbol{\sigma}_i = (\mathbf{r}_c^{(i+1)} - \mathbf{r}_c^{(i)})/c^{i+1}$. The vector Padé-type approximants satisfy the same property as $\mathbf{r}_c$ and the iterates of the power method, namely, we have the following.

PROPERTY 13.

$$\begin{aligned} &\mathbf{e}^T(k-1/k-1)_{\mathbf{f}}(c) = 1 \quad \forall c \in [0,1], \\ &(k-1/k-1)_{\mathbf{f}}(c) \geq 0 \quad \forall c \in [0,\delta], \quad \delta \in [0,1]. \end{aligned}$$

[illegible]. We have $\mathbf{e}^T\boldsymbol{\sigma}_0 = 1$, and $\mathbf{e}^T\boldsymbol{\sigma}_i = 0$ for $i \geq 1$. Thus, multiplying the relations (4) by $\mathbf{e}^T$ gives $\mathbf{e}^T\mathbf{a}_i = b_i$ for $i = 0, \ldots, k-1$, which shows the first result.

Since $(k-1/k-1)_{\mathbf{f}}(c)$ approximates $\mathbf{r}_c$ near zero, and for all $c \in [0,1), \mathbf{r}_c \geq 0$, then $\exists \delta \in [0,1]$ such that $(k-1/k-1)_{\mathbf{f}}(c) \geq 0$ for $c \in [0,\delta]$. Let us mention that the value of δ is unknown but that is should be 1. □

Let us also mention that the value of c in Padé-style approximants can be complex.

6. Acceleration of the power method. Let us begin by recalling some general issues about sequence transformations for accelerating the convergence.

Let $(\mathbf{x}^{(n)})$ be a sequence converging to a limit $\mathbf{x}$. The idea behind a convergence acceleration method is extrapolation. It is assumed that the sequence $(\mathbf{x}^{(n)})$ behaves in a certain way or, in other terms, as a certain function of n depending on some

unknown parameters including its limit $\mathbf{x}$. These parameters (and so also the limit $\mathbf{x}$) are determined by interpolation starting from an index n, and then the function is extrapolated at infinity. If the sequence $(\mathbf{x}^{(n)})$ behaves exactly as the extrapolation function, then the estimated limit obtained by the extrapolation process gives its exact limit $\mathbf{x}$. If $(\mathbf{x}^{(n)})$ does not behave exactly as the extrapolation function, then the estimated limit given by the extrapolation process is only an approximation of $\mathbf{x}$, denoted by $\mathbf{y}^{(n)}$ since it depends on n. So, an extrapolation process transforms the sequence $(\mathbf{x}^{(n)})$ into a new sequence $(\mathbf{y}^{(n)})$ which, under certain assumptions (that is, if the extrapolation function closely follows the exact behavior of $(\mathbf{x}^{(n)})$), converges to $\mathbf{x}$ faster than $(\mathbf{x}^{(n)})$, that is, $\lim_{n\to\infty} \|\mathbf{y}^{(n)} - \mathbf{x}\|/\|\mathbf{x}^{(n)} - \mathbf{x}\| = 0$. A sequence transformation for accelerating the convergence can always be considered as an extrapolation procedure, and conversely an extrapolation procedure always leads to an acceleration method for some classes of sequences. A sequence transformation does not modify the sequence to accelerate, and the way it is generated is taken into account only for obtaining acceleration results.

The set $\mathcal{K}_T$ of sequences such that, for all n, $\mathbf{y}^{(n)} = \mathbf{x}$ is called the [illegible] of the transformation $T : (\mathbf{x}^{(n)}) \longmapsto (\mathbf{y}^{(n)})$. So, when a sequence belongs to the kernel of a transformation, its limit is exactly obtained. An important conjecture about sequence transformations is that, if a sequence is [illegible], in a sense to be specified, to the kernel of a transformation, then its convergence will be accelerated by this transformation. Many numerical results point out that this conjecture is true. However, theoretical results in this direction are very partial, and no general ones exist.

Thus, for constructing an efficient acceleration process for a given sequence, one must first study its behavior with respect to n, and then construct an extrapolation process based on an extrapolation function as close as possible (in some sense) to the exact one. This extrapolation function will define the kernel of the transformation.

On sequence transformations for accelerating the convergence by extrapolation, and, in particular, for the ϵ-algorithm and the other algorithms that will be used below, see [13].

For defining such an acceleration process for the iterates of the power method for the computation of the PageRank vector, we use the idea behind Krylov's method. As proved in Property 5, $\mathbf{r}_c = Q_{m-1}(A_c)\mathbf{v}$, where $\Pi_m(\lambda) = (\lambda - 1)Q_{m-1}(\lambda)$ is the minimal polynomial of A_c for the vector $\mathbf{v}$. We also have

$$\mathbf{r}_c = A_c^n \mathbf{r}_c = A_c^n Q_{m-1}(A_c)\mathbf{v} = Q_{m-1}(A_c)A_c^n \mathbf{v} = Q_{m-1}(A_c)\mathbf{r}_c^{(n)}. \tag{5}$$

Thus, replacing, in this relation, Q_{m-1} by an approximating polynomial Q_{k-1} of degree $k - 1 \leq m - 1$ leads to polynomial approximations of $\mathbf{r}_c$ of the form

$$\mathbf{r}_c^{(k,n)} = Q_{k-1}(A_c)\mathbf{r}_c^{(n)}. \tag{6}$$

As will be seen below, the polynomials Q_{k-1} will be constructed from the vectors $\mathbf{r}_c^{(i)}$ for $i \geq n$. Under certain assumptions, the new sequences $(\mathbf{r}_c^{(k,n)})$ will converge to $\mathbf{r}_c$ faster than the sequence $(\mathbf{r}_c^{(n+k)})$ produced by the power method, that is, such that the sequence $(\|\mathbf{r}_c^{(k,n)} - \mathbf{r}_c\|/\|\mathbf{r}_c^{(n)} - \mathbf{r}_c\|)$ tends to zero, for k fixed and n tending to infinity. When n is fixed and k increases, then for $k = m$, the degree of the minimal polynomial of A_c for the vector $\mathbf{v}$, the exact result $\mathbf{r}_c$ is obtained. Obviously, since m is very large this is not a procedure that could be used in practice. However, when k grows, the $(\mathbf{r}_c^{(k,n)})$'s become more accurate approximations of $\mathbf{r}_c$ since we are getting closer to the kernel of the transformation as explained above.

Another procedure, called [illegible], consists of computing $\mathbf{r}_c^{(k,0)}$ as above, and then starting again the iterations (1) of the power method from $\mathbf{r}_c^{(0)} = \mathbf{r}_c^{(k,0)}$. This was the strategy used in [31].

In both cases, it must be noted that the higher k, the greater the number of vectors $\mathbf{r}_c^{(i)}$ obtained by the power method to store, thus limiting the value of k to be used in practice. This value depends on the numbers of vectors one could store. This is an important point to take into consideration since, if many vectors have to be stored for accelerating the power method, one might as well use a more powerful eigenvalue algorithm, like the restarted Arnoldi's method [23]. Even if the PageRank vector is computed only for a subset of the pages, again Arnoldi's method may be more interesting than the accelerated power method. This is a remark to take into account when considering convergence acceleration procedures.

In section 6.1, we will explain in a different way, simplifying, unifying, and generalizing the Quadratic Extrapolation presented in [31]. This generalization will be related to Krylov subspace methods, and some properties will be given. Then, in section 6.2, this generalization will also be included in the framework of the method of moments, where a polynomial P_k approximating the minimal polynomial Π_m will be constructed. In section 6.3, we will discuss the various ϵ-algorithms and recover the Aitken Extrapolation, as well as the Epsilon Extrapolation given in [31]. These algorithms are generalizations of the well-known Aitken's Δ^2 process. The section closes by reviewing some other possible acceleration methods.

6.1. Vector least squares extrapolation. Let us give a first procedure for computing the coefficients of the polynomial P_k which approximates the minimal polynomial Π_m. We set $P_k(\lambda) = a_0 + \cdots + a_k\lambda^k$, where the a_i's depend on k and another index denoted by n, as we will see, and $P_k(1) = a_0 + \cdots + a_k = 0$. Considering the iterates of the power method, we set

$$R_n = [\mathbf{r}_c^{(n)}, \ldots, \mathbf{r}_c^{(n+k-1)}].$$

Since, for all $n, \mathbf{r}_c^{(n)} = A_c^n\mathbf{v}$, it holds that

$$A_c^n P_k(A_c)\mathbf{v} = P_k(A_c)\mathbf{r}_c^{(n)} = a_0\mathbf{r}_c^{(n)} + \cdots + a_k\mathbf{r}_c^{(n+k)} \simeq 0. \tag{7}$$

Since the coefficients a_i are defined apart from a multiplying factor, and since P_k has exact degree k, we can assume that $a_k = 1$ without restricting the generality. Thus, (7) can be rewritten as

$$R_n\mathbf{a} \simeq -\mathbf{r}_c^{(n+k)},$$

with $\mathbf{a} = (a_0, \ldots, a_{k-1})^T$. Solving this system in the least squares sense gives

$$\mathbf{a} = -(R_n^T R_n)^{-1} R_n^T \mathbf{r}_c^{(n+k)}. \tag{8}$$

Let us remark, in connection with [31], that $(R_n^T R_n)^{-1} R_n^T$ is the pseudoinverse of R_n.

By taking into account that $P_k(1) = 0$, the computation can be simplified as in [31]. We have $a_0 = -a_1 - \cdots - a_{k-1} - 1$. Replacing a_0 by this expression in (7) gives

$$R'_n\mathbf{a}' = -(\mathbf{r}_c^{(n+k)} - \mathbf{r}_c^{(n)}) \tag{9}$$

with $R'_n = [\mathbf{r}_c^{(n+1)} - \mathbf{r}_c^{(n)}, \ldots, \mathbf{r}_c^{(n+k-1)} - \mathbf{r}_c^{(n)}]$ and $\mathbf{a}' = (a_1, \ldots, a_{k-1})^T$. This system is then solved in the least squares sense, that is, $\mathbf{a}' = -(R_n'^T R'_n)^{-1} R_n'^T (\mathbf{r}_c^{(n+k)} - \mathbf{r}_c^{(n)})$.

. . . . 3. Instead of formula (8), any other left inverse of R_n could be used, thus leading to

$$\mathbf{a} = -(Z_n^T R_n)^{-1} Z_n^T \mathbf{r}_c^{(n+k)},$$

where $Z_n = [\mathbf{z}_n, \ldots, \mathbf{z}_{n+k-1}]$ is a $p \times k$ matrix such that $Z_n^T R_n$ is nonsingular. The system (9) can be solved in a similar way.

We now have to compute $\mathbf{r}_c^{(k,n)}$ by (6). We set

$$Q_{k-1}(\lambda) = b_0 + b_1\lambda + \cdots + b_{k-1}\lambda^{k-1}.$$

Since $P_k(\lambda) = (\lambda - 1)Q_{k-1}(\lambda)$, it follows that

$$b_i = -(a_0 + \cdots + a_i) = a_{i+1} + \cdots + a_k, \qquad i = 0, \ldots, k-1. \tag{10}$$

Note that since $a_0 + \cdots + a_k = 0$ and $a_k = 1$, we also have $b_0 = -a_0$ and $b_{k-1} = a_k = 1$. Let $\mathbf{b} = (b_0, \ldots, b_{k-1})^T$. Thus, $\mathbf{r}_c^{(k,n)} = R_n\mathbf{b}$. Denoting by L the $k \times k$ lower triangular matrix whose elements are equal to 1, then, from (10), $\mathbf{b} = -L\mathbf{a}$, and it follows that

$$\mathbf{r}_c^{(k,n)} = R_n\mathbf{b} = -R_nL\mathbf{a} = R_nL(R_n^TR_n)^{-1}R_n^Tr_c^{(n+k)}.$$

We also have $A_c\mathbf{r}_c^{(k,n)} = R_{n+1}\mathbf{b}$.

Thus, from what precedes, we obtain

$$\mathbf{r}_c^{(k,n)} = Q_{k-1}(A_c)\mathbf{r}_c^{(n)} = b_0\mathbf{r}_c^{(n)} + b_1\mathbf{r}_c^{(n+1)} + \cdots + b_{k-1}\mathbf{r}_c^{(n+k-1)}. \tag{11}$$

Since $\mathbf{r}_c^{(n+i)} = A_c^i\mathbf{r}_c^{(n)}$, this relation shows that $\mathbf{r}_c^{(k,n)} \in K_k(A_c, \mathbf{r}_c^{(n)})$, the Krylov subspace of dimension k spanned by the vectors $\mathbf{r}_c^{(n)}, \ldots, A_c^{k-1}\mathbf{r}_c^{(n)}$. More precisely, since $b_{k-1} = 1$, $\mathbf{r}_c^{(k,n)} \in \mathbf{r}_c^{(n+k-1)} + K_{k-1}(A_c, \mathbf{r}_c^{(n)})$. Moreover, the vector

$$\mathbf{e}^{(k,n)} = P_k(A_c)\mathbf{r}_c^{(n)} = (A_c - I)Q_{k-1}(A_c)\mathbf{r}_c^{(n)} = A_c\mathbf{r}_c^{(k,n)} - \mathbf{r}_c^{(k,n)}$$

belongs to $K_{k+1}(A_c, \mathbf{r}_c^{(n)})$, more precisely, since $b_{k-1} = 1$, $\mathbf{e}^{(k,n)} \in \mathbf{r}_c^{(n+k)} + K_k(A_c, \mathbf{r}_c^{(n)})$. From (11), we also see that $\mathbf{e}^{(k,n)} = \Delta R_n\mathbf{b} \in K_k(A_c, \Delta\mathbf{r}_c^{(n)})$; more precisely, it belongs to $\Delta\mathbf{r}_c^{(n+k-1)} + K_{k-1}(A_c, \Delta\mathbf{r}_c^{(n)})$ (Δ is the usual forward difference operator). Since $\mathbf{r}_c^{(k,n)}$ approximates the eigenvector $\mathbf{r}_c$ of A_c, the vector $\mathbf{e}^{(k,n)}$ plays the role of a residual. We have

$$\begin{aligned} R_n^T\mathbf{e}^{(k,n)} &= R_n^TR_n\mathbf{a} + R_n^T\mathbf{r}_c^{(n+k)} \\ &= -R_n^TR_n(R_n^TR_n)^{-1}R_n^T\mathbf{r}_c^{(n+k)} + R_n^T\mathbf{r}_c^{(n+k)} \\ &= 0. \end{aligned}$$

Thus, $\mathbf{e}^{(k,n)}$ is orthogonal to the columns of R_n, and we have the following.

THEOREM 1.

$$\begin{aligned} \mathbf{r}_c^{(k,n)} &\in \mathbf{r}_c^{(n+k-1)} + K_{k-1}(A_c, \mathbf{r}_c^{(n)}), \\ A_c\mathbf{r}_c^{(k,n)} - \mathbf{r}_c^{(k,n)} &\perp K_k(A_c, \mathbf{r}_c^{(n)}). \end{aligned}$$

This result shows that vector least squares extrapolation can be considered as a Krylov subspace method for computing $\mathbf{r}_c$.

Moreover, since $K_k(A_c, \mathbf{r}_c^{(n)}) \subseteq K_{k+1}(A_c, \mathbf{r}_c^{(n)})$, we have the following.

COROLLARY 1.

$$\|\mathbf{e}^{(k+1,n)}\| \leq \|\mathbf{e}^{(k,n)}\|.$$

Obviously, when $k = m$, $\mathbf{e}^{(m,n)} = 0$.

Writing down the conditions of Theorem 1, we immediately obtain several determinantal expressions. Such expressions have no direct practical use, but they could be of interest in proving theoretical results about our vector least squares extrapolation, and in obtaining recursive algorithms for the computation of the vectors $\mathbf{r}_c^{(k,n)}$.

COROLLARY 2.

$$\mathbf{e}^{(k,n)} = (-1)^{k-1} \frac{\begin{vmatrix} \mathbf{r}_c^{(n)} & \cdots & \mathbf{r}_c^{(n+k)} \\ (\mathbf{r}_c^{(n)}, \mathbf{r}_c^{(n)}) & \cdots & (\mathbf{r}_c^{(n)}, \mathbf{r}_c^{(n+k)}) \\ \vdots & & \vdots \\ (\mathbf{r}_c^{(n+k-1)}, \mathbf{r}_c^{(n)}) & \cdots & (\mathbf{r}_c^{(n+k-1)}, \mathbf{r}_c^{(n+k)}) \end{vmatrix}}{\begin{vmatrix} (\mathbf{r}_c^{(n)}, \mathbf{r}_c^{(n)}) & \cdots & (\mathbf{r}_c^{(n)}, \mathbf{r}_c^{(n+k-1)}) \\ \vdots & & \vdots \\ (\mathbf{r}_c^{(n+k-1)}, \mathbf{r}_c^{(n)}) & \cdots & (\mathbf{r}_c^{(n+k-1)}, \mathbf{r}_c^{(n+k-1)}) \end{vmatrix}}.$$

The determinant in the numerator denotes the vector obtained by expanding it with respect to its first row by the classical rules for expanding a determinant.

Let $D_k^{(n)}$ be the determinant in the denominator of $\mathbf{e}^{(k,n)}$. Comparing this result with (11) shows, since $\mathbf{r}_c^{(n+i)} = A_c^i \mathbf{r}_c^{(n)}$, that we have the following.

COROLLARY 3. $\mathbf{e}^{(k,n)} = \widetilde{Q}_{k-1}(A_c)\mathbf{r}_c^{(n)}/D_k^{(n)}$, where

$$\widetilde{Q}_{k-1}(\lambda) = (-1)^{k-1} \begin{vmatrix} 1 & \cdots & \lambda^k \\ (\mathbf{r}_c^{(n)}, \mathbf{r}_c^{(n)}) & \cdots & (\mathbf{r}_c^{(n)}, \mathbf{r}_c^{(n+k)}) \\ \vdots & & \vdots \\ (\mathbf{r}_c^{(n+k-1)}, \mathbf{r}_c^{(n)}) & \cdots & (\mathbf{r}_c^{(n+k-1)}, \mathbf{r}_c^{(n+k)}) \end{vmatrix}.$$

Note that the polynomial $\widetilde{Q}_{k-1}(\lambda)/D_k^{(n)}$ is monic. Moreover, the ratio of determinants given in Corollary 2 shows that $\mathbf{e}^{(k,n)}$ can also be expressed as a Schur complement (see [9, p. 150] or [52]), thus leading to the following.

COROLLARY 4.

$$\mathbf{e}^{(k,n)} = \mathbf{r}_c^{(n+k)} - R_n \begin{pmatrix} (\mathbf{r}_c^{(n)}, \mathbf{r}_c^{(n)}) & \cdots & (\mathbf{r}_c^{(n)}, \mathbf{r}_c^{(n+k-1)}) \\ \vdots & & \vdots \\ (\mathbf{r}_c^{(n+k-1)}, \mathbf{r}_c^{(n)}) & \cdots & (\mathbf{r}_c^{(n+k-1)}, \mathbf{r}_c^{(n+k-1)}) \end{pmatrix}^{-1} \begin{pmatrix} (\mathbf{r}_c^{(n)}, \mathbf{r}_c^{(n+k)}) \\ \vdots \\ (\mathbf{r}_c^{(n+k-1)}, \mathbf{r}_c^{(n+k)}) \end{pmatrix}.$$

Let us now express the vectors $\mathbf{r}_c^{(k,n)}$ as a ratio of determinants. We have the following theorem.

THEOREM 2.

$$\mathbf{r}_c^{(k,n)} = (-1)^{k-1} \frac{\begin{vmatrix} \mathbf{r}_c^{(n)} & \cdots & \mathbf{r}_c^{(n+k-1)} \\ (\Delta\mathbf{r}_c^{(n)}, \Delta\mathbf{r}_c^{(n)}) & \cdots & (\Delta\mathbf{r}_c^{(n)}, \Delta\mathbf{r}_c^{(n+k-1)}) \\ \vdots & & \vdots \\ (\Delta\mathbf{r}_c^{(n+k-2)}, \Delta\mathbf{r}_c^{(n)}) & \cdots & (\Delta\mathbf{r}_c^{(n+k-2)}, \Delta\mathbf{r}_c^{(n+k-1)}) \end{vmatrix}}{\begin{vmatrix} (\Delta\mathbf{r}_c^{(n)}, \Delta\mathbf{r}_c^{(n)}) & \cdots & (\Delta\mathbf{r}_c^{(n)}, \Delta\mathbf{r}_c^{(n+k-2)}) \\ \vdots & & \vdots \\ (\Delta\mathbf{r}_c^{(n+k-2)}, \Delta\mathbf{r}_c^{(n)}) & \cdots & (\Delta\mathbf{r}_c^{(n+k-2)}, \Delta\mathbf{r}_c^{(n+k-2)}) \end{vmatrix}}.$$

[illegible]. We have $\mathbf{e}^{(k,n)} = \Delta R_n \mathbf{b}$. Taking into account that $b_{k-1} = 1$, this relation can also be written as $\mathbf{e}^{(k,n)} = \Delta\widetilde{R}_n\widetilde{\mathbf{b}} + \Delta\mathbf{r}_c^{(n+k-1)}$, with $\widetilde{R}_n = [\mathbf{r}_c^{(n)}, \ldots, \mathbf{r}_c^{(n+k-2)}]$ and $\widetilde{\mathbf{b}} = (b_0, \ldots, b_{k-2})^T$. Solving, as above, the system $\mathbf{e}^{(k,n)} = 0$ in the least squares sense gives $\widetilde{\mathbf{b}} = -(\Delta\widetilde{R}_n^T\Delta\widetilde{R}_n)^{-1}\Delta\widetilde{R}_n^T\Delta\mathbf{r}_c^{(n+k-1)}$. Thus, since $\mathbf{r}_c^{(k,n)} = \widetilde{R}_n\widetilde{\mathbf{b}} + \mathbf{r}_c^{(n+k-1)}$, we get

$$\mathbf{r}_c^{(k,n)} = \mathbf{r}_c^{(n+k-1)} - \widetilde{R}_n(\Delta\widetilde{R}_n^T\Delta\widetilde{R}_n)^{-1}\Delta\widetilde{R}_n^T\Delta\mathbf{r}_c^{(n+k-1)}.$$

This relation shows that $\mathbf{r}_c^{(k,n)}$ is a Schur complement, and the result follows from Schur's determinantal formula. □

Since $A_c\mathbf{r}_c^{(k,n)} = R_{n+1}\mathbf{b}$, we immediately have the following.

COROLLARY 5. $\mathbf{e}^{(k,n)}$ [illegible] $\Delta\mathbf{r}_c^{(n)}, \ldots, \Delta\mathbf{r}_c^{(n+k-1)}$ [illegible] $(\Delta\mathbf{r}_c^{(n+i)}, \mathbf{e}^{(k,n)}) = 0$ [illegible] $i = 0, \ldots, k-2$ [illegible] $\Delta\widetilde{R}_n^T\mathbf{e}^{(k,n)} = 0$

Note that $R_n = [\widetilde{R}_n, \mathbf{r}_c^{(n+k-1)}]$, and $\mathbf{b} = (\widetilde{\mathbf{b}}^T, b_{k-1})^T$. Polynomial expressions for $\mathbf{r}_c^{(k,n)}$ and $\mathbf{e}^{(k,n)}$ similar to that of Corollary 3 can easily be deduced from Theorem 2 and Corollary 5. The preceding results can be easily modified if R_n is replaced by Z_n.

Thus, in this section, we have generalized to an arbitrary value of k the Quadratic Extrapolation presented in [31] which corresponds to $k = 3$. Moreover, it has been related to Krylov subspace methods.

In practice, the value of k is limited by the dimension p of the problem and by the number of vectors to store for computing the vector $\mathbf{r}_c^{(k,n)}$. For $k = 2$, we obtain the new vector sequence transformation

$$\mathbf{r}_c^{(2,n)} = (A_c - \alpha_n I)\mathbf{r}_c^{(n)} = \mathbf{r}_c^{(n+1)} - \alpha_n\mathbf{r}_c^{(n)} \quad \text{with} \quad \alpha_n = \frac{(\Delta\mathbf{r}_c^{(n)}, \Delta\mathbf{r}_c^{(n+1)})}{(\Delta\mathbf{r}_c^{(n)}, \Delta\mathbf{r}_c^{(n)})}.$$

This relation corresponds to the ratio of determinants given in Theorem 2.

These vector least squares extrapolation procedures follow an idea similar to that used in the least squares extrapolation discussed in [13, sect. 3.10] for scalar sequences and in the vector transformations proposed in [15].

6.2. The method of moments. The generalization of the Quadratic Extrapolation [31] discussed in the previous section could be interpreted as a special case of the method of moments of Vorobyev [50, pp. 14–16] (see also [8, pp. 154–157]). Thus, we will have a different point of view on this generalization, which is always helpful for obtaining theoretical results, such as acceleration properties.

Let $\mathbf{u}_0, \ldots, \mathbf{u}_k$ be linearly independent vectors in $\mathbb{R}^p$ and $\mathbf{z}_0, \ldots, \mathbf{z}_{k-1}$ also, where $k+1 \leq p$. The method of moments consists of constructing the linear mapping A_k on $E_k = \operatorname{span}(\mathbf{u}_0, \ldots, \mathbf{u}_{k-1})$ such that

$$\begin{aligned}
&\mathbf{u}_1 = A_k \mathbf{u}_0, \\
&\mathbf{u}_2 = A_k \mathbf{u}_1 = A_k^2 \mathbf{u}_0, \\
&\cdots\cdots\cdots\cdots\cdots \\
&\mathbf{u}_{k-1} = A_k \mathbf{u}_{k-2} = A_k^{k-1} \mathbf{u}_0, \\
&\mathbf{P}_k \mathbf{u}_k = A_k \mathbf{u}_{k-1} = A_k^k \mathbf{u}_0,
\end{aligned}$$

where $\mathbf{P}_k$ is the projection on E_k orthogonal to $F_k = \operatorname{span}(\mathbf{z}_0, \ldots, \mathbf{z}_{k-1})$.

These relations completely determine the mapping A_k. Indeed, for any $\mathbf{u} \in E_k$, there exist numbers $b_0, \ldots, b_{k-1}$ such that

$$\mathbf{u} = b_0 \mathbf{u}_0 + \cdots + b_{k-1} \mathbf{u}_{k-1}. \tag{12}$$

Thus,

$$\begin{aligned}
A_k \mathbf{u} &= b_0 A_k \mathbf{u}_0 + \cdots + b_{k-2} A_k \mathbf{u}_{k-2} + b_{k-1} A_k \mathbf{u}_{k-1} \\
&= b_0 \mathbf{u}_1 + \cdots + b_{k-2} \mathbf{u}_{k-1} + b_{k-1} \mathbf{P}_k \mathbf{u}_k \in E_k.
\end{aligned} \tag{13}$$

Since $\mathbf{P}_k \mathbf{u}_k \in E_k$, there exist numbers $a_0, \ldots, a_{k-1}$ such that

$$\mathbf{P}_k \mathbf{u}_k = -a_0 \mathbf{u}_0 - \cdots - a_{k-1} \mathbf{u}_{k-1}, \tag{14}$$

that is,

$$a_0 \mathbf{u}_0 + \cdots + a_{k-1} \mathbf{u}_{k-1} + \mathbf{P}_k \mathbf{u}_k = (a_0 + a_1 A_k + \cdots + a_{k-1} A_k^{k-1} + A_k^k) \mathbf{u}_0 = 0.$$

But

$$(\mathbf{z}_i, \mathbf{u}_k - \mathbf{P}_k \mathbf{u}_k) = 0 \quad \text{for } i = 0, \ldots, k-1,$$

that is, for $i = 0, \ldots, k-1$,

$$a_0 (\mathbf{z}_i, \mathbf{u}_0) + \cdots + a_{k-1} (\mathbf{z}_i, \mathbf{u}_{k-1}) + (\mathbf{z}_i, \mathbf{u}_k) = 0.$$

Solving this system gives the a_i's and, thus, A_k is completely determined.

Now, if we set

$$P_k(\xi) = a_0 + \cdots + a_{k-1} \xi^{k-1} + \xi^k,$$

then

$$P_k(A_k) \mathbf{u}_0 = 0,$$

which shows that P_k is an annihilating polynomial of A_k for the vector $\mathbf{u}_0$.

We will be looking for the eigenvectors of A_k belonging to E_k. Let $\mathbf{u} \in E_k$. From (13) and (14), we have

$$\begin{aligned}
A_k \mathbf{u} &= b_0 \mathbf{u}_1 + \cdots + b_{k-2} \mathbf{u}_{k-1} + b_{k-1} (-a_0 \mathbf{u}_0 - \cdots - a_{k-1} \mathbf{u}_{k-1}) \\
&= -a_0 b_{k-1} \mathbf{u}_0 + (b_0 - a_1 b_{k-1}) \mathbf{u}_1 + \cdots + (b_{k-2} - a_{k-1} b_{k-1}) \mathbf{u}_{k-1}.
\end{aligned} \tag{15}$$

If λ is an eigenvalue of A_k and $\mathbf{u}$ is the corresponding eigenvector, then

$$A_k\mathbf{u} = \lambda(b_0\mathbf{u}_0 + \cdots + b_{k-1}\mathbf{u}_{k-1}),$$

and, since $\mathbf{u}_0, \ldots, \mathbf{u}_{k-1}$ are linearly independent in E_k, we see from (15) that we must have

$$\begin{aligned} -a_0 b_{k-1} &= b_0\lambda, \\ b_i - a_{i+1}b_{k-1} &= b_{i+1}\lambda, \qquad i = 0, \ldots, k-2, \end{aligned} \tag{16}$$

that is, in matrix form,

$$\begin{pmatrix} -\lambda & 0 & \cdots & \cdots & 0 & -a_0 \\ 1 & -\lambda & \ddots & & 0 & -a_1 \\ 0 & 1 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & -\lambda & 0 & -a_{k-3} \\ \vdots & & \ddots & 1 & -\lambda & -a_{k-2} \\ 0 & \cdots & \cdots & 0 & 1 & -a_{k-1} - \lambda \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{k-3} \\ b_{k-2} \\ b_{k-1} \end{pmatrix} = 0.$$

Since this system has a nonzero solution, its determinant must be zero, that is,

$$P_k(\lambda) = 0.$$

Moreover, we must have $b_{k-1} \neq 0$, since otherwise all the b_i's would be zero. Since an eigenvector is defined up to a multiplying factor, then b_{k-1} can be arbitrarily set to 1 and, from (16), we have

$$b_i = a_{i+1} + b_{i+1}\lambda, \qquad i = k-2, \ldots, 0.$$

We see that, for $\lambda = 1$, these relations are the same as (10).

As seen above, for $\mathbf{u}$ as in (12), $A_k\mathbf{u}$ is given by (15), and the transformation mapping the coordinates $b_0, \ldots, b_{k-1}$ of $\mathbf{u}$ in the basis formed by the elements $\mathbf{u}_0, \ldots, \mathbf{u}_{k-1}$ into the coordinates of $A_k\mathbf{u}$ in the same basis is given by the matrix $\widetilde{A}_k$ of the system

$$\begin{pmatrix} 0 & \cdots & \cdots & 0 & -a_0 \\ 1 & \ddots & & \vdots & -a_1 \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & -a_{k-2} \\ 0 & \cdots & 0 & 1 & -a_{k-1} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{k-2} \\ b_{k-1} \end{pmatrix} = \begin{pmatrix} -a_0 b_{k-1} \\ b_0 - b_{k-1}a_1 \\ \vdots \\ b_{k-3} - b_{k-2}a_{k-2} \\ b_{k-2} - b_{k-1}a_{k-1} \end{pmatrix}.$$

Thus, the polynomial P_k is the characteristic polynomial of the $k \times k$ matrix $\widetilde{A}_k$ which represents the mapping A_k in E_k. Consequently, $\widetilde{A}_k$ is regular if and only if $a_0 \neq 0$, and the rank of A_k is equal to the rank of $\widetilde{A}_k$.

In the particular case where $\mathbf{u}_i = A_c^i\mathbf{u}_0$, $i = 0, 1, \ldots$, which is the case we treated, it is possible to obtain an expression for A_k. Let $\mathbf{u}$ be as in (12). Then

$$\begin{aligned} A_c\mathbf{u} &= b_0A_c\mathbf{u}_0 + \cdots + b_{k-2}A_c\mathbf{u}_{k-2} + b_{k-1}A_c\mathbf{u}_{k-1} \\ &= b_0A_c\mathbf{u}_0 + \cdots + b_{k-2}A_c^{k-1}\mathbf{u}_0 + b_{k-1}A_c^k\mathbf{u}_0 \\ &= b_0A_k\mathbf{u}_0 + \cdots + b_{k-2}A_k^{k-1}\mathbf{u}_0 + b_{k-1}A_c^k\mathbf{u}_0, \end{aligned}$$

and it follows that

$$\begin{aligned}\mathbf{P}_k A_c \mathbf{u} &= b_0 A_k \mathbf{u}_0 + \cdots + b_{k-2} A_k^{k-1} \mathbf{u}_0 + b_{k-1} \mathbf{P}_k \mathbf{u}_k \\ &= b_0 A_k \mathbf{u}_0 + \cdots + b_{k-2} A_k^{k-1} \mathbf{u}_0 + b_{k-1} A_k^k \mathbf{u}_0 \\ &= A_k (b_0 \mathbf{u}_0 + \cdots + b_{k-1} \mathbf{u}_{k-1}) = A_k \mathbf{u},\end{aligned}$$

which shows that $A_k = \mathbf{P}_k A_c$ on E_k. Since, if $\mathbf{u} \in E_k$, $\mathbf{P}_k \mathbf{u} \in E_k$, then the domain of A_k can be extended to the whole space $\mathbb{R}^p$ by setting

$$A_k = \mathbf{P}_k A_c \mathbf{P}_k.$$

Now let $\mathbf{u} \in \mathbb{R}^p$. Setting $U_k = [\mathbf{u}_0, \ldots, \mathbf{u}_{k-1}]$, $Z_k = [\mathbf{z}_0, \ldots, \mathbf{z}_{k-1}]$, and $\mathbf{a} = (a_0, \ldots, a_{k-1})^T$, the conditions $(\mathbf{P}_k \mathbf{u} - \mathbf{u}, \mathbf{z}_i) = 0$ for $i = 0, \ldots, k-1$ can be written as $Z_k^T U_k \mathbf{a} = -Z_k^T \mathbf{u}$ and it follows that $\mathbf{P}_k \mathbf{u} = -U_k \mathbf{a} = U_k (Z_k^T U_k)^{-1} Z_k^T \mathbf{u}$, which gives

$$\mathbf{P}_k = U_k (Z_k^T U_k)^{-1} Z_k^T.$$

It must be noted that A_k is not an injection since $\mathbf{P}_k$ is not.

6.3. The ϵ-algorithms. The ϵ-algorithms are sequence transformations which map a given sequence into new ones which, under certain assumptions, converge faster to the same limit. Let us now discuss the various ϵ-algorithms for vector sequences.

As above, let $(\mathbf{x}^{(n)})$ be a vector sequence converging to $\mathbf{x}$. The vector ϵ-algorithm consists in the recursive rule

$$\begin{aligned}\boldsymbol{\epsilon}_{-1}^{(n)} &= 0, \\ \boldsymbol{\epsilon}_0^{(n)} &= \mathbf{x}^{(n)}, \\ \boldsymbol{\epsilon}_{j+1}^{(n)} &= \boldsymbol{\epsilon}_{j-1}^{(n+1)} + \left[\boldsymbol{\epsilon}_j^{(n+1)} - \boldsymbol{\epsilon}_j^{(n)}\right]^{-1}\end{aligned}$$

for $j = 0, 1, \ldots$ and $n = 0, 1, \ldots$, where the inverse of a vector $\mathbf{y}$ is defined by $\mathbf{y}^{-1} = \mathbf{y}/(\mathbf{y}, \mathbf{y})$. The vectors with an odd lower index are intermediate computations without any interesting meaning, while those with an even lower index approximate $\mathbf{x}$. These rules are also valid for the scalar ϵ-algorithm (in which case the ϵ's are not in bold in what follows) with $\epsilon_0^{(n)} = (\mathbf{x}^{(n)})_i$, the ith component of $\mathbf{x}^{(n)}$. The rules of the topological ϵ-algorithm are slightly different, and can be found, for example, in [13, sect. 4.2]. The computation of $\boldsymbol{\epsilon}_{2k}^{(n)}$ needs the knowledge of $\mathbf{x}^{(n)}, \ldots, \mathbf{x}^{(n+2k)}$ and the storage of $2k+1$ vectors, thus restricting k to small values in our case.

The kernels of the scalar ϵ-algorithm (applied separately on each components), of the vector ϵ-algorithm, and of the topological ϵ-algorithm contain the set of sequences satisfying the characteristic relation

$$b_0(\mathbf{x}^{(n)} - \mathbf{x}) + \cdots + b_{k-1}(\mathbf{x}^{(n+k-1)} - \mathbf{x}) = 0, \quad n = 0, 1, \ldots, \tag{17}$$

where the b_i's are any numbers satisfying $b_0 b_{k-1} \neq 0$. Thus, if one of these ϵ-algorithms is applied to a sequence $(\mathbf{x}^{(n)})$ satisfying (17), then, by construction, $\boldsymbol{\epsilon}_{2k-2}^{(n)} = \mathbf{x}$ for $n = 0, 1, \ldots$.

Let us now study our particular case. From (5), we have $\mathbf{r}_c = Q_{m-1}(A_c)\mathbf{r}_c^{(n)}$. Moreover, since $\mathbf{r}_c = A_c^i \mathbf{r}_c$ for all i, $\sum_{i=0}^{m-1} b_i \mathbf{r}_c = \sum_{i=0}^{m-1} b_i A_c^i \mathbf{r}_c = Q_{m-1}(A_c)\mathbf{r}_c = \mathbf{r}_c$, assuming that $\sum_{i=0}^{m-1} b_i = 1$, which does not restrict the generality. Thus, subtracting the second relation from the first one, we get the following.

PROPERTY 14.

$$Q_{m-1}(A_c)(\mathbf{r}_c^{(n)} - \mathbf{r}_c) = b_0(\mathbf{r}_c^{(n)} - \mathbf{r}_c) + \cdots + b_{m-1}(\mathbf{r}_c^{(n+m-1)} - \mathbf{r}_c) = 0, \quad n = 0, 1, \ldots.$$

Thus, applying one of the ϵ-algorithms to the vector sequence $(\mathbf{r}_c^{(n)})$ gives $\boldsymbol{\epsilon}_{2m-2}^{(n)} = \mathbf{r}_c$ for $n = 0, 1, \ldots$ and produces approximations $\boldsymbol{\epsilon}_{2k-2}^{(n)}$ of $\mathbf{r}_c$ for $k < m$. Since, by the theory of the ϵ-algorithms, there exist numbers $b'_0, \ldots, b'_{k-1}$ such that

$$b'_0(\mathbf{r}_c^{(i)} - \boldsymbol{\epsilon}_{2k-2}^{(n)}) + \cdots + b'_{k-1}(\mathbf{r}_c^{(i+k-1)} - \boldsymbol{\epsilon}_{2k-2}^{(n)}) = 0, \quad i = 0, 1, \ldots,$$

then

$$\boldsymbol{\epsilon}_{2k-2}^{(n)} = Q_{k-1}(A_c)\mathbf{r}_c^{(n)}, \qquad n = 0, 1, \ldots,$$

with $Q_{k-1}(\lambda) = b'_0 + \cdots + b'_{k-1}\lambda^{k-1}$. These vectors $\boldsymbol{\epsilon}_{2k-2}^{(n)}$ are rational approximations of $\mathbf{r}_c$ in the Padé style. In the case of the topological ϵ-algorithm, it is well known that the vectors it computes can be represented as a ratio of determinants, and we have (see, for example, [13, p. 221]) the following.

THEOREM 3. [illegible] $\boldsymbol{\epsilon}_{2k-2}^{(n)} = \widetilde{Q}_{k-1}(A_c)\mathbf{r}_c^{(n)}/\widetilde{Q}_{k-1}(1)$ [illegible]

$$\widetilde{Q}_{k-1}(\lambda) = \begin{vmatrix} 1 & \cdots & \lambda^k \\ (\mathbf{y}, \Delta\mathbf{r}_c^{(n)}) & \cdots & (\mathbf{y}, \Delta\mathbf{r}_c^{(n+k-1)}) \\ \vdots & & \vdots \\ (\mathbf{y}, \Delta\mathbf{r}_c^{(n+k-2)}) & \cdots & (\mathbf{y}, \Delta\mathbf{r}_c^{(n+2k-3)}) \end{vmatrix},$$

[illegible] $\mathbf{y}$ [illegible] $\widetilde{Q}_{k-1}(1) \neq 0$

Let us now analyze the behavior of the vectors $\boldsymbol{\epsilon}_{2k-2}^{(n)}$ when k is fixed and n tends to infinity. The relation of Property 14 shows that the vectors $\mathbf{r}_c^{(n)} - \mathbf{r}_c$ satisfy a linear homogeneous difference equation of order $m-1$ with constant coefficients. In the particular case where the zeros $c\widetilde{\lambda}_2, \ldots, c\widetilde{\lambda}_m$ of Q_{m-1} (which are the eigenvalues of A_c) are real and simple, and all the eigenvectors of A_c are present in the spectral decomposition of $\mathbf{v}$, the solution of this difference equation is

$$\mathbf{r}_c^{(n)} = \mathbf{r}_c + \sum_{i=2}^{m} (c\widetilde{\lambda}_i)^n \mathbf{v}_i, \quad n = 0, 1, \ldots, \tag{18}$$

where the vectors $\mathbf{v}_i \in \mathbb{R}^p$ depend on the eigenvectors of A_c. The solution of the relation of Property 14 was studied in its full generality in [12] (see also [13, Thm. 2.18]), but it will not be reproduced here for length reasons. Let us mention only that if an eigenvalue $\widetilde{\lambda}_i$ has multiplicity k_i, then $\mathbf{v}_i$ is replaced in (18) by a polynomial of degree $k_i - 1$ with vector coefficients.

Using (18), we have the following convergence and acceleration results which support, in particular, the numerical results given in [31] for $k = 1$. They follow directly from the acceleration theorems proved by Wynn [51] for the scalar ϵ-algorithm and by Matos for the vector ϵ-algorithm [39]

THEOREM 4. [illegible] A_c [illegible] A_c [illegible] $\mathbf{v}$ [illegible] $1 \leq k \leq m-1$

$$\|\boldsymbol{\epsilon}_{2k}^{(n)} - \mathbf{r}_c\|_2 = \mathcal{O}((c\widetilde{\lambda}_{k+2})^n),$$

$$\lim_{n\to\infty} \frac{\|\boldsymbol{\epsilon}_{2k}^{(n)} - \mathbf{r}_c\|_2}{\|\boldsymbol{\epsilon}_{2k-2}^{(n)} - \mathbf{r}_c\|_2} = 0.$$

If some of the eigenvalues of A_c are multiple, they have to be counted according to their multiplicity, and the polynomial factors in the solution of the relation of Property 14 come into the discussion. However, the theory and the results remain essentially the same (in particular Theorem 4), but they become more complicated to write down (see, for example, Theorem 5 of [39]). These results are also valid for the topological ϵ-algorithm.

Other approximations of the Padé style are the vectors $\mathbf{E}_{k-1}^{(n)}$ computed by the scalar E-algorithm (applied componentwise) or the vector E-algorithm [7] (see also [13, pp. 55–72, 228–232]). Applying the convergence and acceleration results proved in [7, 45, 38], conclusions similar to those of Theorem 4 can be obtained. Let us also mention that the ϵ-algorithm and E-algorithm are related to Schur complements [52, pp. 233–238].

For the scalar ϵ-algorithm, when $k = 2$, the well-known Aitken's Δ^2 process is recovered. The kernel of Aitken's process is the set of scalar sequences $(x^{(n)})$ satisfying

$$b_0(x^{(n)} - x) + b_1(x^{(n+1)} - x) = 0, \qquad n = 0, 1, \ldots,$$

with $b_0 + b_1 \neq 0$, or, equivalently,

$$x^{(n)} = x + \alpha\mu^n, \qquad n = 0, 1, \ldots,$$

with $\mu \neq 1$. Note that the form of the first relation is the same as (17) when $k = 2$.

Aitken's Δ^2 process can be written in different ways. For example, we have the three following equivalent formulae:

$$\epsilon_2^{(n)} = x^{(n)} - \frac{(x^{(n+1)} - x^{(n)})^2}{x^{(n+2)} - 2x^{(n+1)} + x^{(n)}} \tag{19}$$

$$= x^{(n+1)} - \frac{(x^{(n+2)} - x^{(n+1)})(x^{(n+1)} - x^{(n)})}{x^{(n+2)} - 2x^{(n+1)} + x^{(n)}} \tag{20}$$

$$= x^{(n+2)} - \frac{(x^{(n+2)} - x^{(n+1)})^2}{x^{(n+2)} - 2x^{(n+1)} + x^{(n)}}. \tag{21}$$

If each component of the vectors $\mathbf{r}_c^{(n)}$ successively plays the role of $x^{(n)}$, then (19) is exactly the Aitken Extrapolation given by formula (15) of [31], while (20) corresponds to the Epsilon Extrapolation of [31]. Another implementation of the same extrapolation method can be obtained by using (21). However, let us mention that, although these are completely equivalent from the mathematical point of view, the numerical stability of these formulae can be quite different. It is well known that Aitken's process accelerates the convergence of sequences such that $\exists\delta \neq 1$, $\lim_{n\to\infty}(x^{(n+1)}-x)/(x^{(n)}-x) = \delta$, which, by Property 11, is exactly our case with $\delta = c\widetilde{\lambda}_2$. Thus, the effectiveness of the methods proposed in [31] is justified by the preceding discussion and by Theorem 4.

Each of the scalars $\epsilon_2^{(n)}$ produced by Aitken's process applied separately on each component has a different denominator. On the contrary, using the vector or the topological ϵ-algorithm for transforming the vectors $\mathbf{r}_c^{(n)}$ will lead to vectors $\boldsymbol{\epsilon}_2^{(n)}$ with the same denominator for each component, and thus will be more similar to the exact form of $\mathbf{r}_c$.

Let us recall that the ϵ-algorithms are related to various Padé-style approximants [6, 9]. If the scalar ϵ-algorithm is applied to the partial sums of a formal power series f with scalar coefficients, then the quantities $\epsilon_{2k}^{(n)}$ it computes are the Padé approximants $[n+k/k]_f$ of f. Reciprocally, the quantities $\epsilon_{2k}^{(n)}$ given by this algorithm

with $\epsilon_0^{(n)} = x^{(n)}$ are the $[n+k/k]_f$ Padé approximants of the series $f(\xi) = x^{(0)}+(x^{(1)}-x^{(0)})\xi+(x^{(2)}-x^{(1)})\xi^2+\cdots$. In particular, the $\epsilon_2^{(n)}$'s computed by Aitken's process are identical to its Padé approximants $[n/1]_f(1)$. Thus, applying the scalar ϵ-algorithm separately on each component of a series with vector coefficients, as in [31], produces Padé approximants with, in general, a different denominator for each component, while, in our case, all denominators should be identical by Properties 3 and 4. On the contrary, the vector and the topological ϵ-algorithms provide rational approximations of the series (2), but with the same denominator for all components. Thus, they seem to be better adapted to the acceleration of the power method. Moreover, it would also be interesting to consider the approximants $[k-1/k-1]$ for increasing values of k instead of the approximants $[n/1]$. It is possible to use the vector E-algorithm, which also leads to rational vector approximations of $\mathbf{r}_c$ with a unique denominator for all components, and allows more flexibility by an arbitrary choice of auxiliary vector sequences; see [13, sect. 4.3].

6.4. Other algorithms. Of course, the acceleration procedures studied above are not the only possible ones. Among them, there exist several other acceleration methods whose kernel is the set of sequences satisfying (17), where the vectors $\mathbf{x}^{(n)}$ are those obtained by the power method, and $\mathbf{x}$ is the vector $\mathbf{r}_c$ we are looking for. In this section, we will briefly review some of them, since they probably are those having the best acceleration properties, as explained at the beginning of section 6.

The relation of Property 14 can also be written as

$$\mathbf{r}_c^{(n)} = \mathbf{r}_c^{(k,n)} + \alpha_0\Delta\mathbf{r}_c^{(n)} + \cdots + \alpha_{k-2}\Delta\mathbf{r}_c^{(n+k-2)}, \quad n = 0, 1, \ldots,$$

that is,

$$\mathbf{r}_c^{(n)} = \mathbf{r}_c^{(k,n)} + \Delta\widetilde{R}_n\boldsymbol{\alpha}, \tag{22}$$

with $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_{k-2})^T$ and $\widetilde{R}_n = [\mathbf{r}_c^{(n)}, \ldots, \mathbf{r}_c^{(n+k-2)}]$, as in section 6.1.

There are several ways to compute this vector $\boldsymbol{\alpha}$. One of them leads to a sequence transformation named the [illegible] (VTT) defined in [14] by

$$\mathbf{r}_c^{(k,n)} = \mathbf{r}_c^{(n)} - \Delta\widetilde{R}_n(Z_n^T\Delta^2\widetilde{R}_n)^{-1}Z_n^T\Delta^2\mathbf{r}_c^{(n)}, \quad n = 0, 1, \ldots, \quad k > 2,$$

where $Z_n \in \mathbb{R}^{p\times(k-1)}$.

The [illegible] (BVTT) transformation is a particular case of the VTT, with the same kernel [14].

A general methodology, based on various strategies, for constructing sequence transformations whose kernel contains sequences of the form (22) is described in [10]. These transformations can be implemented either by one of the ϵ-algorithms given in the preceding section, by the RPA [13, sect. 4.4], or by the $S\beta$-algorithm [27]. The case where the matrix to be inverted is singular is treated similarly in [15] by using pseudoinverses and pseudo-Schur complements, whose properties are studied in [43]. Another vector sequence transformation related to the method of moments is the [illegible] (MMPE) of Pugachev [42]. Application of other vector extrapolation methods, such as the RRE and the MPE, to PageRank computations are discussed in [46], but numerical experiments have yet to be carried out.

7. Conclusions. In this paper, we analyzed the PageRank problem and its solution by the power method. Several procedures for accelerating the convergence of its iterates were proposed, and some theoretical results were given. However, no results for comparing these algorithms exist so far. When the parameter k in these acceleration methods increases, in general their efficiency increases, but the number of vectors to store also increases, thus putting a restriction on their practical use due to the huge dimension of the problem. Moreover, the behaviors of these algorithms are quite similar, and the choice between them is, more or less, a matter of taste. Thus, extensive numerical experiments have to be carried out, and perhaps they could help in making this choice.

Let us mention another problem related to PageRank computations. When c approaches 1 (which corresponds to the real PageRank vector), Property 12 shows that the speed of convergence of the power method reduces, and, moreover, the matrix A_c becomes more and more ill conditioned (as proved in [28], its condition number behaves as $(1-c)^{-1}$), the conditioning of the eigenproblem becomes poor, and $\mathbf{r}_c$ cannot be computed accurately. So, to avoid these drawbacks, $\mathbf{r}_c$ can be computed for several values of c far away from 1 by any procedure, and then these vectors can be extrapolated at the point $c = 1$ (or at any other point). In order for an extrapolation procedure to work well, the extrapolating function has to mimic as closely as possible the behavior of $\mathbf{r}_c$ with respect to the parameter c. Extrapolation algorithms based on the analysis of this dependence, given in [44], are described in [17]; see [16] for more developments.

Acknowledgments. We are grateful to Stefano Serra-Capizzano for pointing out the problem of accelerating PageRank computations and the discussions we had with him. We would like to thank Gene Golub for reading a first draft of the manuscript and suggesting additional references. We appreciated several valuable and constructive comments by Daniel Szyld and are grateful to the referees for their comments, which greatly helped us look deeper into the problem. Finally, we would like to thank Valeria Simoncini, who asked us to expand the first version of this paper for making it also a survey.

REFERENCES

[1] A. Arasu, J. Novak, A. Tomkins, and J. Tomlin, *PageRank computation and the structure of the Web: Experiments and algorithms*, in Proceedings of the Eleventh International World Wide Web Conference, ACM Press, New York, 2002; available online from http://www2002.org/CDROM/poster/173.pdf.

[2] P. Berkhin, *A survey on PageRank computing*, Internet Math., 2 (2005), pp. 73–120.

[3] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, SIAM, Philadelphia, 1994.

[4] M. Bianchini, M. Gori, and F. Scarselli, *Inside Pagerank*, ACM Trans. Internet Technol., 5 (2005), pp. 92–128.

[5] P. Boldi, M. Santini, and S. Vigna, *PageRank as a function of the damping factor*, in Proceedings of the Fourteenth International World Wide Web Conference, ACM Press, New York, 2005, pp. 557–566.

[6] C. Brezinski, *Padé-Type Approximation and General Orthogonal Polynomials*, Internat. Ser. Numer. Math. 50, Birkhäuser Verlag, Basel, 1980.

[7] C. Brezinski, *A general extrapolation algorithm*, Numer. Math., 35 (1980), pp. 175–187.

[8] C. Brezinski, *Projections Methods for Systems of Equations*, North-Holland, Amsterdam, 1997.

[9] C. Brezinski, *Computational Aspects of Linear Control*, Kluwer, Dordrecht, 2002.

[10] C. Brezinski, *Vector sequence transformations: Methodology and applications to linear systems*, J. Comput. Appl. Math., 98 (1998), pp. 149–175.

[11] C. Brezinski, *Biorthogonal vector sequence transformations and Padé approximation of vector series*, Appl. Numer. Math., 41 (2002), pp. 437–442.
[12] C. Brezinski and M. Crouzeix, *Remarques sur le procédé Δ^2 d'Aitken*, C. R. Acad. Sci. Paris Sér. A-B, 270 (1970), pp. 896–898.
[13] C. Brezinski and M. Redivo-Zaglia, *Extrapolation Methods. Theory and Practice*, North-Holland, Amsterdam, 1991.
[14] C. Brezinski and M. Redivo-Zaglia, *Vector and matrix sequence transformations based on biorthogonality*, Appl. Numer. Math., 21 (1996), pp. 353–373.
[15] C. Brezinski and M. Redivo-Zaglia, *New vector sequence transformations*, Linear Algebra Appl., 389 (2004), pp. 189–213.
[16] C. Brezinski and M. Redivo-Zaglia, *Rational Extrapolation of the PageRank Vectors*, in progress.
[17] C. Brezinski, M. Redivo-Zaglia, and S. Serra-Capizzano, *Extrapolation methods for PageRank computations*, C. R. Math. Acad. Sci. Paris, 340 (2005), pp. 393–397.
[18] S. Brin and L. Page, *The anatomy of a large-scale hypertextual web search engine*, Comput. Networks ISDN Syst., 30 (1998), pp. 107–117.
[19] G. M. Del Corso, A. Gulli, and F. Romani, *Fast PageRank computation via a sparse linear system*, in Algorithms and Models for the Web-Graph, S. Leonardi, ed., Lecture Notes in Comput. Sci. 3243, Springer-Verlag, New York, 2004, pp. 118–130.
[20] L. Eldén, *The Eigenvalues of the Google Matrix*, Report LiTH-MAT-R-04-01, Department of Mathematics, Linköping University, Sweden, 2003.
[21] F. R. Gantmacher, *The Theory of Matrices*, Vol. 1, Chelsea, New York, 1959.
[22] D. Gleich, L. Zhukov, and P. Berkhin, *Fast Parallel PageRank: A Linear System Approach*, Technical report, Yahoo! Inc., 2004.
[23] G. H. Golub and C. Greif, *Arnoldi-Type Algorithms for Computing Stationary Distribution Vectors, with Application to PageRank*, Technical report SCCM-04-15, Stanford University, Stanford, CA, 2004.
[24] T. H. Haveliwala and S. D. Kamvar, *The Second Eigenvalue of the Google Matrix*, Technical report, Stanford University, Stanford, CA, 2003; available online from http://dbpubs.stanford.edu/pub/2003-20.
[25] I. C. F. Ipsen and S. Kirkland, *Convergence analysis of a PageRank updating algorithm by Langville and Meyer*, SIAM J. Matrix Anal. Appl., 27 (2006), pp. 952–967.
[26] I. C. F. Ipsen and R. S. Wills, *Mathematical properties and analysis of Google's PageRank*, Bol. Soc. Esp. Mat. Apl., 34 (2006), pp. 191–196.
[27] K. Jbilou, *A general projection algorithm for solving systems of linear equations*, Numer. Algorithms, 4 (1993), pp. 361–377.
[28] S. D. Kamvar and T. H. Haveliwala, *The Condition Number of the PageRank Problem*, Technical report, Stanford University, Stanford, CA, 2003; available online from http://dbpubs.stanford.edu/pub/2003-36.
[29] S. Kamvar, T. Haveliwala, and G. Golub, *Adaptive methods for the computation of PageRank*, Linear Algebra Appl., 386 (2004), pp. 51–65.
[30] S. D. Kamvar, T. H. Haveliwala, C. D. Manning, and G. H. Golub, *Exploiting the Block Structure of the Web for Computing PageRank*, Technical report, Stanford University, Stanford, CA, 2003; available online from http://dbpubs.stanford.edu/pub/2003-17.
[31] S. D. Kamvar, T. H. Haveliwala, C. D. Manning, and G. H. Golub, *Extrapolations methods for accelerating PageRank computations*, in Proceedings of the Twelfth International World Wide Web Conference, ACM Press, New York, 2003, pp. 261–270.
[32] G. Kollias, E. Gallopoulos, and D. B. Szyld, *Asynchronous iterative computations with Web information retrieval structures: The PageRank case*, in Proceedings of the Conference on Parallel Computing 2005, Malaga, Spain, 2005.
[33] A. N. Langville and C. D. Meyer, *The use of linear algebra by web search engines*, IMAGE Newsletter, 33 (2004), pp. 2–6.
[34] A. N. Langville and C. D. Meyer, *Deeper inside PageRank*, Internet Math., 1 (2005), pp. 335–400.
[35] A. N. Langville and C. D. Meyer, *Updating PageRank with iterative aggregation*, in Proceedings of the Thirteenth World Wide Web Conference, ACM Press, New York, 2004, pp. 392–393.
[36] A. N. Langville and C. D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, Princeton, NJ, 2006.
[37] C. P.-C. Lee, G. H. Golub, and S. A. Zenios, *A Fast Two-Stage Algorithm for Computing PageRank and Its Extensions*, Technical report SCCM-2003-15, Scientific Computation and Computational Mathematics, Stanford University, Stanford, CA, 2003; available online from http://www-sccm.stanford.edu/nf-publications-tech.html.

[38] A. C. Matos, *Acceleration results for the vector E-algorithm*, Numer. Algorithms, 1 (1991), pp. 237–260.
[39] A. C. Matos, *Convergence and acceleration properties for the vector ϵ-algorithm*, Numer. Algorithms, 3 (1992), pp. 313–320.
[40] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*, SIAM, Philadelphia, 2000.
[41] L. Page, S. Brin, R. Motwani, and T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web*, Technical report, Stanford University, Stanford, CA, 1999; available online from http://dbpubs.stanford.edu/pub/1999-66.
[42] B. P. Pugachev, *Acceleration of convergence of iterative processes and a method of solving systems of nonlinear equations*, USSR Comput. Maths. Maths. Phys., 17 (1978), pp. 199–207.
[43] M. Redivo-Zaglia, *Pseudo-Schur complements and their properties*, Appl. Numer. Math., 50 (2004), pp. 511–519.
[44] S. Serra-Capizzano, *Jordan canonical form of the Google matrix: A potential contribution to the PageRank computation*, SIAM J. Matrix Anal. Appl., 27 (2005), pp. 305–312.
[45] A. Sidi, *On a generalization of the Richardson extrapolation process*, Numer. Math., 57 (1990), pp. 365–377.
[46] A. Sidi, *Approximation of Largest Eigenpairs of Matrices and Applications to PageRank Computation*, Technical report CS-2004-16, Computer Science Department, Technion, Haifa, Israel, 2004.
[47] W. J. Stewart, *Introduction to the Numerical Solution of Markov Chains*, Princeton University Press, Princeton, NJ, 1994.
[48] J. Van Iseghem, *Vector Padé approximants*, in Numerical Mathematics and Applications, R. Vichnevetsky and J. Vignes, eds., North-Holland, Amsterdam, 1986, pp. 73–77.
[49] R. S. Varga, *Matrix Iterative Analysis*, 2nd ed., Springer, New York, 2000.
[50] Yu. V. Vorobyev, *Method of Moments in Applied Mathematics*, Gordon and Breach, New York, 1965.
[51] P. Wynn, *On the convergence and stability of the epsilon algorithm*, SIAM J. Numer. Anal., 3 (1966), pp. 91–122.
[52] F.-Z. Zhang, ed., *The Schur Complement and Its Applications*, Springer, New York, 2005.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 2, pp. 576–595

STABLE FACTORIZATIONS OF SYMMETRIC TRIDIAGONAL AND TRIADIC MATRICES*

HAW-REN FANG† AND DIANNE P. O'LEARY†

Abstract. We call a matrix triadic if it has no more than two nonzero off-diagonal elements in any column. A symmetric tridiagonal matrix is a special case. In this paper we consider LXL^T factorizations of symmetric triadic matrices, where L is unit lower triangular and X is diagonal, block diagonal with 1×1 and 2×2 blocks, or the identity with L lower triangular. We prove that with diagonal pivoting, the LXL^T factorization of a symmetric triadic matrix is sparse, study some pivoting algorithms, discuss their growth factor and performance, analyze their stability, and develop perturbation bounds. These factorizations are useful in computing inertia, in solving linear systems of equations, and in determining modified Newton search directions.

Key words. matrix factorizations, tridiagonal matrices, pivoting, Cholesky decomposition

AMS subject classifications. 65F05, 65F50, 15A23

DOI. 10.1137/050636280

1. Introduction.

A symmetric matrix $A \in R^{n \times n}$ can be factored in the form LXL^T in several ways:

1. LL^T with L lower triangular and X the identity.
2. LDL^T with L unit lower triangular and X diagonal.
3. LBL^T with L unit lower triangular and X block diagonal with block order 1 or 2.

These LXL^T factorizations can be used to solve linear systems [1, 3, 4, 6], to determine a downhill search direction in modified Newton methods [10, 11], and to compute the inertia of a matrix [4].

Not all symmetric matrices have LDL^T factorizations. We allow diagonal pivoting and factor PAP^T, where P is a permutation matrix. With diagonal pivoting, we can ensure the existence of an LBL^T factorization of any symmetric matrix and the existence of an LDL^T factorization if A is positive semidefinite or diagonally dominant. Diagonal pivoting is also used to improve numerical stability of the LBL^T factorization when A is indefinite [1, 3, 4, 6]. Interchanging rows and columns can ruin the sparsity of LXL^T factorizations of band matrices, so for tridiagonal matrices, attempts have been made to develop stable algorithms that do not require interchanges [3, 5, 14].

In this paper, we study the sparsity and stability of LXL^T factorizations for a class of symmetric matrices called triadic. A matrix A is if the number of nonzero off-diagonal elements in each column is bounded by 2. Tridiagonal matrices are a special case of these, but other matrices, such as block diagonal matrices with full 3×3 blocks, and matrices that are tridiagonal except for entries in each corner, are also triadic. These latter matrices arise in the solution of differential equations with periodic boundary conditions.

*Received by the editors July 18, 2005; accepted for publication (in revised form) by P. Benner February 8, 2006; published electronically August 16, 2006. This work was supported by the National Science Foundation under grant CCR 02-04084.

http://www.siam.org/journals/simax/28-2/63628.html

†Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, A. V. Williams Building, College Park, MD 20742 (hrfang@cs.umd.edu, oleary@cs.umd.edu).

In section 2 we show that LXL^T factorizations of a symmetric triadic matrix using diagonal pivoting remain sparse. Section 3 reviews various diagonal pivoting strategies for symmetric matrices, and they are applied to triadic matrices in section 4. In section 5 the perturbation analysis of these factorizations is discussed. Section 6 gives conclusions. A rounding error analysis for these factorizations is given in [8], which also includes analysis when A is rank-deficient.

One application of LXL^T factorizations of triadic matrices is in modified Cholesky algorithms to safeguard the Newton method. Modified Cholesky algorithms replace the Hessian matrix A by $A+E$, for a suitably chosen perturbation matrix E, in order to ensure that we are factoring a positive definite matrix and therefore computing a downhill search direction. In a subsequent paper, we will discuss the usefulness of triadic matrices in such algorithms [9].

2. Diagonal pivoting in LXL^T factorization preserves triadic structure. In this section, we show that diagonal pivoting preserves sparsity in the LXL^T factorizations of symmetric triadic matrices. This is a consequence of the property that for any permutation matrix P, PAP^T is symmetric triadic if and only if A is symmetric triadic.

First we consider the sparsity of LDL^T (and thus LL^T) factorizations. The following lemma on the structure of the Schur complements leads to the desired result. We define e_k to be the column vector that is zero except for a 1 in its kth position.

LEMMA 2.1. $A = \left[\begin{smallmatrix} a_{11} & c_1^T \\ c_1 & A_{22} \end{smallmatrix}\right]$ $a_{11} \neq 0$ $\bar{A} = A_{22} - c_1 c_1^T / a_{11}$.

. Since A is triadic, c_1 has at most two nonzero elements. We denote them by $c_{i1} = \xi$ and $c_{j1} = \eta$. The matrix A_{22} is also triadic and its ith and jth rows have at most one off-diagonal element each. Moreover,

$$c_1 c_1^T = \xi^2 e_i e_i^T + \xi\eta(e_i e_j^T + e_j e_i^T) + \eta^2 e_j e_j^T$$

has at most four nonzero elements. Two of these are on the diagonal, and the others are in positions (i,j) and (j,i). Thus the sum of A_{22} and $-c_1 c_1^T/a_{11}$ is triadic. □

THEOREM 2.2. LDL^T L .

. The proof is by finite induction. At the kth step, assume that the remaining $(n-k+1)\times(n-k+1)$ matrix $\bar{A}$ is symmetric triadic. Then the next column of L is computed as c_1/a_{11}, where

$$\bar{A} = \begin{bmatrix} a_{11} & c_1^T \\ c_1 & A_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ c_1/a_{11} & I \end{bmatrix} \begin{bmatrix} a_{11} & 0 \\ 0 & \tilde{A} \end{bmatrix} \begin{bmatrix} 1 & c_1^T/a_{11} \\ 0 & I \end{bmatrix},$$

and $\tilde{A} = A_{22} - c_1 c_1^T/a_{11}$ is the Schur complement of $\bar{A}$. Notice that c_1 has at most two elements. By Lemma 2.1, the matrix $\tilde{A}$, which becomes $\bar{A}$ for the next iteration, is triadic, so we can continue the induction. □

Now we establish the same result for the LBL^T factorization. The algorithm for LBL^T factorization is the same as for LDL^T factorization with diagonal pivoting, except when all diagonal elements of the Schur complement are zeros. In such a case, we diagonally pivot some nonzero off-diagonal element in the lower triangular part to be at the second row and first column in the Schur complement and pivot with respect to the 2×2 block. This decomposition can be used to control element growth for numerical stability, even if we find a nonzero diagonal element [1, 3, 4, 6].

LEMMA 2.3. $A = \begin{bmatrix} A_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix}$ $A_{11} = \begin{bmatrix} \sigma_1 & a \\ a & \sigma_2 \end{bmatrix}$ $a \neq 0$ $\det(A_{11}) \neq 0$ $\bar{A} = A_{22} - A_{21}A_{11}^{-1}A_{21}^T$ 2×2 A_{11} . Since $\det(A_{11}) \neq 0$, $A_{11}^{-1} = \frac{1}{\det(A_{11})}\begin{bmatrix} \sigma_2 & -a \\ -a & \sigma_1 \end{bmatrix}$. Since A has at most two nonzero off-diagonal elements in each column and A_{11} already has one nonzero off-diagonal element in each column, A_{21} has at most one nonzero element in each column, so we denote it as $A_{21} = \begin{bmatrix} \xi e_i & \eta e_j \end{bmatrix}$. Then

$$\begin{aligned} A_{21}A_{11}^{-1}A_{21}^T &= \frac{1}{\det(A_{11})} \begin{bmatrix} \xi e_i & \eta e_j \end{bmatrix} \begin{bmatrix} \sigma_2 & -a \\ -a & \sigma_1 \end{bmatrix} \begin{bmatrix} \xi e_i^T \\ \eta e_j^T \end{bmatrix} \\ &= \frac{1}{\det(A_{11})}(\sigma_2\xi^2 e_i e_i^T - a\xi\eta e_j e_i^T + \sigma_1\eta^2 e_j e_j^T - a\xi\eta e_i e_j^T). \end{aligned}$$

Thus the only two off-diagonal elements of this matrix are in positions (i, j) and (j, i). Since A is triadic, A_{22} has at most one nonzero element in each of ith and jth rows, so the sum of A_{22} and $A_{21}A_{11}^{-1}A_{21}^T$ is triadic. □

THEOREM 2.4. LBL^T L . Again the proof is by finite induction. At the kth step, assume that the remaining matrix $\bar{A}$ is triadic. If the next pivot is 1×1, then Lemma 2.1 and the argument in the proof of Theorem 2.2 show that the next column of L is triadic, as is the new remaining matrix. If the next pivot is 2×2, then the factorization produces

$$\bar{A} = \begin{bmatrix} A_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} I_2 & 0 \\ A_{21}A_{11}^{-1} & I_{k-2} \end{bmatrix} \begin{bmatrix} A_{11} & 0 \\ 0 & \tilde{A} \end{bmatrix} \begin{bmatrix} I_2 & A_{11}^{-T}A_{21}^T \\ 0 & I_{k-2} \end{bmatrix}.$$

The off-diagonal part of the two new columns of L is

$$\begin{aligned} A_{21}A_{11}^{-1} &= \frac{1}{\det(A_{11})} \begin{bmatrix} \xi e_i & \eta e_j \end{bmatrix} \begin{bmatrix} \sigma_2 & -a \\ -a & \sigma_1 \end{bmatrix} \\ &= \frac{1}{\det(A_{11})} \begin{bmatrix} \sigma_2\xi e_i - a\eta e_j & -a\xi e_i + \sigma_1\eta e_j \end{bmatrix}, \end{aligned}$$

which is also triadic, and Lemma 2.3 shows that $\tilde{A}$ is triadic, so the induction can be continued. □

Combining these results with the fact that the triadic property of a matrix is preserved under symmetric permutation, we see that the number of nonzero elements is $O(n)$ in all of these factorizations if diagonal pivoting is used. More precisely, by Lemmas 2.1 and 2.3, at most $n - 2$ off-diagonal fill entries can occur.

THEOREM 2.5. LXL^T L .

Although the columns of L are sparse, the number of nonzero elements in each row of L is bounded only by n; if A is tridiagonal, for example, and

$$\tilde{Z} = \begin{bmatrix} 0 & & & 1 \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}$$

is the circular shift-down matrix, then the last row of L in the factorization $\tilde{Z}^T A\tilde{Z} = LDL^T$ is generally full.

3. Diagonal pivoting strategies for symmetric indefinite matrices. If the symmetric matrix $A \in R^{n\times n}$ is positive semidefinite [7], [15, section 10.3] or diagonally dominant [7], [13, section 9.5] (i.e., $|a_{ii}| \geq \sum_{j\neq i} |a_{ij}|$ for $i = 1, \ldots, n$), then the largest magnitude element will appear on the diagonal. Each Schur complement inherits the property of positive semidefiniteness or diagonal dominance. Therefore, in either case, the elements of L in the LDL^T factorization are bounded in magnitude by 1 with pivoting. With or without pivoting, the growth factor for D is $\rho(A) = 1$ if A is symmetric positive semidefinite, and $\rho(A) \leq 2$ if A is diagonally dominant, where $\rho(A)$ is the ratio of the largest magnitude element in the Schur complements to the largest magnitude element in A.

We would like to compute factorizations of symmetric indefinite matrices that also give bounds on the elements of L and B. In order to do this, it is necessary to pivot. There are three kinds of pivoting strategies in the literature: Bunch–Parlett [6] (complete pivoting); fast Bunch–Parlett and bounded Bunch–Kaufman [1] (rook pivoting); and Bunch–Kaufman [4] (partial pivoting). For full matrices, complete pivoting requires $O(n^3)$ comparisons, partial pivoting requires $O(n^2)$, and the cost of rook pivoting varies between $O(n^2)$ and $O(n^3)$. Therefore, it is important to uncover the advantages of the more expensive strategies. We consider each strategy in turn, applying each to the current Schur complement matrix A, noting that each depends on a preset constant $0 < \alpha < 1$.

3.1. Complete pivoting. Bunch and Parlett [6] devised the pivoting strategy presented in Algorithm 1.

Algorithm 1 Bunch–Parlett pivot selection.

Let a_{kk} be the largest magnitude diagonal element.
Let a_{ij} $(i < j)$ be the largest magnitude off-diagonal element.
if $|a_{kk}| \geq \alpha |a_{ij}|$ **then**
  Use a_{kk} as a 1×1 pivot.
else
  Use $\begin{bmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{bmatrix}$ as a 2×2 block pivot.
end if

The process continues until $a_{kk} = a_{ij} = 0$ or the factorization completes. The resulting pivot satisfies the following [illegible]:

1. If a 1×1 pivot a_{kk} is chosen, then $|a_{kk}| \geq \alpha |a_{pk}|$ for $p \neq k$.
2. If a 2×2 block pivot $\begin{bmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{bmatrix}$ is chosen, then each of the 1×1 pivots a_{ii} and a_{jj} satisfy $|a_{ii}| < \alpha |a_{ij}|$ and $|a_{jj}| < \alpha |a_{ij}|$, and a_{ij} is the element of maximum magnitude in both column i and column j.

For any algorithm satisfying the strong condition, the elements in L are bounded and the element growth in B during the factorization is well controlled, as we will show in section 3.5.

3.2. Rook pivoting. The cost for finding a pivot satisfying the strong condition can be reduced by the iterative process in Algorithm 2.

If the initial pivot index $i = 1$, this is called a [illegible] pivot selection, while if a_{ii} is the maximal magnitude diagonal element, it is called a [illegible] pivot selection [1]. Note that for a fast Bunch–Parlett selection, we do not need to test whether a_{jj} is a 1×1 pivot, because if the initial maximum magnitude diagonal element a_{ii} failed to be a pivot at the beginning, $|a_{jj}|$ is at most $|a_{ii}|$, and $|a_{ij}|$ is increasing in the loop.

Algorithm 2 Pivot selection by rook pivoting, given an initial pivot index i.

Find the index $j \neq i$ such that $|a_{ji}| = \max_{p \neq i} |a_{pi}|$.
if $|a_{ii}| \geq \alpha |a_{ji}|$ **then**
  Use a_{ii} as a 1×1 pivot.
else
  Find the index $k \neq j$ such that $|a_{kj}| = \max_{p \neq j} |a_{pj}|$.
  repeat
    if $|a_{jj}| \geq \alpha |a_{kj}|$ **then**
      Use a_{jj} as a 1×1 pivot.
    else if $|a_{ij}| = |a_{kj}|$ **then**
      Use $\begin{bmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{bmatrix}$ as a 2×2 pivot.
    else
      Set $i := j$ and $j := k$.
      Find index $k \neq j$ such that $|a_{kj}| = \max_{p \neq j} |a_{pj}|$.
    end if
  until a pivot is chosen.
end if

3.3. Partial pivoting. Bunch and Kaufman [4] devised the efficient pivoting strategy shown in Algorithm 3.

Algorithm 3 Bunch–Kaufman pivot selection, given an initial pivot index i.

Find the index $j \neq i$ such that $|a_{ji}| = \max_{p \neq i} |a_{pi}| =: \lambda$.
if $|a_{ii}| \geq \alpha \lambda$ **then**
  Use a_{ii} as a 1×1 pivot.
else
  Compute $\sigma := \max_{p \neq j} |a_{pj}| \geq \lambda$.
  if $|a_{ii}| \sigma \geq \alpha \lambda^2$ **then**
    Use a_{ii} as a 1×1 pivot.
  else if $|a_{jj}| \geq \alpha \sigma$ **then**
    Use a_{jj} as a 1×1 pivot.
  else
    Use $\begin{bmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{bmatrix}$ as a 2×2 pivot.
  end if
end if

Bunch–Kaufman pivoting does not guarantee the strong condition, but satisfies the following weak condition:

1. If a 1×1 pivot a_{kk} is chosen, then
   - $|a_{kk}| \max_{p \neq q} \{|a_{pq}| : (a_{qk} \neq 0 \text{ or } q = k)\} \geq \alpha \max_{p \neq k} |a_{pk}|^2$.
2. If a 2×2 block pivot $\begin{bmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{bmatrix}$ is chosen, then
   - $|a_{ii}| < \alpha \lambda$,
   - $|a_{ii}| \sigma < \alpha \lambda^2$,
   - $|a_{jj}| < \alpha \sigma$,

   where $\lambda = \max_{k \neq i} |a_{ki}|$ and $\sigma = \max_{k \neq j} |a_{kj}|$.

We compare the weak condition with the strong condition. For 1×1 pivots, $\max\{|a_{pq}| : p \neq q \text{ and } (a_{qk} \neq 0 \text{ or } q = k)\} \geq \max_{p \neq k} |a_{pk}|$ so the strong condition guarantees the weak condition. For 2×2 block pivots, the weak condition meets the

strong condition if $\sigma = \lambda$. We conclude that the strong condition implies the weak condition.

The natural choice of the initial pivot index i in Algorithm 3 is $i = 1$, which achieves the least cost to satisfy the weak condition [4].

Ashcraft, Grimes, and Lewis [1] argued that a bounded L can improve stability. We can improve the probability that the Bunch–Kaufman algorithm has a bounded L by choosing the largest magnitude diagonal entry as the search starting point at each pivot step [4]. The additional number of comparisons is $\frac{n^2}{2} + O(n)$, so the total comparison count remains $O(n^2)$. By making this change, we usually find a 1×1 pivot at the very first test at each step of pivot selection. The strong condition usually holds, but it is not guaranteed, as shown in the following example [13]:

$$A = \begin{bmatrix} \epsilon^2 & \epsilon & \epsilon \\ \epsilon & 0 & 1 \\ \epsilon & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & & \\ \frac{1}{\epsilon} & 1 & \\ \frac{1}{\epsilon} & 0 & 1 \end{bmatrix} \begin{bmatrix} \epsilon^2 & & \\ & -1 & \\ & & -1 \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{\epsilon} & \frac{1}{\epsilon} \\ & 1 & 0 \\ & & 1 \end{bmatrix} = LBL^T,$$

where L is unbounded as $\epsilon \to 0$.

3.4. The weak condition controls the growth factor. In summary, the Bunch–Parlett, fast Bunch–Parlett, and bounded Bunch–Kaufman pivoting strategies satisfy the strong condition, whereas the Bunch–Kaufman pivoting strategy and that of Ashcraft et al. satisfy the weak condition. The weak condition controls element growth during the factorization, as shown by an argument similar to those in [1, 4, 6, 13], [15, Chapter 11]. The [illegible] in factoring $A \in R^{n\times n}$ is defined by

$$\rho(A) = \frac{\max_{i,j,k} |a_{ij}^{(k)}|}{\max_{i,j} |a_{ij}|}, \tag{3.1}$$

where a_{ij} and $a_{ij}^{(k)}$ are the (i,j) entries of A and of the kth Schur complement, respectively, and $\|\cdot\|_M$ is the maximum magnitude element in the given matrix.

When a 1×1 pivot is chosen, we have

$$\begin{aligned} \frac{\max_{p\neq k} |a_{pk}|^2}{|a_{kk}|} &\leq \frac{1}{\alpha} \max\{|a_{pq}| : p \neq q \text{ and } (a_{qk} \neq 0 \text{ or } q = k)\} \\ &\leq \frac{1}{\alpha} \max_{p\neq q} |a_{pq}|. \end{aligned} \tag{3.2}$$

Therefore, the element growth is bounded by $1 + \frac{1}{\alpha}$.

If a 2×2 block pivot is chosen, the weak condition guarantees $|a_{ii}a_{jj}| < \alpha^2\lambda^2$. Then

$$\left| \det \left(\begin{bmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{bmatrix} \right) \right| = |a_{ij}^2 - a_{ii}a_{jj}| > (1 - \alpha^2)\lambda^2. \tag{3.3}$$

Since $0 < \alpha < 1$,

$$\left| \begin{bmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{bmatrix}^{-1} \right| < \frac{1}{(1-\alpha^2)\lambda^2} \begin{bmatrix} |a_{jj}| & \lambda \\ \lambda & |a_{ii}| \end{bmatrix}.$$

Therefore, the increase of each element in magnitude for the 2×2 block decomposition is bounded by

$$\begin{aligned}\frac{1}{(1-\alpha^2)\lambda^2}\begin{bmatrix}\lambda & \sigma\end{bmatrix}\begin{bmatrix}|a_{jj}| & \lambda\\ \lambda & |a_{ii}|\end{bmatrix}\begin{bmatrix}\lambda\\ \sigma\end{bmatrix} &= \frac{1}{(1-\alpha^2)\lambda^2}(\lambda^2(|a_{jj}|+\sigma)+(\lambda^2+\sigma|a_{ii}|)\sigma)\\ &< \frac{1}{(1-\alpha^2)\lambda^2}(\lambda^2(\alpha\sigma+\sigma)+(\lambda^2+\alpha\lambda^2)\sigma)\\ &= \frac{2(1+\alpha)\sigma}{1-\alpha^2} = \frac{2\sigma}{1-\alpha}, \end{aligned} \tag{3.4}$$

and the element growth for the 2×2 block decomposition is bounded by $1+\frac{2}{1-\alpha}$.

Therefore, element growth is bounded by

$$g = \max\left\{1+\frac{1}{\alpha}, \sqrt{1+\frac{2}{1-\alpha}}\right\}.$$

The minimum of g is $\frac{1+\sqrt{17}}{2}\approx 2.562$, which is attained when $\alpha=\frac{1+\sqrt{17}}{8}\approx 0.640$.

Thus

$$\rho(A) \le g^{n-1}. \tag{3.5}$$

The attainability of the last inequality is a research problem [15, Problem 11.10].

With complete pivoting (the Bunch–Parlett pivoting strategy), we can bound the growth factor of $A\in R^{n\times n}$ as

$$\rho(A) \le 3nf(n), \quad \text{where } f(n) = \left(\prod_{k=2}^{n} k^{1/(k-1)}\right)^{1/2} \le 1.8n^{(\ln n)/4}$$

with the pivoting argument $\alpha=\frac{1+\sqrt{17}}{8}$. This was shown by Bunch [2] with an analysis similar to Wilkinson's for Gaussian elimination with complete pivoting [16].

We note that the bounds on element increases in (3.2) and (3.4) are in terms of off-diagonal elements. Therefore, the growth factor $\bar{\rho}(A)$ for off-diagonal elements is bounded by g^{n-2}, i.e.,

$$\bar{\rho}(A) = \frac{\max_{i\ne j,k}|a_{ij}^{(k)}|}{\max_{i\ne j}|a_{ij}|} \le g^{n-2} \quad (n>1). \tag{3.6}$$

This is attainable, for example, with $\alpha=\frac{1+\sqrt{17}}{8}$ and

$$A = \begin{bmatrix} -\alpha & 1 & 1 & \cdots & 1\\ 1 & -\alpha g-\frac{1}{\alpha} & 1 & \cdots & 1\\ 1 & 1 & -\alpha g^2-\frac{g}{\alpha}-\frac{1}{\alpha} & \ddots & \vdots\\ \vdots & \vdots & \ddots & \ddots & 1\\ 1 & 1 & \cdots & 1 & -\alpha g^{n-1}-\frac{g^{n-2}}{\alpha}-\frac{g^{n-3}}{\alpha}-\cdots-\frac{1}{\alpha}\end{bmatrix}.$$

The weak condition is stronger than necessary to bound the growth factor; we need only

$$|a_{kk}|\max_{p\ne q}|a_{pq}| \ge \alpha\max_{p\ne k}|a_{pk}|^2$$

FIG. 3.1. *Experimental maximum growth factor for factoring a symmetric matrix.*

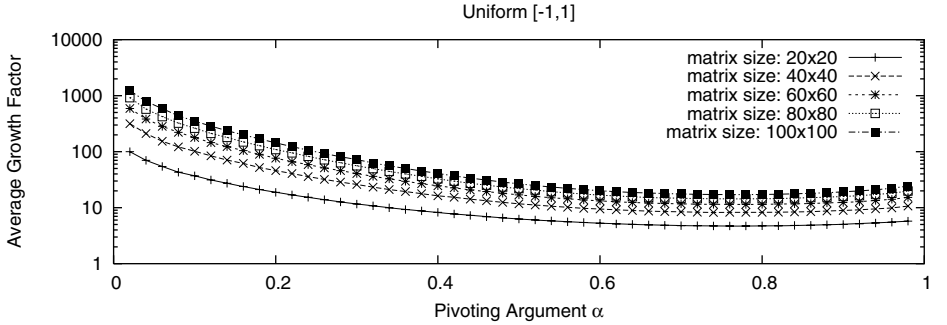


FIG. 3.2. *Experimental average growth factor for factoring a symmetric matrix, Bunch–Kaufman.*

for 1×1 pivots, but our version of the weak condition is useful for the triadic case considered in section 4.2.

In practice, the average growth factors for both tridiagonal and full matrices are far from this bound. Figure 3.1 shows the maximum growth factor of 20,000 random symmetric $n \times n$ matrices for each $n = 1, \ldots, 100$ with $\alpha = \frac{1+\sqrt{17}}{8} \approx 0.640$. In our experiments, all matrix elements are drawn independently from a uniform distribution on $[-1, 1]$; results for a normal distribution are similar. Although $\alpha \approx 0.640$ minimizes the a priori bound on the growth factor, our experiments show that the best α to minimize the average growth factor with Bunch–Kaufman pivoting is usually between 0.74 and 0.78, as shown in Figure 3.2, where 20,000 random matrices are generated for each matrix size and each α.

3.5. The strong condition bounds elements in L. The weak condition does not bound L for general matrices. For example [13], [15, section 11.1.2],

$$A = \begin{bmatrix} 0 & \epsilon & \\ \epsilon & 0 & 1 \\ & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & & \\ 0 & 1 & \\ \frac{1}{\epsilon} & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & \epsilon & \\ \epsilon & 0 & \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & \frac{1}{\epsilon} \\ & 1 & 0 \\ & & 1 \end{bmatrix} = LBL^T,$$

when the Bunch–Kaufman pivoting strategy is applied. As $\epsilon \to 0$, L is unbounded.

In contrast, the strong condition does ensure a bound on elements in L. When a 1×1 pivot is chosen, then the magnitude of elements in the pivot column of L is bounded by $\frac{1}{\alpha}$. If a 2×2 block pivot is chosen, the strong condition implies $\lambda = \sigma$ and

TABLE 3.1
The element growth bound g and the bound γ for L (when complete or rook pivoting is used) with two optimal choices of α.

	α	γ	g
Minimize g	$\frac{1+\sqrt{17}}{8} \approx 0.640$	$\frac{7+\sqrt{17}}{4} \approx 2.781$	$\frac{1+\sqrt{17}}{2} \approx 2.562$
Minimize γ	$\frac{1}{2}$	2	3

therefore the two columns of L corresponding to this 2×2 block pivot have elements bounded by

$$\frac{1}{(1-\alpha^2)\lambda^2} \begin{bmatrix} \lambda & \sigma \end{bmatrix} \begin{bmatrix} |a_{jj}| & \lambda \\ \lambda & |a_{ii}| \end{bmatrix} < \frac{1}{(1-\alpha^2)\lambda^2} \begin{bmatrix} \lambda & \lambda \end{bmatrix} \begin{bmatrix} \alpha\lambda & \lambda \\ \lambda & \alpha\lambda \end{bmatrix}$$
$$= \frac{1+\alpha}{1-\alpha^2} \begin{bmatrix} 1 & 1 \end{bmatrix} = \frac{1}{1-\alpha} \begin{bmatrix} 1 & 1 \end{bmatrix}.$$

Therefore, the elements in L are bounded in magnitude by

$$\gamma = \max\left\{\frac{1}{\alpha}, \frac{1}{1-\alpha}\right\}.$$

3.6. The growth factor and element bounds. We summarize the results on element growth in the following theorem, which extends some previous results to general α.

THEOREM 3.1. [illegible] LBL^T [illegible] $A \in R^{n\times n}$ [illegible] $\rho(A)$ [illegible] (3.1) [illegible]

$$\rho(A) \le g^{n-1},$$

[illegible]

$$g = \max\left\{1 + \frac{1}{\alpha}, \sqrt{1 + \frac{2}{1-\alpha}}\right\},$$

[illegible] α [illegible] L [illegible]

$$\gamma = \max\left\{\frac{1}{\alpha}, \frac{1}{1-\alpha}\right\}.$$

As shown above, $\alpha = \frac{1+\sqrt{17}}{8}$ minimizes g, the element growth bound. But $\alpha = 0.5$ minimizes the bound γ on the elements of L. The consequences of each of these choices are summarized in Table 3.1.

4. Diagonal pivoting strategies for triadic symmetric matrices. In section 2, we showed that sparsity is preserved in the LXL^T factorization of a symmetric triadic matrix with any diagonal pivoting strategy. In this section, we study a pivoting strategy particular to symmetric tridiagonal matrices [3] and also apply the pivoting strategies from the previous section to triadic matrices.

Algorithm 4 Bunch's pivot selection.

$\alpha = \frac{\sqrt{5}-1}{2} \approx 0.618$
if $|a_{11}|\sigma \geq \alpha |a_{21}|^2$ **then**
  Use a_{11} as a 1×1 pivot.
else
  Use $\left[\begin{smallmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{smallmatrix}\right]$ as a 2×2 block pivot.
end if

4.1. Pivoting strategies specific to symmetric tridiagonal matrices. One pivoting strategy has been proposed for LBL^T factorizations of irreducible tridiagonal matrices. Consider the variant proposed by Higham [14] of the algorithm of Bunch [3] represented in Algorithm 4, with parameter $\sigma = \max_{i,j} |a_{ij}|$. The algorithm's great advantage is that there are interchanges of rows and columns, yet the growth factor is bounded by

$$\rho(A) = \max\left\{1 + \frac{1}{\alpha}, \frac{1}{1-\alpha}\right\},$$

whose minimum is achieved by choosing $\alpha = \frac{\sqrt{5}-1}{2}$. This method is excellent for applications relying on B (e.g., computing inertia), but there is no element bound on L, illustrated, for example, as $\epsilon \to 0$ and

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

A similar example is given in [14]. Therefore, this algorithm is not well suited to computing Newton-like directions or solving tridiagonal systems of equations with corner elements. Nevertheless, Higham showed that it is a stable method for solving linear symmetric tridiagonal systems [14].

Note that Algorithm 4 requires computing the maximum magnitude element σ of the original matrix in advance. Recently Bunch and Marcia [5] developed another LBL^T factorization algorithm for symmetric tridiagonal matrices that does not need the whole matrix a priori and requires no interchanges of rows and columns. It is favored in some applications.

4.2. Pivoting strategies from those for dense matrices. All the pivoting strategies from section 3 can be applied to a symmetric triadic matrix $A \in R^{n \times n}$. The growth factor is constrained because of the triadic structure, and we obtain a sharper result for $\rho(A)$ than that of Theorem 3.1, although the bound γ on the elements of L remains the same.

THEOREM 4.1. LBL^T $A \in R^{n \times n}$ $\rho(A)$ (3.1)

$$\rho(A) \leq \begin{cases} \frac{4g(g^{(n-3)/2}-1)}{g-1} + 2(g^{(n-1)/2} + g^{(n+1)/2}) + 1 & n \\ \frac{4g(g^{(n-2)/2}-1)}{g-1} + 2g^{n/2} + 1 & n \end{cases}.$$

$\rho(A) = O(g^{n/2})$

$$\rho(A) \leq 2ng^{\lfloor \lg(n-1) \rfloor} \leq 2n(n-1)^{\lg g} = O(n^{1+\lg g}),$$

. . $n > 1$. $g = \max\{\frac{1}{\alpha}, \frac{1}{1-\alpha^2}\}$

The proof of this theorem is given in the appendix.

If we choose $\alpha = \frac{\sqrt{5}-1}{2}$ to minimize g, then $\lg g \approx 0.694$, and therefore the bound for the strong condition is subquadratic. Even linear growth is rare, but it is possible; for example, if we take the circulant matrix A with second row equal to $[1, -2, 1, 0, \ldots, 0]$ and change its $(1,1)$ element to -1, then $\rho(A) = n/2 + O(1)$.

For the weak condition, exponential growth is achievable; define the $n \times n$ matrix

$$A = \begin{bmatrix} -a & -1 & 0 & 1 & 0 & 0 & & & 0 \\ -1 & -a & 0 & 0 & 0 & 0 & & & 1 \\ 0 & 0 & -a & -1 & 0 & 1 & & & 0 \\ 1 & 0 & -1 & (g-1)a & 0 & 0 & & & 0 \\ 0 & 0 & 0 & 0 & -a & -1 & & \ddots & \\ 0 & 0 & 1 & 0 & -1 & (g-1)a & & & 0 \\ & & & & & & \ddots & \ddots & \vdots \\ & & & & \ddots & & \ddots & \ddots & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}, \tag{4.1}$$

$|a| < \alpha = \frac{\sqrt{5}-1}{2}$ and n odd. Then when $a \to \alpha^-$, the $(n, 2j)$ entry becomes g^{j-1} after $(j-1)$ 2×2 pivots for $j = 1, \ldots, \frac{n-1}{2}$, and therefore $\rho(A) = O(g^{n/2})$.

Despite these examples, in our experiments, $\rho(A)$ is almost always bounded by a constant for both the weak and the strong conditions.

Although $\alpha = \frac{\sqrt{5}-1}{2} \approx 0.618$ minimizes the a priori bound on the relative element increase, our experiments show that the best α to minimize the average growth factor is usually between 0.82 and 0.86 for Bunch–Kaufman pivoting, as illustrated in Figure 4.1, where 20,000 random matrices are generated for each matrix size and α.

With pivoting argument $\alpha = \frac{\sqrt{5}-1}{2}$, there are symmetric triadic matrices A having $\rho(A) = O(g^{n/2})$ and $\rho(A) = O(n)$ for the weak and strong conditions, respectively. But our experiments show that, in practice, LBL^T factorizations of symmetric tridiagonal or symmetric tridiagonal matrices with corner elements added usually show only constant growth in $\rho(A)$, whenever any of the four pivoting strategies are applied. Figure 4.2 shows the maximum growth factor of 20,000 random symmetric tridiagonal $n \times n$ matrices for each $n = 50, 100, \ldots, 1000$ and for random symmetric tridiagonal matrices with corner elements.

4.3. Pivoting cost. When the Bunch–Parlett algorithm is applied, it is natural to search the whole matrix instead of only the lower (or upper) triangular part due to the data structure for sparse matrices. So the number of comparisons is at most $3k + O(1)$ to select a pivot in a $k \times k$ Schur complement. Therefore, the total number of comparisons is bounded by $\frac{3}{2}n^2 + O(n)$ for a symmetric triadic $A \in R^{n \times n}$, which is more expensive than the $O(n)$ cost of the factorization. The Bunch–Kaufman algorithm requires at most $5n + O(1)$ comparisons for a symmetric triadic $A \in R^{n \times n}$. For the fast Bunch–Parlett and bounded Bunch–Kaufman pivoting strategies, the worst-case number of comparisons is the same as that of Bunch–Parlett pivoting. The average number of element comparisons is between that for the Bunch–Kaufman and Bunch–Parlett pivoting strategies. Figure 4.3 shows the average number of comparisons of 20,000 symmetric matrices for each $n = 50, 100, \ldots, 1000$.

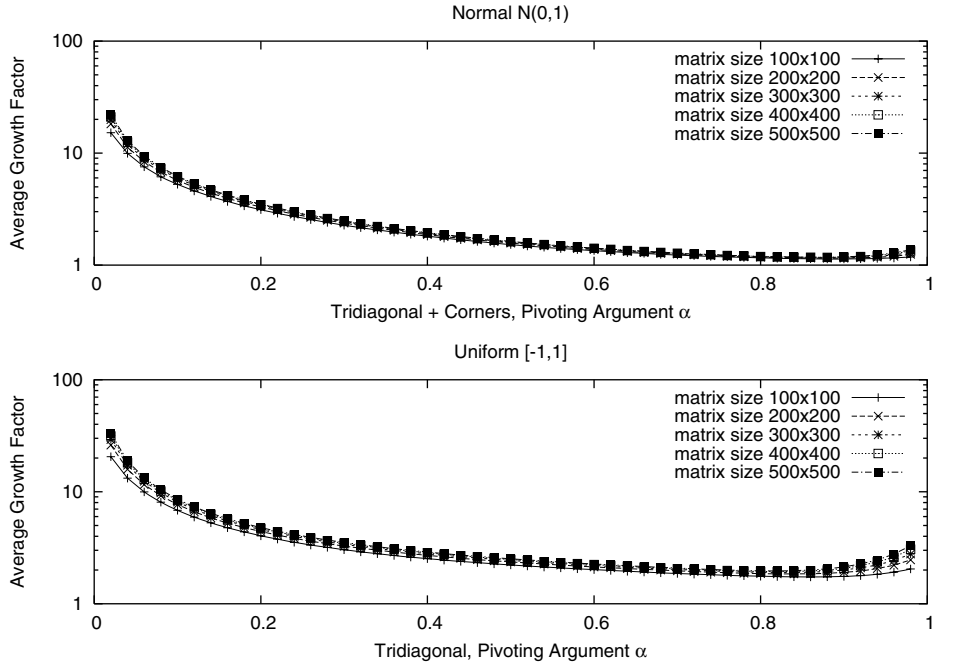


FIG. 4.1. *Experimental average growth factor for Bunch–Kaufman pivoting on a symmetric triadic matrix or a symmetric tridiagonal matrix.*

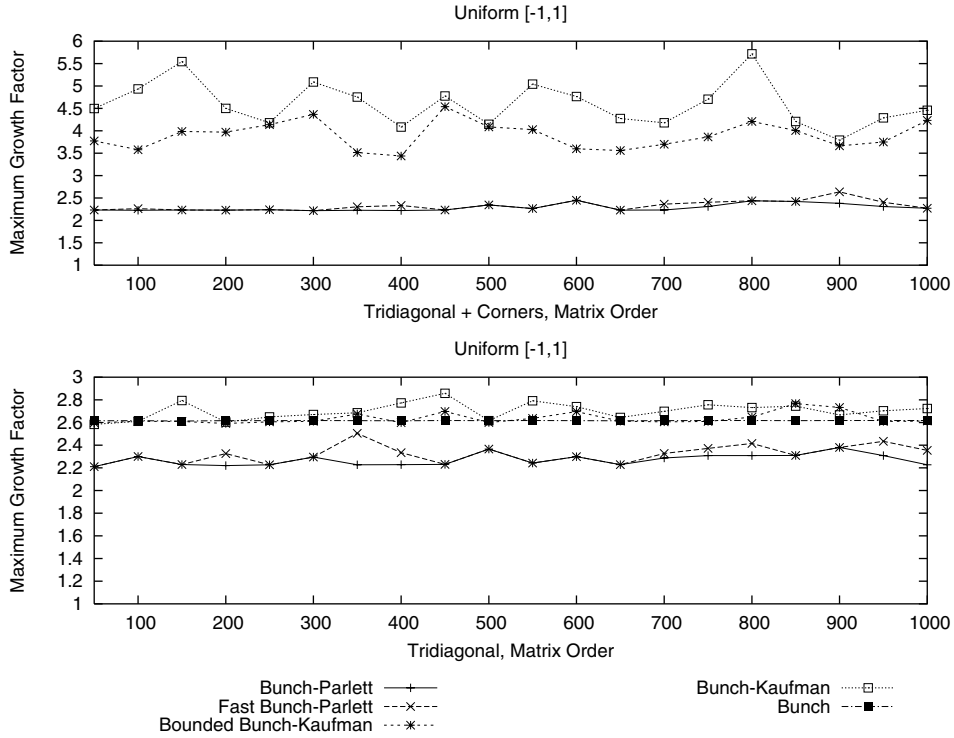


FIG. 4.2. *Experimental average growth factor for factoring a symmetric triadic matrix or a symmetric tridiagonal matrix.*

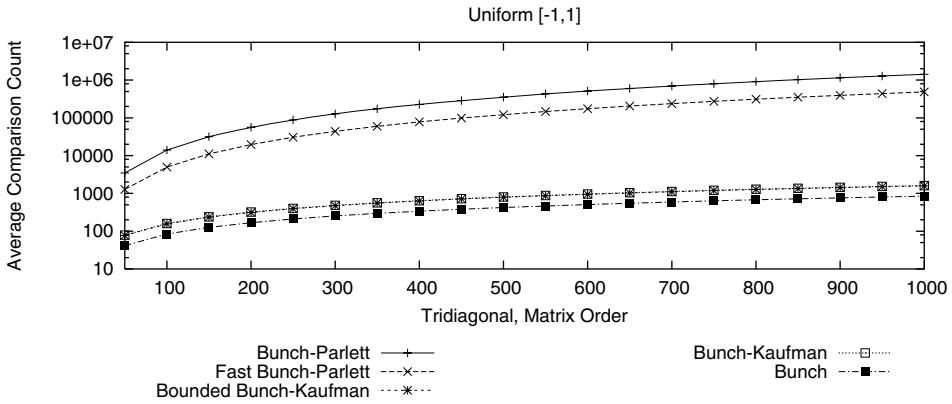


FIG. 4.3. *Experimental average number of comparisons to factor a symmetric tridiagonal matrix.*

5. Perturbation theory. The perturbation analysis of LL^T factorization of a positive semidefinite symmetric matrix with complete pivoting is discussed in [12]. Partition A as

$$A = \begin{bmatrix} A_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix},$$

where $A_{11} \in R^{k\times k}$, and partition L and E accordingly. Assume that both A_{11} and $A_{11} + E_{11}$ are nonsingular, and let $W = A_{11}^{-1}A_{21}^T = L_{11}^{-T}L_{21}^T$. In [12], Higham showed that with complete pivoting applied to a general positive semidefinite matrix,

$$\|W\|_{2,F} \le \sqrt{\frac{1}{3}(n-k)(4^k-1)}.$$

We give bounds on $\|W\|_{2,F}$ for LXL^T factorization of both full symmetric and symmetric triadic matrices.

THEOREM 5.1. $S_k(A)$ LXL^T A k k $k < n$ A E $\|\cdot\|$ p $|\cdot|$ A_{11} $A_{11} + E_{11}$

$$S_k(A+E) - S_k(A) = E_{22} - (E_{21}W + W^T E_{21}^T) + W^T E_{11} W + O(\|E\|^2),$$

$$|S_k(A+E) - S_k(A)| \le |E_{22}| + |E_{21}|\,|W| + |W^T|\,|E_{21}^T| + |W^T|\,|E_{11}|\,|W| + O(\|E\|^2)$$

$$\|S_k(A+E) - S_k(A)\| \le \|E\|(1+\|W\|^2)^2 + O(\|E\|^2),$$

$$\|W\|_{2,F} \le \sqrt{\frac{\gamma}{\gamma+2}(n-k)\left((1+\gamma)^{2k}-1\right)}$$

γ L
A

$$\|W\|_{2,F} \le \frac{2\gamma\Phi_\gamma}{\Phi_\gamma - 1}\sqrt{\frac{\Phi_\gamma^{2k}-1}{\Phi_\gamma^2-1}} = O(\Phi_\gamma^k),$$

$$\Phi_\gamma = \frac{1+\sqrt{1+4/\gamma}}{2}\,\gamma.$$

The proof of the theorem is contained in the following series of lemmas. We begin by generalizing to LXL^T factorizations a result of Higham [12] for LL^T factorization.

LEMMA 5.2. $S_k(A)$ LXL^T A k k $k < n$ A E A

$$A = \begin{bmatrix} A_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix},$$

$A_{11} \in R^{k\times k}$ E A_{11} $A_{11}+E_{11}$

$$S_k(A+E) = S_k(A) + E_{22} - (E_{21}W + W^T E_{21}^T) + W^T E_{11} W + O(\|E\|^2),$$

$W = A_{11}^{-1}A_{21}^T$. The factorization takes the form

$$A = \begin{bmatrix} A_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} L_{11} & \\ L_{21} & I_{n-k} \end{bmatrix} \begin{bmatrix} X & \\ & S_k(A) \end{bmatrix} \begin{bmatrix} L_{11}^T & L_{21}^T \\ & I_{n-k} \end{bmatrix},$$

where $L_{11} \in R^{k\times k}$ is lower triangular and the symmetric matrix $X \in R^{k\times k}$ is block diagonal with block order 1 or 2. The matrix X is either the identity, a diagonal matrix, or a block diagonal matrix, depending on the factorization. In any case, $A_{11} = L_{11}XL_{11}^T$ and $A_{21} = L_{21}XL_{11}^T$. Therefore, $W = A_{11}^{-1}A_{21}^T = L_{11}^{-T}L_{21}^T$. We also know that $S_k(A) = A_{22} - A_{21}A_{11}^{-1}A_{21}^T$, and since A_{11} is nonsingular, $(A_{11}+E_{11})^{-1} = (I + A_{11}^{-1}E_{11})^{-1}A_{11}^{-1} = (I - A_{11}^{-1}E_{11})A_{11}^{-1} + O(\|E_{11}\|^2)$. The result is obtained by substituting the previous two equations into $S_k(A+E) = (A_{22}+E_{22}) - (A_{21}+E_{21})(A_{11}+E_{11})^{-1}(A_{21}+E_{21})^T$ and collecting the $O(\|E\|^2)$ terms. □

Next, we bound the elements in $W = L_{11}^{-T}L_{21}^T$.

LEMMA 5.3. L γ

$$|W| = |L_{11}^{-T}L_{21}^T| \le ye^T,$$

$y_{k-j} = \gamma(1+\gamma)^j$ e

. The matrix W satisfies $L_{11}^T W = L_{21}^T$, so let's consider a single column of this relationship. Let r be a column of W. We will compute a vector y satisfying $|r| \le y$. Note that $|r_k|/\gamma$ is bounded by 1, and $|r_{k-j}|/\gamma$ is bounded by 1 plus the sum of the later entries in r. If we let s_{k-j} be a bound on the sum of the entries $k-j,\ldots,k$, then for $j = 1,2,\ldots,k-1$, we have the recursions

$$y_{k-j} = \gamma(1+s_{k-j+1}),$$
$$s_{k-j} = s_{k-j+1} + y_{k-j},$$

with $y_k = \gamma$ and $s_k = \gamma$. The solution to these recursions is

$$y_{k-j} = \gamma(1+\gamma)^j,$$
$$s_{k-j} = (1+\gamma)^{j+1} - 1.$$

Therefore, each column of W is bounded in absolute value by y and the result follows. □

The bound on $\|W\|$ follows immediately. This takes care of the general case and leaves only the triadic bound to be demonstrated. We begin with two simple lemmas and then proceed to the main result.

LEMMA 5.4. $F_\gamma(n) = \sum_{i=1}^{\lceil n/2 \rceil} \binom{n-i}{i-1} \gamma^{n-i}$ $\Phi_\gamma = \frac{1+\sqrt{1+4/\gamma}}{2}\gamma$

$$\frac{1}{1+(1/\gamma)}\Phi_\gamma^{n-1} \le F_\gamma(n) \le \Phi_\gamma^{n-1}$$

$n = 1, 2, \ldots$ $\gamma > 0$

. We first observe that $F_\gamma(n) = \gamma(F_\gamma(n-1) + F_\gamma(n-2))$ for $n > 2$, with $F_\gamma(1) = 1$ and $F_\gamma(2) = \gamma$. Note that $\gamma + \gamma\Phi_\gamma = \Phi_\gamma^2$. The result can be obtained by mathematical induction. □

LEMMA 5.5. $C \ge 0$ $m \times n$ $n \ge 2$ $\|C\|_p \le \|C\hat{I}\|_p$ $1 \le p \le \infty$ $p = F$ $n \times (n-1)$ $\hat{I}$ $n-1$

. The cases of $p = F$ (Frobenius norm) and $p = \infty$ (∞-norm) are trivial. When $0 \le p < \infty$, $\|C\|_p = \max_{\|x\|_p=1} \|Cx\|_p = \|Cz\|_p$, and this value is achieved for some z with $\|z\|_p = 1$. Note that $z_i \ge 0$ for $i = 1, \ldots, n$, since all the elements of C are nonnegative. Let $\hat{z} = [z_1, \ldots, z_{n-2}, \max(z_{n-1}, z_n)]^T$. Then $\|\hat{z}\|_p \le 1$, and

$$\|C\|_p = \|Cz\|_p \le \|C\hat{I}\hat{z}\|_p \le \|C\hat{I}(\hat{z}/\|\hat{z}\|_p)\|_p \le \max_{\|x\|_p=1} \|C\hat{I}x\|_p = \|C\hat{I}\|_p. \quad \square$$

LEMMA 5.6. LBL^T

$$\|L_{11}^{-T} L_{21}^T\|_{2,F} \le \frac{2\gamma\Phi_\gamma}{\Phi_\gamma - 1}\sqrt{\frac{\Phi_\gamma^{2k}-1}{\Phi_\gamma^2-1}} = O(\Phi_\gamma^k),$$

$\gamma \ge 1$ L $\Phi_\gamma = \frac{1+\sqrt{1+4/\gamma}}{2}\gamma$

. The proof of Lemma 5.4 shows $F_\gamma(i) = \gamma(F_\gamma(i-1) + F_\gamma(i-2))$ for $i > 2$. Observing the elements in $L_{11}^{-1} L_{11} = I$, we obtain

$$|L_{11}^{-1}| \le \sum_{i=1}^{k} F_\gamma(k) Z^{k-1} = \begin{bmatrix} F_\gamma(1) & & & & \\ F_\gamma(2) & F_\gamma(1) & & & \\ F_\gamma(3) & F_\gamma(2) & F_\gamma(1) & & \\ \vdots & \ddots & \ddots & \ddots & \\ F_\gamma(k) & \cdots & F_\gamma(3) & F_\gamma(2) & F_\gamma(1) \end{bmatrix},$$

where $Z \in R^{k \times k}$ is the shift-down matrix. Note that this bound is attainable with $L_{11} = I - \gamma Z - \gamma Z^2$. By Lemma 5.4,

$$|L_{11}^{-T}| e \le \left[\frac{\Phi_\gamma^k - 1}{\Phi_\gamma - 1}, \frac{\Phi_\gamma^{k-1} - 1}{\Phi_\gamma - 1}, \ldots, 1 \right]^T. \tag{5.1}$$

Since L is triadic, each row of L_{21}^T has at most two nonzero elements. Let $|L_{21}^T| = R_1 + R_2$, where R_1 and R_2 contain the first and the second nonzero elements in each row, respectively. Then

$$\|L_{11}^{-T} L_{21}^T\| \le \||L_{11}^{-T}||L_{21}^T|\| \le \||L_{11}^{-T}|R_1\| + \||L_{11}^{-T}|R_2\|.$$

By Lemma 5.5,

$$\begin{aligned}\||L_{11}^{-T}|R_1\| &\le \||L_{11}^{-T}|R_1\hat{I}_{n-k}\| \\ &\le \||L_{11}^{-T}|R_1\hat{I}_{n-k}\hat{I}_{n-k-1}\| \\ &\le \cdots \le \||L_{11}^{-T}|R_1\hat{I}_{n-k}\hat{I}_{n-k-1}\cdots\hat{I}_2\| \\ &\le \||L_{11}^{-T}|(\gamma e)\| = \gamma\||L_{11}^{-T}|e\|.\end{aligned}$$

Similarly, $\||L_{11}^{-T}|R_2\| \le \gamma\||L_{11}^{-T}|e\|$. By (5.1), for $\gamma \ge 1$

$$\|L_{11}^{-T} L_{21}^T\|_{2,F} \le 2\gamma\||L_{11}^{-T}|e\|_{2,F} \le \frac{2\gamma\Phi_\gamma}{\Phi_\gamma - 1}\sqrt{\frac{\Phi_\gamma^{2k} - 1}{\Phi_\gamma^2 - 1}}.$$

Note that this bound is halved when $n - k = 1$. □

For positive semidefinite triadic matrices and complete pivoting, $\gamma = 1$ so $\Phi_\gamma^k = O((\frac{1+\sqrt{5}}{2})^k)$.

In the LBL^T factorization of a symmetric triadic matrix with diagonal pivoting, γ can be 2 or $\frac{7+\sqrt{17}}{4} \approx 2.781$, to minimize the element bound of matrix L or the element growth factor, respectively.

6. Concluding remarks. We have studied various pivoting strategies in computing the LXL^T factorizations of symmetric triadic matrices. We denote the strategies as follows: BT (Bunch's pivoting strategy for a symmetric tridiagonal matrix), BP (Bunch–Parlett), FBP (fast Bunch–Parlett), BBK (bounded Bunch–Kaufman), and BK (Bunch–Kaufman). We summarize our results as follows:

1. The LL^T, LDL^T, and LBL^T factors of a symmetric triadic matrix with any diagonal pivoting strategy remain sparse.
2. We have analyzed the boundedness of the growth factors in case the pivoting strategy satisfies either a strong or a weak condition.
3. We have presented a new choice of the α parameter that better controls the growth factor.
4. In the LBL^T factorization with various pivoting strategies, L is bounded for BP, FBP, and BBK pivoting strategies, whereas the BK pivoting strategy may result in L unbounded. All four pivoting strategies have the growth factor controlled for full symmetric matrices. The bound on the growth factor is smaller for symmetric triadic matrices.
5. For symmetric matrices, pivoting strategies BT and BK produce an L matrix with no bounds on its elements, whereas the magnitude of elements in L from pivoting strategies BBK, BP, and FBP is bounded by a constant γ given in Table 3.1, depending on the parameter α in the algorithm.
6. For LDL^T factorization of a positive definite symmetric matrix A with complete pivoting, the magnification factor in the error bound for the Schur complement after k steps is $\sqrt{1/3(n-k)(4^k - 1)}$ if A is full [12], and $O((\frac{1+\sqrt{5}}{2})^k)$ if A triadic.

7. For two pivoting strategies D and E, we will say $D \succ E$, $D \succeq E$, and $D \simeq E$ if D is better than, slightly better than, or similar to E, respectively. Our experimental results with pivoting argument $\alpha = \frac{\sqrt{5}-1}{2} \approx 0.618$ are as follows: For LBL^T factorizations of uniformly distributed tridiagonal matrices, the maximum growth factors satisfy BP $\succeq$ FBP $\succ$ BT $\succeq$ BBK $\succeq$ BK, as shown in Figure 4.2, whereas the average number of comparisons satisfies BT $\succ$ BK $\simeq$ BBK $\succ$ FBP $\succ$ BP, as shown in Figure 4.3. Thus, more expensive pivoting usually yields a smaller growth factor.

Appendix. Proof of bounds for pivoting on triadics in Theorem 4.1.

THEOREM A.1. LBL^T $A \in R^{n\times n}$ $\bar{\rho}(A)$ (3.6)

$$\bar{\rho}(A) \le \begin{cases} 2g^{\lfloor \lg(n-1) \rfloor} \le 2(n-1)^{\lg g} & \\ 2g^{\lfloor (n-1)/2 \rfloor} & \end{cases}$$

$n > 1$

$$g = \max\left\{\frac{1}{\alpha}, \frac{1}{1-\alpha^2}\right\}.$$

. Without loss of generality, we assume the required interchanges of rows and columns for pivoting are done prior to the factorization. Let $S_k(A)$ be the Schur complement of A after reducing k rows and k columns, and let

$$A^{(k+1)} = \begin{array}{c} \\ k \\ n-k \end{array} \begin{array}{c} \begin{array}{cc} k & n-k \end{array} \\ \begin{bmatrix} 0 & 0 \\ 0 & S_k(A) \end{bmatrix} \end{array}.$$

By Lemmas 2.1 and 2.3, at most two diagonal and two off-diagonal elements are changed in the Schur complement. We denote them by $a_{ii}^{(k+1)}$, $a_{jj}^{(k+1)}$, $a_{ij}^{(k+1)}$, and $a_{ji}^{(k+1)}$. In addition to $a_{ij}^{(k+1)}$ and $a_{ji}^{(k+1)}$, $A^{(k+1)}$ has at most one nonzero off-diagonal element in each of ith and jth rows, inherited from $A^{(k-p)}$. Let $p = 1$ or $p = 2$ for the previous selection being 1×1 or 2×2, respectively.

Assume for now that

$$a_{ij}^{(k+1-p)} = a_{ji}^{(k+1-p)} = 0 \tag{A.1}$$

for each k. Later we will show that if this assumption breaks, the bounds on the off-diagonal growth factor are at most doubled.

For a 1×1 pivot, (A.1) implies that the weak condition coincides with the strong condition. Therefore,

$$|a_{ij}^{(k+1)}| = \frac{|a_{ik}^{(k)}||a_{jk}^{(k)}|}{|a_{kk}^{(k)}|} \le \frac{1}{\alpha}\min\{|a_{ik}^{(k)}|, |a_{jk}^{(k)}|\} \le g\min\{|a_{ik}^{(k)}|, |a_{jk}^{(k)}|\}. \tag{A.2}$$

For a 2×2 pivot $\begin{bmatrix} a_{k-1,k-1}^{(k-1)} & a_{k-1,k}^{(k-1)} \\ a_{k,k-1}^{(k-1)} & a_{kk}^{(k-1)} \end{bmatrix}$, there are at most two nonzero off-diagonal elements under the pivot, denoted by $a_{i,k-1}^{(k-1)}$ and $a_{jk}^{(k-1)}$. If $i = j$, then the only element changed in $A^{(k+1)}$ from $A^{(k-1)}$ is $a_{ii}^{(k+1)}$. In this case, the matrix size is

reduced without increasing the off-diagonal elements. In order to maximize $\bar{\rho}(A)$, we assume $i \neq j$. The weak condition ensures (3.3). Therefore,

$$
\begin{aligned}
|a_{ij}^{(k+1)}| &\leq \frac{1}{(1-\alpha^2)|a_{k,k-1}^{(k-1)}|^2} \begin{bmatrix} |a_{i,k-1}^{(k-1)}| & 0 \end{bmatrix} \begin{bmatrix} |a_{kk}^{(k-1)}| & |a_{k-1,k}^{(k-1)}| \\ |a_{k,k-1}^{(k-1)}| & |a_{k-1,k-1}^{(k-1)}| \end{bmatrix} \begin{bmatrix} 0 \\ |a_{jk}^{(k-1)}| \end{bmatrix} \\
&= \frac{|a_{i,k-1}^{(k-1)}||a_{jk}^{(k-1)}|}{(1-\alpha^2)|a_{k,k-1}^{(k-1)}|} \\
&\leq \begin{cases} g \min\{|a_{i,k-1}^{(k-1)}|, |a_{jk}^{(k-1)}|\} & \text{if strong condition holds,} \\ g|a_{jk}^{(k-1)}| & \text{if weak condition holds.} \end{cases} \qquad \text{(A.3)}
\end{aligned}
$$

Since the Schur complement is symmetric, we consider the elements in the lower triangular. Let $G(m) = g^m \max_{i \neq j} |a_{ij}|$.

Consider the case that the strong condition holds. By (A.3) for a 2×2 pivot, an off-diagonal element of size $G(m)$ requires three $G(m-1)$ elements: $|a_{i,k-1}^{(k-1)}|$, $|a_{jk}^{(k-1)}|$, and $|a_{k,k-1}^{(k-1)}|$. Note that the strong condition guarantees $|a_{k,k-1}^{(k-1)}| \geq |a_{i,k-1}^{(k-1)}|$. By (A.2) for a 1×1 pivot, if $|a_{ij}^{(k+1)}| \geq G(m)$, then $|a_{ik}^{(k)}|, |a_{jk}^{(k)}| \geq G(m-1)$. In other words, for a 1×1 pivot, an off-diagonal element of size $G(m)$ requires two off-diagonal supporting elements of size $G(m-1)$. Therefore, the bound on element growth using 1×1 pivots is higher than that using 2×2 pivots. Note that each 1×1 elimination step introduces at most one fill-in entry. Considering a sequence of 2^{m-1} pivots of size 1×1, we see by induction that a $G(m)$ element requires 2^m $G(0)$ elements that cannot contribute to the growth of other elements, and thus the growth must be logarithmic. We illustrate this in the following diagram for obtaining a $G(3)$ element with the smallest number of pivots. The last column indicates the Schur complements as the sources of the two off-diagonal elements in each row if they were not present initially. Note that $G(0)$ elements are from the original matrix A, whereas $G(1)$, $G(2)$, and $G(3)$ elements are fill-in entries during the factorization.

$*$		$G(0)$						$G(0)$	
	$*$	$G(0)$				$G(0)$			
$G(0)$	$G(0)$	$*$				$G(1)$		$G(1)$	$A^{(2)}, A^{(1)}$
			$*$		$G(0)$		$G(0)$		
				$*$	$G(0)$	$G(0)$			
			$G(0)$	$G(0)$	$*$	$G(1)$	$G(1)$		$A^{(5)}, A^{(4)}$
	$G(0)$	$G(1)$		$G(0)$	$G(1)$	$*$	$G(2)$	$G(2)$	$A^{(6)}, A^{(3)}$
			$G(0)$		$G(1)$	$G(2)$	$*$	$G(3)$	$A^{(7)}$
$G(0)$		$G(1)$				$G(2)$	$G(3)$	$*$	

The number of pivots is $2^{m-1} + 2^{m-2} + \cdots + 2^0 = 2^m - 1$. The last 2×2 Schur complement, with or without a row/column reduced afterward, cannot contribute to off-diagonal element growth. Therefore, the dimension of the smallest matrix that can have a $G(m)$ off-diagonal element is $(2^m - 1) + 2 = 2^m + 1$. If A has dimension less than $2^m + 1$ but larger than 2^{m-1}, then the off-diagonal elements in the Schur complements are at most $G(m-1)$ in magnitude. In other words,

$$\bar{\rho}(A) \leq g^{\lfloor \lg(n-1) \rfloor} \leq (n-1)^{\lg g}. \qquad \text{(A.4)}$$

Consider the case that the weak condition holds. Recall that for a 1×1 pivot, the weak condition coincides with the strong condition, and an off-diagonal element of

size $G(m)$ requires two $G(m-1)$ elements. By (A.3) for a 2×2 pivot, an off-diagonal element of size $G(m)$ requires only one $G(m-1)$ element. For maximal growth from $G(0)$ to $G(1)$ we use a 1×1 pivot. Otherwise, the bound on element growth using 2×2 pivots is at least as big as that using 1×1 pivots. The bound can increase by a factor of g for every two rows reduced during the decomposition, except from $G(0)$ to $G(1)$ (one row/column reduced). The last Schur complement cannot contribute to off-diagonal element growth. Therefore,

$$\bar{\rho}(A) \le g^{\lfloor (n-1)/2 \rfloor}, \tag{A.5}$$

where $A \in R^{n\times n}$ is symmetric triadic.

So far we assume (A.1) holds. Now we show that if (A.1) breaks, the bounds in (A.4) and (A.5) are at most doubled. If $a_{ij}^{(k+1-p)} = a_{ji}^{(k+1-p)} \neq 0$, then there are no other off-diagonal elements in the ith and jth rows and columns in $A^{(k+1)}$, where $p = 1, 2$ stands for 1×1, 2×2 pivots, respectively. As a result, $A^{(k+1)}$ is a reducible matrix. After diagonally interchanging rows and columns, $A^{(k+1)}$ consists of two diagonal blocks: $\left[\begin{smallmatrix} a_{ii}^{(k+1)} & a_{ij}^{(k+1)} \\ a_{ji}^{(k+1)} & a_{jj}^{(k+1)} \end{smallmatrix}\right]$ and the remaining matrix, in which all the elements are taken from $A^{(k+1-p)}$. The bound on $a_{ji}^{(k+1)}$ in the 2×2 block is at most doubled, since it is a sum of two terms, each of which is bounded as (A.4) or (A.5), depending on whether the condition satisfied is strong or weak. Note that no off-diagonal element growth occurs afterward in this 2×2 block, and the other block is intact. Therefore, we obtain the result by safely declaring that the bounds in (A.4) and (A.5) are at most doubled if (A.1) breaks. ∎

THEOREM A.2. [illegible] LBL^T [illegible] $A \in R^{n\times n}$ [illegible] $\rho(A)$ [illegible] (3.1) [illegible]

$$\rho(A) \le \begin{cases} \frac{4g(g^{(n-3)/2}-1)}{g-1} + 2(g^{(n-1)/2} + g^{(n+1)/2}) + 1 & \text{[illegible] } n \text{ [illegible]}, \\ \frac{4g(g^{(n-2)/2}-1)}{g-1} + 2g^{n/2} + 1 & \text{[illegible] } n \text{ [illegible]}. \end{cases}$$

[illegible] $\rho(A) = O(g^{n/2})$ [illegible]

$$\rho(A) \le 2ng^{\lfloor \lg(n-1) \rfloor} \le 2n(n-1)^{\lg g} = O(n^{1+\lg g}),$$

[illegible] $n > 1$ [illegible] $g = \max\{\frac{1}{\alpha}, \frac{1}{1-\alpha^2}\}$ [illegible]. The major difference between $\rho(A)$ and $\bar{\rho}(A)$ is that the diagonal element increases can accumulate, whereas the accumulation of two off-diagonal element increases results in a reducible Schur complement, so further accumulation is impossible. Therefore, the diagonal element growth factor is bounded by the sum of n elements, each of which is bounded by Theorem 4.1. So we obtain the bound on $\rho(A)$ for the strong condition. Though this approach also gives a bound for the weak condition, a tighter bound can be obtained, as follows.

The proof of Theorem A.1 shows that the off-diagonal element bound in the Schur complement depends on the number of rows/columns reduced. We follow the notation in the proof of Theorem A.1.

If the weak condition holds, the off-diagonal elements $a_{ij}^{(k+1)}$ in $A^{(k+1)}$ (after reducing k rows/columns) are bounded as $|a_{ij}^{(k+1)}| \le 2g^{\lfloor (k+1)/2 \rfloor} \max |a_{ij}|$ for $i \neq j$ and k from 1 to $n-2$. This is also the bound on the diagonal element increase of A_{k+1} from the previous iteration. We sum up all the relative element increases during

the decomposition to obtain a bound on $\rho(A)$, where $A \in R^{n\times n}$ is symmetric triadic:

$$\rho(A) \le \underline{1} + 2g^{\lfloor 2/2\rfloor} + 2g^{\lfloor 3/2\rfloor} + \cdots + 2g^{\lfloor (n-1)/2\rfloor} + \underline{2g^{\lfloor (n-1)/2\rfloor+1}}$$
$$= \begin{cases} \frac{4g(g^{(n-3)/2}-1)}{g-1} + 2(g^{(n-1)/2} + g^{(n+1)/2}) + 1 & \text{if } n \text{ odd}, \\ \frac{4g(g^{(n-2)/2}-1)}{g-1} + 2g^{n/2} + 1 & \text{if } n \text{ even}. \end{cases}$$

The underlined 1 occurs because each diagonal element in the initial A can be $G(0)$. The reason for the last term $\underline{2g^{\lfloor (n-1)/2\rfloor+1}}$ is as follows. If a 1×1 pivot is chosen in the last 2×2 Schur complement or a 2×2 pivot is chosen in the last 3×3 Schur complement, the reduction can still increase the very last diagonal element, but there is no off-diagonal element growth. If (A.1) breaks, the reduced 2×2 block can have diagonal element growth but no off-diagonal element growth. This case is also taken into account in $\underline{2g^{\lfloor (n-1)/2\rfloor+1}}$. In a similar vein, we can also obtain a slightly tighter bound for the strong condition, but it is also $O(n^{1+\lg g})$:

$$\rho(A) \le 1 + 2g^{\lfloor \lg 2\rfloor} + 2g^{\lfloor \lg 3\rfloor} + \cdots + 2g^{\lfloor \lg(n-1)\rfloor} + 2g^{\lfloor \lg(n-1)\rfloor+1} = O(n^{1+\lg g}). \qquad \square$$

Acknowledgments. The authors are grateful to the referees for their careful reading of the manuscript. It is a pleasure to thank Nick Higham for his encouragement in this work.

REFERENCES

[1] C. Ashcraft, R. G. Grimes, and J. G. Lewis, *Accurate symmetric indefinite linear equation solvers*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 513–561.

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

[3] J. R. Bunch, *Partial pivoting strategies for symmetric matrices*, SIAM J. Numer. Anal., 11 (1974), pp. 521–528.

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

[5] J. R. Bunch and R. F. Marcia, *A pivoting strategy for symmetric tridiagonal matrices*, Numer. Linear Algebra Appl., 12 (2005), pp. 911–922.

[6] J. R. Bunch and B. N. Parlett, *Direct methods for solving symmetric indefinite systems of linear equations*, SIAM J. Numer. Anal., 8 (1971), pp. 639–655.

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

[8] H.-R. Fang, *Backward Error Analysis of Factorization Algorithms for Symmetric and Symmetric Triadic Matrices*, Tech. Report CS-4734, Computer Science Department, University of Maryland, College Park, MD, 2005.

[9] H.-R. Fang and D. P. O'Leary, *Modified Cholesky Factorizations: A Catalog with New Approaches*, Tech Report TR-4807, Computer Science Department, University of Maryland, College Park, MD, 2006.

[10] A. Forsgren, P. E. Gill, and W. Murray, *Computing modified Newton directions using a partial Cholesky factorization*, SIAM J. Sci. Comput., 16 (1995), pp. 139–150.

[11] P. E. Gill, W. Murray, and M. H. Wright, *Unconstrained methods*, in Practical Optimization, Academic Press, 1981, Chap. 4, pp. 105–116.

[12] N. J. Higham, *Analysis of the Cholesky decomposition of a semi-definite matrix*, in Reliable Numerical Computation, M. G. Cox and S. J. Hammarling, eds., Oxford University Press, New York, 1990, pp. 161–185.

[13] N. J. Higham, *Stability of the diagonal pivoting method with partial pivoting*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 52–65.

[14] N. J. Higham, *Stability of block LDL^T factorization of a symmetric tridiagonal matrix*, Linear Algebra Appl., 287 (1999), pp. 181–189.

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

[16] J. H. Wilkinson, *Error analysis of direct methods of matrix inversion*, J. ACM, 8 (1961), pp. 281–330.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 2, pp. 596–601

INERTIAL SIGNATURES OF HERMITIAN POLYNOMIAL MATRICES*

HARALD K. WIMMER†

Abstract. Signs associated with real characteristic roots of a hermitian polynomial matrix $L(s)$ are determined from signatures of Hankel matrices of $L(s)^{-1}$.

Key words. hermitian matrix pencil, polynomial matrix, Weierstraß canonical form, inertial signs, signature, Hankel matrix

AMS subject classifications. 15A57, 15A21, 15A54, 15A22, 93B15

DOI. 10.1137/050639041

1. Introduction.

We first review the notion of inertial signs of a hermitian pencil $P(s) = A_0 + A_1 s$. Suppose $A_0, A_1 \in \mathbb{C}^{n\times n}$ are hermitian, and A_1 is nonsingular. Then $\det P \neq 0$ (zero polynomial) and P^{-1} is strictly proper rational. Let $\sigma(P) = \{\lambda \,|\, \det P(\lambda) = 0\}$ denote the set of [illegible] of P. The following result goes back to Weierstraß (see [7], [5], [4]).

LEMMA 1.1. [illegible] $A_0 + A_1 s \in \mathbb{C}^{n\times n}[s]$ [illegible] A_1 [illegible] $T \in \mathbb{C}^{n\times n}$ [illegible] $T(A_0 + A_1 s)T^*$ [illegible] I [illegible] II [illegible]

$$\epsilon\, D_r(s,\alpha) = \epsilon \begin{pmatrix} 0 & 0 & \cdots & -1 & s-\alpha \\ 0 & 0 & \cdots & s-\alpha & 0 \\ \vdots & \vdots & \iddots & \vdots & \vdots \\ -1 & s-\alpha & \cdots & 0 & 0 \\ s-\alpha & 0 & \cdots & 0 & 0 \end{pmatrix}_{r\times r} \tag{I}$$

[illegible] $\alpha \in \mathbb{R}$ [illegible] $\epsilon \in \{1, -1\}$ [illegible]

$$G_{2k}(s,\beta) = \begin{pmatrix} 0 & D_k(s,\beta) \\ D_k(s,\bar\beta) & 0 \end{pmatrix}_{2k\times 2k} \tag{II}$$

[illegible] $\beta \notin \mathbb{R}$ [illegible] $T(A_0 + A_1 s)T^* =$

$$\operatorname{diag}\big(\dots, \epsilon\, D_r(s,\alpha), \dots, G_{2k}(s,\beta), \dots\big) \tag{1.1}$$

[illegible]

The block diagonal matrix in (1.1) is the [illegible] of the pencil $A_0 + A_1 s$. We observe that a number $\epsilon = \pm 1$ is attached to each block of type I. Thus a sign can be associated with each elementary divisor corresponding to a real characteristic root α.

DEFINITION 1.2. [illegible] $A_0 + A_1 s$ [illegible] π_i [illegible] $(s-\alpha)^i$ [illegible] $\epsilon_{i1}, \dots, \epsilon_{i\pi_i}$ [illegible]

$$(\dots, \epsilon_{i1}, \dots, \epsilon_{i\pi_i}, \dots) \tag{1.2}$$

*Received by the editors August 27, 2005; accepted for publication (in revised form) by P. Benner February 23, 2006; published electronically August 16, 2006.
http://www.siam.org/journals/simax/28-2/63904.html
†Mathematisches Institut, Universität Würzburg, Am Hubland, D-97074 Würzburg, Germany (wimmer@mathematik.uni-wuerzburg.de).

α

$$\eta_i = \epsilon_{i1} + \cdots + \epsilon_{i\pi_i}, \tag{1.3}$$

$\eta_i = 0$, $\pi_i = 0$ η_i α

If $A \in \mathbb{C}^{n\times n}$ is a hermitian matrix with ℓ positive and ν negative eigenvalues (counting multiplicities), then the difference $\ell - \nu$ is the of A. It will be denoted by $\operatorname{sgn} A$. If $\pi_i \neq 0$, then

$$\eta_i = \operatorname{sgn} \operatorname{diag}(\epsilon_{i1}, \ldots, \epsilon_{i\pi_i}).$$

Definition 1.2 is motivated by [6]. The terminology is not uniform. The Cauchy characteristic in [2] includes elementary divisors and signs, and the term sign characteristic can be found in [3].

Now consider a nonsingular hermitian polynomial matrix $L \in \mathbb{C}^{n\times n}[s]$ such that

$$L(s) = A_0 + A_1\, s + \cdots + A_t\, s^t$$

and $A_i = A_i^*$, $i = 0, \ldots, t$. Assume that L^{-1} is strictly proper rational. Set $m = \deg \det L$. A factorization

$$L^{-1}(s) = C(A_0 + A_1 s)^{-1} C^* \tag{1.4}$$

is a of L^{-1} if $P(s) = A_0 + A_1 s \in \mathbb{C}^{m\times m}[s]$ is a hermitian pencil and $C \in \mathbb{C}^{n\times m}$. The following observation (see, e.g., [9]) is an immediate consequence of Kalman's state space isomorphism theorem.

LEMMA 1.3. $L \in \mathbb{C}^{n\times n}[s]$ $\deg \det L = m$ L^{-1} L^{-1} (1.4)

$$L^{-1}(s) = \tilde{C}(\tilde{A}_0 + \tilde{A}_1 s)^{-1} \tilde{C}^*$$

A_1 $\tilde{A}_1$ $T \in \mathbb{C}^{m\times m}$

$$\tilde{A}_0 + \tilde{A}_1 s = T(A_0 + A_1 s)T^* \quad \textit{and} \quad C = \tilde{C}T^*. \tag{1.5}$$

Set $\sigma(L) = \{\lambda \,|\, \det L(\lambda) = 0\}$. If (1.4) is a minimal hermitian realization, then $\sigma(L) = \sigma(A_0 + A_1 s)$, and (see, e.g., [1]) the elementary divisors of the pencil $A_0 + A_1 s$ are the same as those of the polynomial matrix L. Moreover the preceding lemma shows that the pencil $A_0 + A_1 s$ is determined by L up to congruence. This leads to the following definition of inertial signs and signatures of polynomial matrices.

DEFINITION 1.4. $L \in \mathbb{C}^{n\times n}[s]$ L^{-1} $L^{-1}(s) = C(A_0 + A_1 s)^{-1} C^*$ $\alpha \in \sigma(L)$ $\alpha \in \mathbb{R}$ α α $A_0 + A_1 s$

It is the purpose of this paper to determine inertial signatures of real characteristic values of L using Laurent expansions of L^{-1}. Without loss of generality we may assume $0 \in \sigma(L)$. We focus on the inertial signatures at $\alpha = 0$. Let

$$W_{L^{-1}}(s) = s^{-1}[W_0 + s^{-1}W_1 + \cdots + s^{-(k-1)}W_{k-1}] \tag{1.6}$$

be the principal part of the Laurent expansion of $L^{-1}(s)$ at $\alpha = 0$. Define the Hankel matrices

$$H(L^{-1}) = \begin{pmatrix} W_0 & W_1 & \cdot & \cdot & \cdot & W_{k-2} & W_{k-1} \\ W_1 & W_2 & \cdot & \cdot & \cdot & W_{k-1} & \\ \cdot & \cdot & \cdot & \cdot & \cdot & & \\ W_{k-2} & W_{k-1} & & & & & \\ W_{k-1} & & & & & & \end{pmatrix}, \tag{1.7}$$

$$H(sL^{-1}) = \begin{pmatrix} W_1 & \cdot & \cdot & \cdot & W_{k-2} & W_{k-1} \\ W_2 & \cdot & \cdot & \cdot & W_{k-1} & \\ \cdot & \cdot & \cdot & \cdot & & \\ W_{k-1} & & & & & \end{pmatrix}, \ldots, \; H(s^{k-1}L^{-1}) = W_{k-1}.$$

Our main result is the following.

THEOREM 1.5. Let $L \in \mathbb{C}^{n\times n}[s]$ [illegible] $\alpha = 0$ [illegible] L [illegible] 0 [illegible] k [illegible] L^{-1} [illegible] η_i [illegible] L [illegible] $\alpha = 0$ [illegible]

$$\begin{pmatrix} \eta_1 \\ \cdot \\ \cdot \\ \cdot \\ \eta_{k-1} \\ \eta_k \end{pmatrix} = \begin{pmatrix} 1 & 0 & -1 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot \\ & 1 & 0 & -1 & 0 & & \cdot & \cdot & \cdot \\ & & 1 & 0 & -1 & 0 & \cdot & \cdot & \cdot \\ & & & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & & & & 0 & -1 & 0 \\ & & & & & & 1 & 0 & -1 \\ & & & & & & & 1 & 0 \\ & & & & & & & & 1 \end{pmatrix} \begin{pmatrix} \operatorname{sgn} H(s^{k-1}L^{-1}) \\ \cdot \\ \cdot \\ \cdot \\ \operatorname{sgn} H(sL^{-1}) \\ \operatorname{sgn} H(L^{-1}) \end{pmatrix}. \tag{1.8}$$

The proof of the theorem will be given in section 3. It is based on a result of Turnbull [8].

2. Turnbull's signature test. In this section we deal with a hermitian pencil $P(s) = A_0 + A_1 s$. We describe a modified form of Turnbull's signature test [8]. The following notation will be used. We set $D_r(s) = D_r(s, 0)$, and define $r \times r$ matrices

$$E_r = (\delta_{i,r+1-i}) = \begin{pmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{pmatrix},$$

$$N_r = (\delta_{i+1,i}) = \begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{pmatrix}.$$

Then $D_r(s) = (sI - N_r^T)E_r$.

LEMMA 2.1. Let $P(s) = A_0 + A_1 s$ [illegible] s^k [illegible] $\pi_i = 0$ [illegible] $i > k$ [illegible] $\alpha = 0$ [illegible] (1.2) (1.3)

$$W_{P^{-1}}(s) = s^{-1}[M_0 + s^{-1}M_1 + \cdots + s^{k-1}M_{k-1}] \tag{2.1}$$

[illegible] $P^{-1}(s)$ [illegible] $\alpha = 0$ [illegible]

$$\begin{pmatrix} \operatorname{sgn} M_0 \\ \operatorname{sgn} M_1 \\ \cdot \\ \cdot \\ \cdot \\ \operatorname{sgn} M_{k-1} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & . & . & . \\ & 1 & 0 & 1 & 0 & 1 & . & . & . \\ & & 1 & 0 & 1 & 0 & . & . & . \\ & & & . & . & . & . & . & . \\ & & & & . & . & . & . & . \\ & & & & & . & . & . & . \\ & & & & & & 1 & 0 & 1 \\ & & & & & & & 1 & 0 \\ & & & & & & & & 1 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \cdot \\ \cdot \\ \cdot \\ \eta_{k-1} \\ \eta_k \end{pmatrix}. \tag{2.2}$$

[illegible]. We may assume

$$P(s) = \operatorname{diag}\big(\hat{P}(s),\, Q(s)\big),\ \sigma(\hat{P}) = \{0\},\ 0 \notin \sigma(Q). \tag{2.3}$$

Let $\hat{P}(s)$ be in Weierstraß canonical form such that

$$\hat{P}(s) = \operatorname{diag}\big(\dots, \epsilon_{i1} D_i(s), \dots, \epsilon_{i\pi_i} D_i(s), \dots\big). \tag{2.4}$$

Then

$$W_{P^{-1}}(s) = \hat{P}^{-1}(s) = \operatorname{diag}(\dots, \epsilon_{i1} D_i^{-1}(s), \dots, \epsilon_{i\pi_i} D_i^{-1}(s), \dots).$$

We first deal with the case $\hat{P}(s) = \epsilon\, D_k(s)$ and proceed as in [8]. From

$$\epsilon\, D_k^{-1}(s) = \epsilon \sum_{i=0}^{k} N_k^i E_k s^{-i-1} = \epsilon \begin{pmatrix} 0 & 0 & 0 & . & . & . & 0 & s^{-1} \\ 0 & 0 & 0 & . & . & . & s^{-1} & s^{-2} \\ . & . & . & . & . & . & . & . \\ 0 & s^{-1} & s^{-2} & . & . & . & s^{-(k-2)} & s^{-(k-1)} \\ s^{-1} & s^{-2} & s^{-3} & . & . & . & s^{-(k-1)} & s^{-k} \end{pmatrix}$$

follows

$$M_i = \epsilon\, N_k^i E_k = \epsilon\, \operatorname{diag}\big(0_{i\times i}, E_{k-i}\big).$$

Therefore $\operatorname{sgn} E_k = 0$ if k is even, and $\operatorname{sgn} E_k = 1$ if k is odd. Hence

$$\operatorname{sgn} M_{k-1} = \epsilon, \quad \operatorname{sgn} M_{k-2} = 0, \quad \operatorname{sgn} M_{k-3} = \epsilon, \dots, \tag{2.5}$$

and (2.2) holds with $(\eta_k, \eta_{k-1}, \dots, \eta_1) = (\epsilon, 0, \dots, 0)$. In the general case, with $\hat{P}(s)$ given as in (2.4), we obtain (2.2) by inspecting $\hat{P}^{-1}(s)$ and using (2.5). □

3. Proof of the theorem. We shall need a generalization of Sylvester's law of inertia [2, p. 200].

LEMMA 3.1. [illegible] $A \in \mathbb{C}^{n\times n}$ [illegible] $Y \in \mathbb{C}^{t\times n}$ [illegible] A [illegible] YAY^* [illegible].

The proof of Theorem 1.5 starts from a minimal hermitian realization

$$L^{-1}(s) = C(A_0 + A_1 s)^{-1} C^*, \tag{3.1}$$

where $P(s) = A_0 + A_1 s$ is given by (2.3). Then $\hat{P}(s) = \hat{A}_0 + \hat{A}_1 s$, and $\hat{A}_1$ is nonsingular, and $\hat{N} = -\hat{A}_1^{-1}\hat{A}_0$ is nilpotent with $\hat{N}^k = 0$. Let $C = (\hat{C}, D)$ be partitioned in accordance with (2.3). Then

$$L^{-1}(s) = \hat{C}\hat{P}^{-1}(s)\hat{C}^* + DQ(s)D^*$$

and

$$W_{L^{-1}}(s) = \hat{C}\hat{P}^{-1}(s)\hat{C}^* = \hat{C}[\hat{A}_1(-\hat{N} + sI)]^{-1}\hat{C}^*.$$

Hence we have (1.6) with

$$W_i = \hat{C}\hat{N}^i\hat{A}_1^{-1}\hat{C}^*, \quad i = 0, \dots, k-1.$$

Let $H(\hat{P}^{-1})$ be the Hankel matrix associated with $\hat{P}^{-1}(s)$. Then

$$H(\hat{P}^{-1}) = \begin{pmatrix} \hat{A}_1^{-1} & \hat{N}\hat{A}_1^{-1} & . & . & . & \hat{N}^{k-2}\hat{A}_1^{-1} & \hat{N}^{k-1}\hat{A}_1^{-1} \\ \hat{N}\hat{A}_1^{-1} & \hat{N}^2\hat{A}_1^{-1} & . & . & . & \hat{N}^{k-1}\hat{A}_1^{-1} & \\ . & . & . & . & . & & \\ \hat{N}^{k-1}\hat{A}_1^{-1} & & & & & & \end{pmatrix}.$$

Because of $\hat{N}^k = 0$ and $\hat{N}\hat{A}_1^{-1} = \hat{A}_1^{-1}\hat{N}^T$ we obtain

$$H(\hat{P}^{-1}) = \begin{pmatrix} I \\ \hat{N} \\ \vdots \\ \hat{N}^{k-1} \end{pmatrix} \hat{A}_1^{-1} \begin{pmatrix} I & \hat{N}^T & \dots & (\hat{N}^{k-1})^T \end{pmatrix}. \tag{3.2}$$

Let

$$\mathcal{O} = \mathcal{O}(\hat{N}, \hat{C}) = \begin{pmatrix} \hat{C} \\ \hat{C}\hat{N} \\ \vdots \\ \hat{C}\hat{N}^{k-1} \end{pmatrix}$$

be the observability matrix of the pair $\big(\hat{N}, \hat{C}\big)$. Then $H(L^{-1}) = \mathcal{O}\hat{A}_1^{-1}\mathcal{O}^*$, and similarly

$$H(s^iL^{-1}) = \mathcal{O}\hat{N}^i\hat{A}_1^{-1}\mathcal{O}^*, \quad i = 1, \dots, k-1.$$

The realization (3.1) is minimal. Hence $\mathcal{O}$ has full column rank [1], and Lemma 3.1 implies

$$\operatorname{sgn} H(s^iL^{-1}) = \operatorname{sgn} \hat{N}^i\hat{A}_1^{-1}, \quad i = 0, \dots, k-1. \tag{3.3}$$

Recall $W_{P^{-1}}(s) = \hat{P}^{-1}(s)$. Therefore the matrices M_i in (2.1) are given by $M_i = \hat{N}^i\hat{A}_1^{-1}$. Thus we have $\operatorname{sgn} M_i = \operatorname{sgn} H(s^iL^{-1})$. Then (2.1) yields

$$\big(\operatorname{sgn} H(L^{-1}), \dots, \operatorname{sgn} H(s^iL^{-1})\big) = (\eta_1, \dots, \eta_k)(I + N_k^2 + N_k^4 + \cdots),$$

and because of

$$(I + N_k^2 + N_k^4 + \cdots)^{-1} = I - N_k^2$$

the proof is complete.

REFERENCES

[1] P. A. Fuhrmann, *Linear Operators and Systems in Hilbert Space*, McGraw-Hill, New York, 1981.
[2] P. A. Fuhrmann, *On symmetric rational transfer functions*, Linear Algebra Appl., 50 (1983), pp. 167–250.
[3] I. Gohberg, P. Lancaster, and L. Rodman, *Spectral analysis of selfadjoint matrix polynomials*, Ann. of Math. (2), 112 (1980), pp. 33–71.
[4] I. Gohberg, P. Lancaster, and L. Rodman, *Matrices and Indefinite Scalar Products*, Birkhäuser, Basel, 1983.
[5] A. I. Malcev, *Foundations of Linear Algebra*, Freeman, San Francisco, 1963.
[6] J. Milnor, *On isometries of inner product spaces*, Invent. Math., 8 (1969), pp. 83–97.
[7] R. C. Thompson, *Pencils of complex and real symmetric and skew matrices*, Linear Algebra Appl., 147 (1991), pp. 323–371.
[8] H. W. Turnbull, *On the equivalence of pencils of Hermitian forms*, Proc. Lond. Math. Soc. II Ser., 39 (1935), pp. 232–248.
[9] J. C. Willems, *Realizations of systems with internal passivity and symmetry constraints*, J. Franklin Inst., 301 (1976), pp. 605–621.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 3, pp. 603–622

A FAST *ULV* DECOMPOSITION SOLVER FOR HIERARCHICALLY SEMISEPARABLE REPRESENTATIONS*

S. CHANDRASEKARAN†, M. GU‡, AND T. PALS†

Abstract. We consider an algebraic representation that is useful for matrices with off-diagonal blocks of low numerical rank. A fast and stable solver for linear systems of equations in which the coefficient matrix has this representation is presented. We also present a fast algorithm to construct the hierarchically semiseparable representation in the general case.

Key words. fast multipole method, hierarchically semiseparable, fast algorithms, orthogonal factorizations

AMS subject classification. 65F05

DOI. 10.1137/S0895479803436652

1. Introduction.

In this paper we consider a representation of structured dense matrices that we term [illegible] (HSS). This representation is a direct generalization of the one presented in [3]. It is also a special case of the FMM (fast multipole method) representations [20, 2, 29, 30, 31]. It has also been discussed as H^2 matrices in [24].

This representation is useful for matrices characterized by a hierarchical low numerical rank structure in the off-diagonal blocks. Examples of such matrices are shown in Figure 1. The matrix in Figure 1(a) is obtained,[1] for example, for the matrix $[\log|x_i - x_j|]$, where $0 \le x_i < x_{i+1} \le 1$. Similarly the matrix in Figure 1(b) is obtained for the matrix $[\log|\sin\pi(x_i - x_j)|]$. This class of matrices arises frequently in the numerical solution of partial differential and integral equations.

This work arose in an effort to stabilize the fast solver presented in [30, 31]. Our initial efforts in this direction were presented in [5, 6, 7, 8, 9, 27, 28]. During this time we learned about some work in linear time-varying systems theory [13], and other independent work [17, 15, 16, 23, 24], that encouraged us to generalize our ideas and present them in a more algebraic framework [3]. The corresponding technical report [4] is more comprehensive and will give some indication to the reader of how far the methods presented in this paper can be taken.

However, the FMM [20, 2, 29] is our most direct motivation for this work. In fact, the ideas presented here can be viewed as a stable approach to a fast inverse multipole method. There has been some significant work in this regard in the computational electromagnetics literature [1, 12, 19, 22, 26, 30, 31]. The method presented here is the first stable fast solver. In addition it also presents an algebraic generalization.

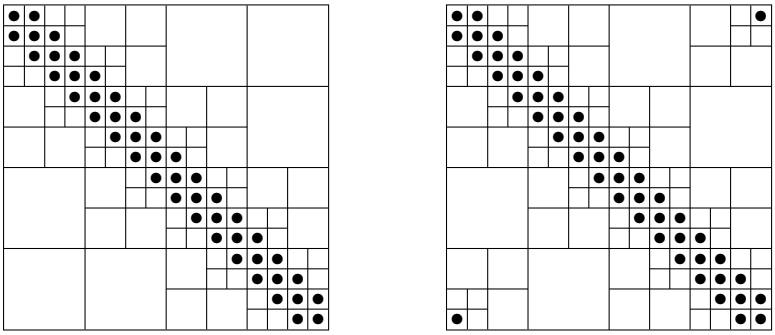
For applications of the fast solver we currently refer to [30, 31] and [4]. However, the applications are much wider than indicated in these references. Some of these will be presented in forthcoming papers.

*Received by the editors October 23, 2003; accepted for publication (in revised form) by V. Mehrmann October 25, 2005; published electronically August 25, 2006.
http://www.siam.org/journals/simax/28-3/43665.html

†Electrical and Computer Engineering Department, University of California, Santa Barbara, CA 93106-9560 (shiv@ece.ucsb.edu, tim@kipling.ece.ucsb.edu).

‡Mathematics Department, University of California, Berkeley, CA (mgu@math.berkeley.edu).

[1] In both cases the particular choice of the diagonal entries is not important.



(a) $\log|x-y|$ on a straight line.

(b) $\log|x-y|$ on a circle.

FIG. 1. *Submatrices labeled with a black dot are full-rank. All other submatrices have low numerical rank.*

For completeness we also present an algorithm to construct a numerical HSS representation of a general matrix. This algorithm requires $O(n^2)$ flops and $O(n)$ space. It is similar to the sequentially semiseparable (SSS) construction algorithm presented in [3, 13]. A construction algorithm for H^2 matrices is presented in [25]. The algorithm we present also ensures that the computed HSS representation satisfies some properties required to enable stability in the solver. We also present two cases where the construction can be carried out in $O(n)$ flops.

2. HSS representation. The representation is identical to the one presented in [30, 31], except that we view it more generically.

The HSS representation is a hierarchical representation that is based on a recursive row and column partitioning of the matrix. For example, for a 2×2 block partitioning of the matrix A its HSS representation is given by

$$A = \begin{pmatrix} D_{1;1} & U_{1;1}B_{1;1,2}V_{1;2}^H \\ U_{1;2}B_{1;2,1}V_{1;1}^H & \end{pmatrix},$$

where the subscripted D, U, V, and B matrices are in the representation. To see the recursive hierarchical nature we consider a block 4×4 partitioning of A and the resultant two-level HSS representation:

$$A = \begin{pmatrix} \begin{pmatrix} D_{2;1} & U_{2;1}B_{2;1,2}V_{2;2}^H \\ U_{2;2}B_{2;2,1}V_{2;1}^H & D_{2;2} \end{pmatrix} & \begin{pmatrix} U_{1;1}B_{1;1,2}V_{1;2}^H \end{pmatrix} \\ \begin{pmatrix} U_{1;2}B_{1;2,1}V_{1;1}^H \end{pmatrix} & \begin{pmatrix} D_{2;3} & U_{2;3}B_{2;3,4}V_{2;4}^H \\ U_{2;4}B_{2;4,3}V_{2;3}^H & D_{2;4} \end{pmatrix} \end{pmatrix}.$$

We first observe that only the two diagonal blocks $D_{1;1}$ and $D_{1;2}$ from the one-level HSS representation have been partitioned at the second level, each of them seemingly assigned their own HSS representations. However, that view is slightly misleading. In fact, in the two-level HSS representation of the matrix A, we do not store the matrices $U_{1;i}$ and $V_{1;i}$ for $i = 1, 2$. Rather we store only the $U_{2;i}$ and $V_{2;i}$ for $i = 1, 2, 3, 4$ and the translation operators $W_{2;i}$ and $R_{2;i}$ for $i = 1, 2, 3, 4$, which can be used to reconstruct the missing $U_{1;i}$ and $V_{1;i}$ via the defining relations

$$U_{1;1} = \begin{pmatrix} U_{2;1}R_{2;1} \\ U_{2;2}R_{2;2} \end{pmatrix},$$

$$U_{1;2} = \begin{pmatrix} U_{2;3}R_{2;3} \\ U_{2;4}R_{2;4} \end{pmatrix},$$
$$V_{1;1} = \begin{pmatrix} V_{2;1}W_{2;1} \\ V_{2;2}W_{2;2} \end{pmatrix},$$
$$V_{1;2} = \begin{pmatrix} V_{2;3}W_{2;3} \\ V_{2;4}W_{2;4} \end{pmatrix}.$$

These translation operators are an integral part of the FMM representation and literature, and their use in getting linear complexity algorithms is well known.

In general in a multilevel HSS representation the diagonal blocks at the ith level are labeled $D_{i;j}$. The $U_{i;j}$ at the lowest levels are used in conjunction with the translation operators $R_{i;j}$ at that level to reconstruct the $U_{i-1,j}$'s at the higher levels via

$$U_{i-1;j} = \begin{pmatrix} U_{i;2j}R_{i;2j-1} \\ U_{i;2j}R_{i;2j} \end{pmatrix}. \tag{1}$$

Similarly for $V_{i;j}$ we have

$$V_{i-1;j} = \begin{pmatrix} V_{i;2j}W_{i;2j-1} \\ V_{i;2j}R_{i;2j} \end{pmatrix}. \tag{2}$$

At every level only the diagonal blocks are eligible for partitioning. The off-diagonal blocks in the upper-triangular part of the ith level are of the form $U_{i;2j-1}B_{i;2j-1,2j}V_{i;2j}^H$ and in the lower-triangular part of the form $U_{i;2j}B_{i;2j,2j-1}V_{i;2j-1}^H$. Therefore, the complete HSS representation of the matrix A consists of the $D_{i;j}$, $U_{i;i}$, and $V_{i;j}$ at the lowest levels along with $B_{i;2j,2j-1}$, $B_{i;2j-1,2j}$, $R_{i;j}$, and $W_{i;j}$ at every level.

In the FMM literature it has been convenient to use a (binary in this case) tree on which all these matrices can be represented. In this notation the root of the tree corresponds to the whole matrix; the two children of the root correspond to the two row (and column) partitions, and so on. In Figure 2 we depict the HSS tree (also called a merge tree) for a uniform three-level HSS representation. We will refer to the ith node at the k-level of the tree as Node(k, i).

It should be observed that every matrix has an HSS representation. However, HSS representations are useful only when the translation operators are small compared to the size of the original matrix. These representations can be constructed in $O(n^2)$ flops and $O(n)$ space; see [10, 24]. In special cases these representations can be constructed in $O(n)$ flops. The FMM literature is rife with such results. Some other interesting instances can also be found in [10].

In this paper we restrict ourselves to HSS trees that are (almost) complete binary trees. In a future paper we will generalize our methods to incomplete binary trees.

3. Fast multiplication. In this section, for the convenience of the reader, we present the standard FMM algorithm for multiplying a matrix in HSS form with a regular vector (or unstructured dense matrix). In particular consider the matrix-vector product $z = Ab$, where A is in HSS form, with $K + 1$ levels in its HSS tree, and m_i indices in Node(K, i). Let $(\, b_{k;i}\,)$ denote a block row partitioning of b such that $b_{k;i}$ has the rows whose indices belong to Node(k, i). We partition z similarly.

We begin by observing that we need to do the multiplication

$$U_{1;1}B_{1;1,2}V_{1;2}^H b_{1;2} \tag{3}$$

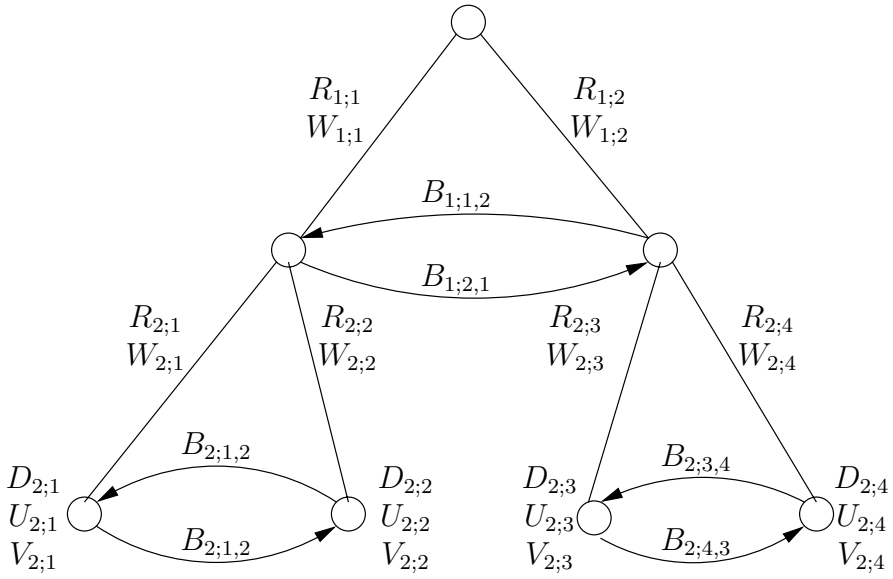


FIG. 2. *Three-level HSS representation on a binary tree.*

and the multiplication

$$U_{2;3}B_{2;3,4}V_{2;4}^H b_{2;4}.$$

Since

$$V_{1;2}^H b_{1;2} = \begin{pmatrix} V_{2;3}W_{2;3} \\ V_{2;4}W_{2;4} \end{pmatrix}^H \begin{pmatrix} b_{2;3} \\ b_{2;4} \end{pmatrix} = W_{2;3}^H V_{2;3}^H b_{2;3} + W_{2;4}^H V_{2;4}^H b_{2;4},$$

we can reduce the number of flops required to compute $V_{1;2}^H b_{1;2}$ in (3) if the number of columns in $W_{2;3}$ and $W_{2;4}$ is small compared to the number of rows in $V_{1;2}$. (This is the basis of all the super-fast algorithms and was first used by Greengard and Rokhlin in the design of FMMs.)

To formalize this idea we define the intermediate quantities

$$G_{k;i} = V_{k;i}^H b_{k;i}$$

and observe that the following recursions, as deduced from the preceding discussion, are available to compute them:

$$G_{K;i} = V_{K;i}^H b_{K;i}, \tag{4}$$

$$G_{k;i} = W_{k+1;2i-1}^H G_{k+1;2i-1} + W_{k+1;2i}^H G_{k+1;2i}. \tag{5}$$

With this notation we have that

$$U_{1;1}B_{1;1,2}V_{1;2}^H b_{1;2} = U_{1;1}B_{1;1,2}G_{1;2}.$$

We now observe that we need to perform the multiplication $U_{1;1}B_{1;1,2}G_{1;2}$. We also observe that we need to do the multiplication

$$U_{2;1}B_{2;1,2}G_{2;2}.$$

Using (1) we see that

$$U_{1;1}B_{1;1,2}G_{1;2} = \begin{pmatrix} U_{2;1}R_{2;1}B_{1;1,2}G_{1;2} \\ U_{2;2}R_{2;2}B_{1;1,2}G_{1;2} \end{pmatrix}.$$

Therefore the computation of $U_{1;1}B_{1;1,2}G_{1;2}$ can be merged with the computations of $U_{2;1}B_{2;1,2}G_{2;2}$ when computing

$$z_{2;1} = \cdots + U_{2;1}\left(B_{2;1,2}G_{2;2} + R_{2;1}B_{1;1,2}G_{1;2}\right) + \cdots,$$

where $\cdots$ denotes other terms that have to be added to produce the correct $z_{2;1}$. Similarly the term $U_{2;2}R_{2;2}B_{2;1,2}G_{2;2}$ from the computation of $A_{1;1,2}b_{1;2}$ can be merged into other terms involving $U_{2;2}$ in the computation of $z_{2;2}$. Clearly there is a recursive process occurring here. This motivates us to define the following intermediate quantities recursively:

$$F_{0;1} = 0, \tag{6}$$

$$F_{k,2i-1} = B_{k;2i-1,2i}G_{k;2i} + R_{k;2i-1}F_{k-1,i}, \tag{7}$$

$$F_{k,2i} = B_{k;2i,2i-1}G_{k;2i-1} + R_{k;2i}F_{k-1;i}. \tag{8}$$

We then observe that

$$z_{K;i} = D_{K;i}b_{K;i} + U_{K;i}F_{K;i}.$$

With that we have described how to compute $z = Ab$ rapidly when A has an HSS representation. Equations (4) and (5) are called the up-sweep recursions, and equations (6), (7), and (8) are called the down-sweep recursions for multiplication in the FMM literature.

4. Fast backward stable solver. In this section we present our fast solver. The algorithm we describe computes a ULV^H decomposition implicitly, where U and V are unitary matrices, and L is a lower-triangular matrix. By implicit, we mean that the factors are not computed and stored explicitly. However, the algorithm and techniques can be modified to compute the factors explicitly if so desired. That will be the subject of a future paper. The algorithm can also be easily modified to permit U and V to be represented as a product of elementary Gauss transforms and permutation matrices. This would lead to a more efficient algorithm but with some chance of numerical instability.

The basic idea of the algorithm is akin to that for the SSS representation. The one major difference is that we operate on all block rows at the same time, whereas in the SSS representation, each block row is operated on in a sequential fashion.

The algorithm is recursive in nature, and the recursion takes one of three possible forms.

4.1. Compressible off-diagonal blocks. This is the first possible way in which the recursion can proceed.

We begin by observing that block row i, excluding the diagonal block $D_{K;i}$, has its column space spanned by the columns of $U_{K;i}$. Hence if the number of columns of $U_{K;i}$, denoted by $n_{K;i}$, is strictly smaller than m_i, the number of rows in that block, we can find a unitary transformation $q_{K;i}$ such that

$$\bar{U}_{K;i} = q_{K;i}^H U_{K;i} = \begin{matrix} m_i - n_{K;i} \\ n_{K;i} \end{matrix} \begin{pmatrix} 0 \\ \hat{U}_{K;i} \end{pmatrix}.$$

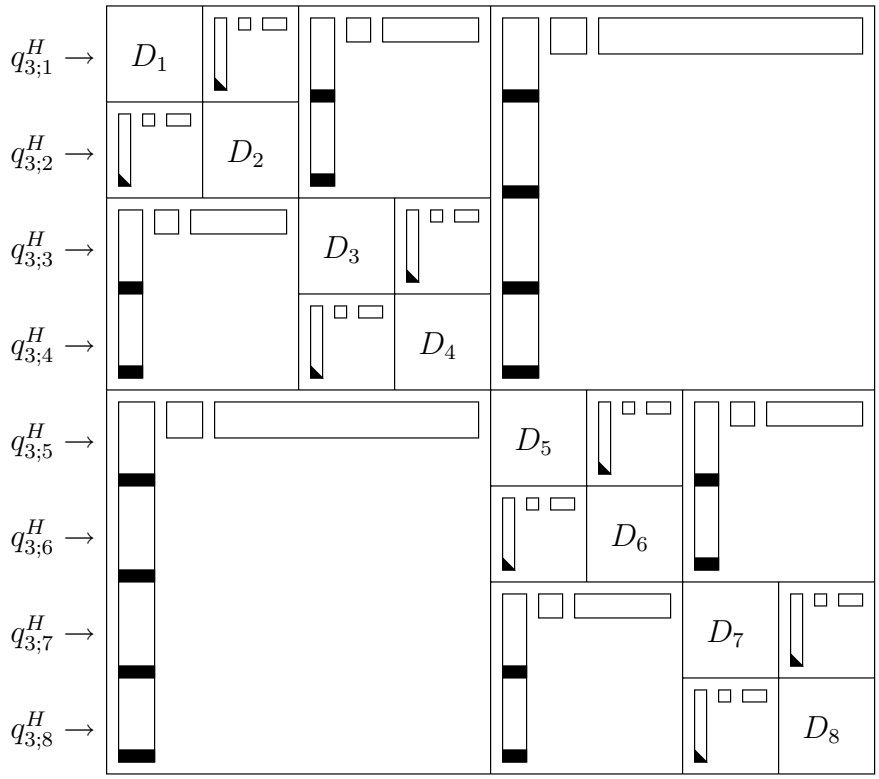


FIG. 3. *A pictorial representation showing the $q_{K;i}$'s compressing the off-diagonal portions of each block row. The black rectangles and triangles show the nonzero positions in the column bases of each off-diagonal block after compression by the $q_{K;i}$'s.*

In the above expression (and in the rest of the paper), variables written to the left of block matrices in parenthesis denote row partition sizes.

We now multiply block row i by q_i^H. (See Figure 3.) The change in the off-diagonal blocks is represented by the above equation since all of them have $U_{K;i}$ as the leading term. The ith block of the right-hand side changes to become

$$q^H_{K;i} b_{K;i} = \begin{matrix} m_i - n_{K;i} \\ n_{K;i} \end{matrix} \begin{pmatrix} \beta_{K;i} \\ \gamma_{K;i} \end{pmatrix}.$$

We also observe that $D_{K;i}$, the diagonal block, has become $q^H_{K;i} D_{K;i}$. Now we pick a unitary transformation $w_{K;i}$ such that

$$\bar{D}_{K;i} = (q^H_{K;i} D_{K;i}) w^H_{K;i} = \begin{matrix} \\ m_i - n_{K;i} \\ n_{K;i} \end{matrix} \begin{matrix} \begin{matrix} m_i - n_{K;i} & \quad n_{K;i} \end{matrix} \\ \begin{pmatrix} D_{K;i;1,1} & 0 \\ D_{K;i;2,1} & D_{K;i;2,2} \end{pmatrix} \end{matrix}.$$

We then multiply the block column i from the right by $w^H_{K;i}$. (See Figure 4.) The change in the diagonal block is represented by the above equation. The off-diagonal blocks in block column i have $V^H_{K;i}$ as the common last term. Hence we just need to multiply $V_{K;i}$ to obtain

$$\bar{V}_{K;i} = w_{K;i} V_{K;i} = \begin{matrix} m_i - n_{K;i} \\ n_{K;i} \end{matrix} \begin{pmatrix} \breve{V}_{K;i} \\ \hat{V}_{K;i} \end{pmatrix}.$$

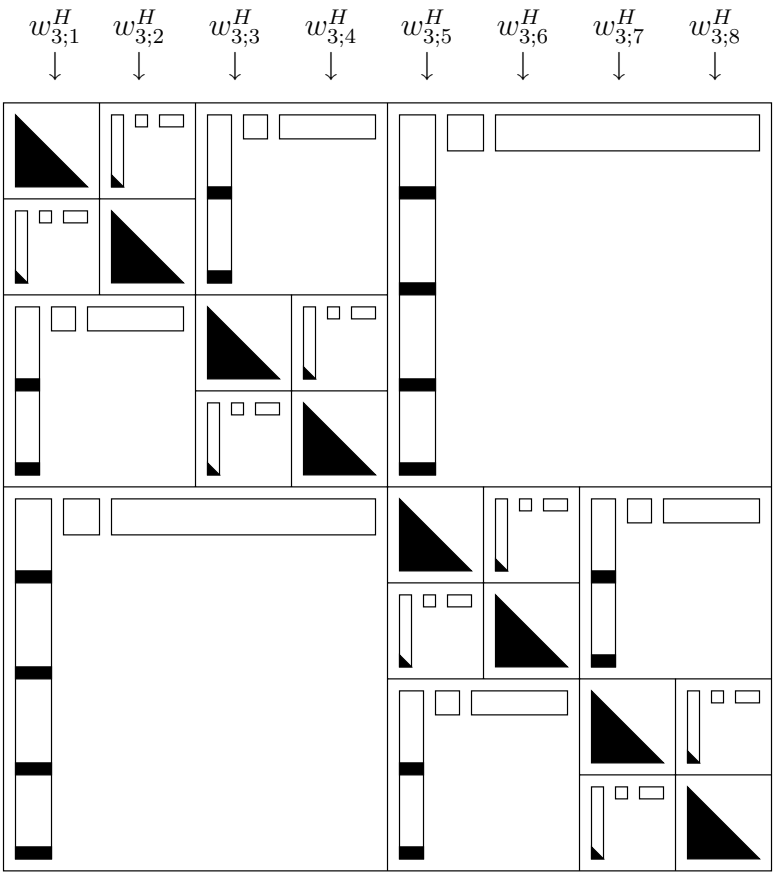


FIG. 4. *A pictorial representation showing the $w_{K;i}$'s lower triangularizing the diagonal blocks after the compression of the off-diagonal blocks by the $q_{K;i}$'s (see Figure 3). The black rectangles and triangles show the nonzero positions in the column bases and the diagonal blocks.*

Since we multiplied block column i from the right by $w^H_{K;i}$, we need to replace the unknowns $x_{K;i}$ by $w_{K;i}x_{K;i}$:

$$w_{K;i}x_{K;i} = \begin{matrix} m_i - n_{K;i} \\ n_{K;i} \end{matrix} \begin{pmatrix} z_{K;i} \\ \hat{x}_{K;i} \end{pmatrix}. \tag{9}$$

At this stage the first $m_i - n_{K;i}$ equations in block row i read as follows:

$$D_{K;i;1,1}z_{K;i} = \beta_{K;i},$$

which can be solved for $z_{K;i}$ to obtain $z_{K;i} = D^{-1}_{K;i;1,1}\beta_{K;i}$. We now need to multiply the first $m_i - n_{K;i}$ columns in the block column i by $z_{K;i}$ and subtract it from the right-hand side. To do this efficiently we observe that the system of equations has been transformed as follows:

$$\left(\operatorname{diag} q^H_{K;i}\, A\, \operatorname{diag} w^H_{K;i}\right)\left(\operatorname{diag} w_{K;i}\, x\right) = \operatorname{diag} q^H_{K;i}\, b.$$

If we define the vector

$$\bar{z}_{K;i} = \begin{matrix} m_i - n_{K;i} \\ n_{K;i} \end{matrix} \begin{pmatrix} z_{K;i} \\ 0 \end{pmatrix},$$

we then observe that the stated subtraction can be rewritten as follows:

$$\bar{b} = \operatorname{diag} q_{K;i}^H b - \left(\operatorname{diag} q_{K;i}^H A \operatorname{diag} w_{K;i}^H\right) \bar{z}.$$

We can do this operation rapidly by observing that

$$\left(\operatorname{diag} q_{K;i}^H A \operatorname{diag} w_{K;i}^H\right)$$

has the HSS representation $\{\bar{D}_{K;i}\}_{i=1}^{2^K}$, $\{\bar{U}_{K;i}\}_{i=1}^{2^K}$, $\{\bar{V}_{K;i}\}_{i=1}^{2^K}$, $\{\{R_{k;i}\}_{i=1}^{2^k}\}_{k=0}^{K}$, $\{\{W_{k;i}\}_{i=1}^{2^k}\}_{k=0}^{K}$, $\{\{B_{k;2i-1,2i}\}_{i=1}^{2^{k-1}}\}_{k=0}^{K}$, $\{\{B_{k;2i,2i-1}\}_{i=1}^{2^{k-1}}\}_{k=0}^{K}$ and by using the fast multiplication algorithm in section 3. Of course, the algorithm can (and should) be modified to take advantage of the zeros in $\bar{D}_{K;i}$, $\bar{U}_{K;i}$, and $\bar{z}_{K;i}$.

Once the subtraction has been done, we discard the first $m_i - n_{K;i}$ columns of block column i and the first $m_i - n_{K;i}$ rows of block row i. We observe that this leads to a new system of equations of the form

$$\hat{A}\hat{x} = \hat{b},$$

where

$$\bar{b}_{K;i} = \begin{matrix} m_i - n_{K;i} \\ n_{K;i} \end{matrix} \begin{pmatrix} * \\ \hat{b}_{K;i} \end{pmatrix},$$

and $\hat{A}$ has the HSS representation $\{D_{K;i;2,2}\}_{i=1}^{2^K}$, $\{\hat{U}_{K;i}\}_{i=1}^{2^K}$, $\{\hat{V}_{K;i}\}_{i=1}^{2^K}$, $\{\{R_{k;i}\}_{i=1}^{2^k}\}_{k=0}^{K}$, $\{\{W_{k;i}\}_{i=1}^{2^k}\}_{k=0}^{K}$, $\{\{B_{k;2i-1,2i}\}_{i=1}^{2^{k-1}}\}_{k=0}^{K}$, $\{\{B_{k;2i,2i-1}\}_{i=1}^{2^{k-1}}\}_{k=0}^{K}$.

Therefore we are left with a system of equations identical to the one we started with, and we can proceed to solve it recursively. Once we have done that we can recover the unknowns x from z and $\hat{x}$ using the formulas

$$x_{K;i} = w_{K;i}^H \begin{pmatrix} z_{K;i} \\ \hat{x}_{K;i} \end{pmatrix}.$$

Note that we have tacitly assumed that all block rows are such that $m_i > n_{K;i}$. However, it is easy to modify the equations so that only those block rows that satisfy $m_i > n_{K;i}$ have their off-diagonal blocks compressed.

4.2. Incompressible off-diagonal blocks. This is the second possibility for the recursion. It occurs when some block rows for the system cannot be compressed any further by invertible transformations from the left. In this case we proceed to merge block rows and columns that correspond to siblings in the merge tree. In particular consider the first two block rows:

$$\begin{pmatrix} D_{3;1;2,2} & U_{3;1}B_{3;1,2}V_{3;2}^H & U_{3;1}R_{3;1}B_{2;1,2}W_{3;3}^H V_{3;3}^H & U_{3;1}R_{3;1}B_{2;1,2}W_{3;4}^H V_{3;4}^H & \cdots \\ U_{3;2}B_{3;2,1}V_{3;1}^H & D_{3;2;2,2} & U_{3;2}R_{3;2}B_{2;1,2}W_{3;3}^H V_{3;3}^H & U_{3;2}R_{3;2}B_{2;1,2}W_{3;4}^H V_{3;4}^H & \cdots \end{pmatrix}.$$

We observe that these two block rows can be rewritten as

$$\left(\begin{pmatrix} D_{3;1;2,2} & U_{3;1}B_{3;1,2}V_{3;2}^H \\ U_{3;2}B_{3;2,1}V_{3;1}^H & D_{3;2;2,2} \end{pmatrix} \quad \left(\begin{pmatrix} U_{3;1}R_{3;1} \\ U_{3;2}R_{3;2} \end{pmatrix} B_{2;1,2} \begin{pmatrix} V_{3;3}W_{3;3} \\ V_{3;4}W_{3;4} \end{pmatrix}^H \right) \quad \cdots \right).$$

This immediately suggests that we merge as follows:

$$\hat{D}_{K-1;i} = \begin{pmatrix} D_{K;2i-1;2,2} & U_{K;2i-1}B_{K;2i-1,2i}V_{K;2i}^H \\ U_{K;2i}B_{K;2i,2i-1}V_{K;2i-1}^H & D_{K;2i;2,2} \end{pmatrix},$$

$$\hat{U}_{K-1;i} = \begin{pmatrix} U_{K;2i-1}R_{K;2i-1} \\ U_{K;2i}R_{K;2i} \end{pmatrix},$$

$$\hat{V}_{K-1;i} = \begin{pmatrix} V_{K;2i-1}W_{K;2i-1} \\ V_{K;2i}W_{K;2i} \end{pmatrix}.$$

We then see that A has an HSS representation (with a merge tree with K, as opposed to $K+1$, levels) given by the sequences $\{\hat{D}_{K-1;i}\}_{i=1}^{2^{K-1}}$, $\{\hat{U}\}_{i=1}^{2^{K-1}}$, $\{\hat{V}\}_{i=1}^{2^{K-1}}$, $\{\{R_{k;i}\}_{i=1}^{2^k}\}_{k=0}^{K-1}$, $\{\{W_{k;i}\}_{i=1}^{2^k}\}_{k=0}^{K-1}$, $\{\{B_{k;2i-1,2i}\}_{i=1}^{2^{k-1}}\}_{k=0}^{K-1}$, $\{\{B_{k;2i,2i-1}\}_{i=1}^{2^{k-1}}\}_{k=0}^{K-1}$. Let us denote by $\hat{A}$ the matrix with this HSS representation (of course, $A=\hat{A}$). We then observe that the system of equations is now in the form

$$\hat{A}x = b, \tag{10}$$

which is exactly in the form we started with, except that the new HSS representation has only K levels in the merge tree. Hence we can solve this system of equations recursively for x. That is, we check if there are compressible off-diagonal blocks. If so, we use the algorithm in section 4.1. If it is not compressible, we use the algorithm in this section. If the tree is just a leaf, we use the algorithm in section 4.3.

4.3. No off-diagonal blocks. Observe that if $K=0$, the equations read $D_1 x = b$, which can be solved by traditional means for x. This case terminates the recursion.

With this we have given a complete account of the algorithm.

4.4. Flop count. We use the flop counts in Table 1 of the basic matrix operations that can be found, for example, in [18].

TABLE 1
Flop counts of basic matrix operations.

Operation	Flops
QL factorization of skinny $m \times n$ matrix	$2n^2(m-n/3)$
Q times $m \times k$ matrix	$2kn(2m-n)$
Forward substitution of $n \times n$ matrix with k right-hand sides	n^2k
$m \times n$ times $n \times k$ matrix	$2mnk$

We begin by estimating the flop count for the fast multiplication algorithm, as that is an integral part of the solver. For simplicity we will assume that the ranks $n_{k;i}$ are independent of i, and that there are l indices in each of the leaves of the merge tree.

Computing $G_{K;i}$ will cost us $2ln_K r2^K$ flops, where r is the number of columns in the right-hand side. Computing $G_{k;i}$ from $G_{k+1;i}$ costs $4n_k n_{k+1} r 2^k$ flops. Computing $F_{k+1;i}$ from $F_{k;i}$ costs $42^{k+1} n_k n_{k+1} r$ flops. Finally computing $z_{K;i}$ costs $2lr(l+n_K)2^K$ flops. Summing these costs over k we obtain

$$2lr(l+n_K)2^K + 2ln_K r2^K + 8r\sum_{k=1}^{K-1} n_k n_{k+1} 2^k$$

as the total cost. Letting $N = 2^K l$ be the order of the matrix, we can simplify this to obtain

$$2Nr(l+2n_K) + 8r\sum_{k=1}^{K-1} n_k n_{k+1} 2^k$$

as the number of flops for the fast multiplication algorithm.

We now proceed to estimate the flops for the fast backward stable solver. To keep the calculations simple we will assume that each level of the tree undergoes a

compression step before going through a merge step. We will also assume that $m_{k;i}$, the size of the block rows at the kth stage, is independent of i.

Let us start with the compression step. We first need to compute the QL factorizations of $U_{k;i}$. This will cost us $2n_k^2(m_k - n_k/3)2^k$ flops. Then we need to apply $q_{k;i}$ to the right-hand side. This costs us $2rm_k(2m_k - n_k)2^k$ flops. We also need to apply $q_{k;i}$ to D_i (at the kth level), which costs us $2m_k n_k(2m_k - n_k)2^k$ flops. Next the LQ factorization of the diagonal blocks costs us $4m_k^3 2^k/3$ flops. Applying $w_{k;i}$ to $V_{k;i}$ costs $2n_k m_k^2 2^k$ flops. The partial forward-substitution at level k costs $(m_k - n_k)^2 r 2^k$ flops. Subtracting the computed unknowns from the right-hand side costs

$$2^k 2m_k r(m_k + 2n_k) + 8r\sum_{s=1}^{k-1} n_s n_{s+1} 2^s$$

flops. Recovering $x_{k;i}$ from $z_{k;i}$ will cost $2rm_k^2 2^k$ flops. That completes the compression stage.

For the merge step, forming the new diagonal blocks costs $8m_k n_k^2 2^k$ flops. Merging $U_{k;i}$ and $V_{k;i}$ costs $8m_k n_k^2 2^k$ flops.

Therefore the total cost of the fast backward stable solver is

$$\begin{aligned}\sum_{k=1}^{K} \Bigg(& 2n_k^2(m_k - n_k/3)2^k + 2rm_k(2m_k - n_k)2^k + 2m_k n_k(2m_k - n_k)2^k \\ & + 4m_k^3 2^k/3 + 2n_k m_k^2 2^k + (m_k - n_k)^2 r 2^k + 2^k 2m_k r(m_k + 2n_k) \\ & + 8r\sum_{s=1}^{k-1} n_s n_{s+1} 2^s + 16 m_k n_k^2 2^k \Bigg),\end{aligned}$$

which can be simplified to

$$\sum_{k=1}^{K} 2^k \left(\frac{4}{3} m_k^3 + 6m_k^2 n_k - \frac{2}{3} n_k^3 + r\left(7m_k^2 + n_k^2 + 8\sum_{s=1}^{k-1} n_s n_{s+1} 2^s \right) \right).$$

The terms not involving r can be thought of as the cost of factorization.

We now observe that under our assumptions $m_k = 2n_{k+1}$ for $k < K$. Making this substitution we can simplify the count to

$$\begin{aligned} & 2^K m_K^2 \left(\frac{4}{3} m_K + 6n_K \right) + 24 \sum_{k=1}^{K-1} 2^k n_{k+1}^2 n_k + \frac{14}{3} \sum_{k=2}^{K} 2^k n_k^3 - \frac{4}{3} n_1^3 \\ & + r\left(7m_K^2 2^K + 15\sum_{k=2}^{K} 2^k n_k^2 + n_1^2 + 8\sum_{k=1}^{K}\sum_{s=1}^{k-1} 2^s n_s n_{s+1} \right).\end{aligned}$$

To simplify further we assume that $n_k \geq n_{k+1}$. Then we can get an upper bound on the flop count

$$\begin{aligned} & 2^K m_K^2 \left(\frac{4}{3} m_K + 6n_K \right) + \frac{86}{3} \sum_{k=1}^{K} 2^k n_k^3 - \frac{4}{3} n_1^3 \\ & + r\left(7m_K^2 2^K + 15\sum_{k=2}^{K} 2^k n_k^2 + n_1^2 + 8\sum_{k=1}^{K}\sum_{s=1}^{k-1} 2^s n_s^2 \right).\end{aligned}$$

We now compute the flop counts for some canonical examples. First we consider the case when $n_k = p$, a constant. In this case the upper bound on the flop count simplifies to

$$2^K m_K^2 \left(\frac{4}{3} m_K + 6p\right) + \frac{86}{3} p^3 2^{K+1} + r\left(7 m_K^2 2^K + 23 p^2 2^{K+1}\right).$$

Using $N = 2^K m_K$, and assuming that $m_K = 2p$, we get

$$46 N p^2 + 37 N p r.$$

As can be seen, the constants are modest. By switching to Gauss transforms rather than Householder transforms we can reduce the constants even further.

In many cases this flop count is sufficient to give an indication of the performance of the algorithm. However, for theoretical purposes we also provide an upper bound on the flop count under the assumption that $n_k \leq \gamma^k n_0$. This model is useful when applying the algorithm to matrices of the form $A_{ij} = f(x_i, x_j)$, when the points x_i lie in high-dimensional spaces.

For example, when $f(x_i, x_j) = \log|x_i - x_j|$, and x_i is a point in the two-dimensional plane, we can take $n_0 = \alpha N^{\frac{1}{2}}$ and $\gamma = \frac{1}{\sqrt{2}}$. For there to be any speed-up possible at all we must have that

$$\alpha \leq \frac{N^{\frac{1}{2}}}{\sqrt{2}}.$$

For simplicity, and since it is common in practice, we assume that $\alpha \geq 1$. We then observe that $m_k = N 2^{-k} \geq 2\alpha N^{1/4} 2^{-k/2}$, provided $k < \log_2 N - 2(\log_2 \alpha + 1)$. Hence we take the depth of the tree to be

$$K = \lfloor \log_2 N - 2 \log_2 \alpha - 1 \rfloor.$$

Note that m_K is approximately $4\alpha^2$ in this scenario. Under these assumptions the flop count for the fast solver is not more than

$$98 N^{\frac{3}{2}} \alpha^3 + 70 N \alpha^4 + N \alpha^2 r (4 \log_2^2 N + 11 \log_2 N + 28).$$

As can be seen the constant is quite sensitive to the size of α.

Next we consider three-dimensional problems. For example, when $f(x_i, x_j) = \|x_i - x_j\|^{-1}$, and x_i is a point in three-dimensional space, we can take $n_0 = \alpha N^{\frac{2}{3}}$ and $\gamma = \frac{1}{\sqrt[3]{4}}$. To obtain any speed-up at all, we must ensure that $\alpha < (N/2)^{1/3}$. For the sake of simplicity we will also assume that $\alpha \geq 1$. We can determine the maximum depth of the tree from the constraint $m_k \geq 2n_k$, which yields

$$K \leq \lfloor \log_2 N - 3 \log_2 \alpha - 1 \rfloor.$$

Under this scenario m_K is approximately $2\alpha^3$. With these assumptions the flop count for the fast solver is less than

$$58 N^2 \alpha^3 + 18 N \alpha^6 + r(39 N^{\frac{4}{3}} \log_2 N \alpha^2 + 74 N^{\frac{4}{3}} \alpha^2 + 14 N \alpha^3).$$

As can be seen the constant is modest.

Observe that in both cases the fast dense solver matches the asymptotic complexity of the corresponding sparse direct finite-element and finite-difference solvers

TABLE 2

CPU run-times in seconds for both the fast stable algorithm and standard solver for random HSS matrices with $m_i = n_{k;i} = p_{k;i}$ for all k and i. Timings are not reported when there was insufficient main memory. (GEPP = Gaussian elimination with partial pivoting.)

	Size							
$m_i/n_{k;i}/p_{k;i}$	256	512	1024	2048	4096	8192	16,384	32,768
16	0.03	0.06	0.13	0.27	0.49	0.99	2.10	4.74
32	0.05	0.12	0.27	0.56	1.14	2.34	4.75	9.81
64	0.09	0.26	0.58	1.27	2.60	5.33	11.11	23.19
128	0.09	0.63	1.71	3.92	8.45	17.14	35.16	74.07
GEPP	0.07	0.33	2.12	14.81	113.75	891.59	...	...

	Size				
$m_i/n_{k;i}/p_{k;i}$	65,536	131,072	262,144	524,288	1,048,576
16	9.93	27.95	64.28	224.73	889.65
32	20.56	44.53	106.35	408.39	...
64	50.24	129.81	405.13	...	...
128	158.97	380.89	...	...	...

of the same dimension. Of course, many times the integral equations corresponding to a particular PDE will be one dimension smaller, frequently yielding the advantage to the integral equation method. However, the quadratic dependence on N for three-dimensional problems makes this algorithm suitable only when the linear system is highly ill-conditioned and a suitable preconditioner is lacking. In fact, this solver can serve as an ideal preconditioner in this and other situations. Another situation where this method is suitable even for three-dimensional problems is when there is a large number of right-hand sides.

We remark that if many of the leaves at level K are empty, then the algorithm we have specified will become inefficient. A more complicated algorithm that does not suffer from this deficiency will be presented in a future paper.

4.5. Experimental run-times. We now present CPU run-times for our fast solver. These timings were obtained on an Apple dual 1GHz PowerPC G4 machine with 1.5GB of RAM, though no explicit use was made of the dual processors. Vendor supplied BLAS [14] (uniprocessor) and LAPACK 3.0 were used in all routines. We report on problem sizes ranging from 256 unknowns to 1,048,576 unknowns. Off-diagonal ranks $n_{k;i}$ and $p_{k;i}$ were chosen to range from 16 to 128. In every instance we chose $m_i = n_{k;i} = p_{k;i}$ for all i. The matrices were generated randomly to these specifications.

The CPU run-times in seconds are reported in Table 2. Also shown are CPU run-times in seconds for the standard Gaussian elimination with row pivoting solver from LAPACK. This routine is highly tuned and essentially runs at peak flop-rates. As can be seen our fast solver breaks even with the standard solver for reasonably small matrix sizes, as predicted by the flop count. Entries marked by ellipses indicate instances where there was insufficient memory to run the test. Again this also indicates another reason why the fast solver might be preferred: memory efficiency.

4.6. Stability. The fast solver we presented can be shown to be numerically backward stable, provided the HSS representation is in the [illegible]. However, the proof would detract from the main ideas of this paper and will be presented elsewhere. By proper form we mean that $\|R_{k;i}\| \leq 1$ and $\|W_{k;i}\| \leq 1$ for a submultiplicative norm. We observe that the HSS construction algorithm presented in section 5 satisfies this requirement for the 2-norm.

However, the algorithm can also be shown to be backward stable to first order in machine precision even if the weaker condition $\|R_{k;i}R_{k+1;2i(-1)}\cdots\| \leq p(n)$ and $\|W_{k;i}W_{k+1;2i(-1)}\cdots\| \leq q(n)$ is satisfied, where $p(n)$ and $q(n)$ are low-degree polynomials in n. This condition is satisfied by the fast HSS construction algorithm presented in subsection 5.1.

The reason for the claimed stability of the fast solver is due to the use of unitary transformations and a single forward substitution. The proof is similar to the one for the sequentially semiseparable representation [3] and will be presented elsewhere.

In Table 3 we present computed experimental backward errors for the fast solver on a wide class of problems which lends credence to our claims of stability. These experiments were carried out in double precision for matrix sizes ranging from 256 to 4096. The ranks of the off-diagonal blocks $n_{k;i}$ and $p_{k;i}$ were chosen to range from 16 to 128. Although the HSS forms were generated randomly, we did not ensure proper form. We only ensured the milder condition that the entries of $W_{k;i}$ and $R_{k;i}$ were no larger than 1 in magnitude. As can be seen from the backward errors presented in Table 3 the fast solver was backward stable even in this case.

TABLE 3

One-norm backward errors $\|Ax-b\|_1/(\epsilon_{mach}(\|A\|_1\|x\|_1+\|b\|_1))$ of the fast solver in double precision with $|W_{k;i}| \leq 1$ and $|R_{k;i}| \leq 1$. Entries much larger than 1 indicate a potential lack of backward stability.

	Size				
$m_i/n_{k;i}/p_{k;i}$	256	512	1024	2048	4096
16	0.31	0.27	0.32	0.25	0.16
32	0.34	0.33	0.24	0.22	0.20
64	0.54	0.38	0.33	0.28	0.25
128	0.47	0.43	0.36	0.28	0.28

5. Computing the HSS representation. In this section we describe an $O(n^2)$ algorithm to compute the HSS representation of an arbitrary matrix to a given tolerance.

The key idea is to compute the singular value decomposition (SVD) of the matrices

$$H_{k;i} = \begin{pmatrix} A_{k;i,1} & A_{k;i,2} & \cdots & A_{k;i,i-1} & A_{k;i,i+1} & A_{k;i,i+2} & \cdots & A_{k;i,2^k} \end{pmatrix}. \tag{11}$$

Notice that $H_{k;i}$ is essentially block row i of the matrix when partitioned according to level k of the merge tree, except that the diagonal block corresponding to that level $A_{k;i,i}$ is missing.

Similarly we also need to compute the SVD of the matrices

$$G_{k;i} = \begin{pmatrix} A^H_{k;1,i} & A^H_{k;2,i} & \cdots & A^H_{k;i-1,i} & A^H_{k;i+1,i} & A^H_{k;i+2,i} & \cdots & A^H_{k;2^k;i} \end{pmatrix}^H. \tag{12}$$

Suppose we have the SVD of $H_{k;i}$ and $G_{k;i}$ for $k = 1$ to K and for $i = 1$ to 2^k:

$$H_{k;i} = U_{k;i}C_{k;i}J^H_{k;i},$$
$$G_{k;i} = L_{k;i}M_{k;i}V^H_{k;i}.$$

Observe that these equations directly define the auxiliary quantities $U_{k;i}$ and $V_{k;i}$ that appear in (1) and (2). In particular we obtain $U_{K;i}$ and $V_{K;i}$. Using (1) and (2) we

can also compute

$$\begin{aligned} R_{k+1;2i-1} &= U^H_{k+1;2i-1}(U_{k;i})_1, \\ R_{k+1;2i} &= U^H_{k+1;2i}(U_{k;i})_2, \\ W_{k+1;2i-1} &= V^H_{k+1;2i-1}(V_{k;i})_1, \\ W_{k+1;2i} &= V^H_{k+1;2i}(V_{k;i})_2, \end{aligned}$$

where we have the conforming partitions

$$\begin{aligned} U_{k;i} &= \begin{pmatrix} (U_{k;i})_1 \\ (U_{k;i})_2 \end{pmatrix} = \begin{pmatrix} U_{k+1;2i-1}R_{k+1;2i-1} \\ U_{k+1;2i}R_{k+1;2i} \end{pmatrix}, \\ V_{k;i} &= \begin{pmatrix} (V_{k;i})_1 \\ (V_{k;i})_2 \end{pmatrix} = \begin{pmatrix} V_{k+1;2i-1}W_{k+1;2i-1} \\ V_{k+1;2i}W_{k+1;2i} \end{pmatrix}. \end{aligned}$$

This leaves us only with determining formulas for $B_{k;2i-1,2i}$ and $B_{k;2i,2i-1}$. Observe that $A_{k;2i-1,2i}$ is the $2i-1$ submatrix of $H_{k;2i-1}$ and $G_{k;2i}$ in the partitioning in (11) and (12). Therefore assuming that $(J_{k;2i-1})_{2i-1}$ and $(L_{k;2i})_{2i-1}$ denote the appropriate submatrices, we have that

$$A_{k;2i-1,2i} = U_{k;2i-1}B_{k;2i-1,2i}V^H_{k;2i} = U_{k;2i-1}C_{k;2i-1}(J_{k;2i-1})^H_{2i-1} = (L_{k;2i})_{2i-1}M_{k;2i}V^H_{k;2i}.$$

This immediately gives us the formulas

$$B_{k;2i-1,2i} = C_{k;2i-1}(J_{k;2i-1})^H_{2i-1}V_{k;2i} = U^H_{k;2i-1}(L_{k;2i})_{2i-1}M_{k;2i}.$$

Similarly $A_{k;2i,2i-1}$ is the $2i-1$ submatrix of $H_{k;2i}$ and $G_{k;2i-1}$ in the partitioning in (11) and (12). Therefore assuming that $(J_{k;2i})_{2i-1}$ and $(L_{k;2i-1})_{2i-1}$ denote the appropriate submatrices, we have that

$$A_{k;2i,2i-1} = U_{k;2i}B_{k;2i,2i-1}V^H_{k;2i-1} = U_{k;2i}C_{k;2i}(J_{k;2i})^H_{2i-1} = (L_{k;2i-1})_{2i-1}M_{k;2i-1}V^H_{k;2i-1}.$$

This immediately gives us the formulas

$$B_{k;2i,2i-1} = C_{k;2i}(J_{k;2i})^H_{2i-1}V_{k;2i-1} = U^H_{k;2i}(L_{k;2i-1})_{2i-1}M_{k;2i-1}.$$

All we need now is an efficient way to compute the needed SVDs. To this end we observe that $H_{k;i}$ is closely related to $H_{k+1;2i-1}$ and $H_{k+1;2i}$. In fact, by dropping the $2i-1$ block column from

$$\begin{pmatrix} H_{k+i;2i-1} \\ H_{k+1;2i} \end{pmatrix},$$

we obtain $H_{k;i}$. Similarly, by dropping the $2i-1$ block row from

$$\begin{pmatrix} G_{k+1;2i-1} & G_{k+1;2i} \end{pmatrix},$$

we obtain $G_{k;i}$. Hence we can obtain the SVD of $H_{k;i}$ efficiently from the SVDs of $H_{k+1;2i-1}$ and $H_{k+1;2i}$. Similarly for $G_{k;i}$.

Assuming that $B_{k;2i-1,2i}$ is an $n_{k;i} \times n_{k;i}$ matrix and $B_{k;2i,2i-1}$ is a $p_{k;i} \times p_{k;i}$ matrix, the complexity of the above algorithm is $O(N(N + \sum_{k;i}(n^2_{k;i} + p^2_{k;i})))$.

The cost of the algorithm can be reduced by replacing the SVD with a rank-revealing QR factorization [11, 21] instead.

5.1. Smooth matrices. When the matrix entry A_{ij} is specified by a function $f(x_i, x_j)$ that is smooth away from the diagonal, the HSS representation can be computed more rapidly than in the general case. In this section we consider the special case when the points x_i lie on the real line. The more general case is beyond the scope of this paper. Important examples of the function $f(x, y)$ include $\log \|g(x) - g(y)\|$ and $\|g(x) - g(y)\|^\alpha$, where $g : \mathcal{R} \to \mathcal{R}^d$ represents a simple closed or nonclosed curve in d-dimensional space. For applications see [30, 31].

Since we are restricting ourselves to uniform HSS representations in this paper, we will assume that the points x_i are distributed uniformly in the interval $[0, 1]$. Note that this does not mean that the points x_i are equispaced. Furthermore, for simplicity, we will assume the function f has at most singularities at 0 and 1, and that it is analytic away from these singularities. A good example to keep in mind is $f(x, y) = \log |x - y|$.

From the basic theory of polynomial approximation of such functions it follows that if $T_k(x)$ denotes the kth Chebyshev polynomial

$$T_k(x) = \cos(k \arccos x), \qquad -1 \le x \le 1,$$

and if

$$\phi_{a,b} : [a, b] \to [-1, 1], \qquad \phi_{a,b}(x) = -1 + 2\frac{x - a}{b - a}$$

denotes the affine-linear function that maps the interval $[a, b]$ to $[-1, 1]$, then on any rectangle $[a, b] \times [c, d]$ such that $a < b < c < d$ and $\min(d - c, b - a) > c - b$, we can find a short two-sided Chebyshev expansion of $f(x, y)$ to a given accuracy:

$$f(x, y) \approx \sum_{p,q} \beta_{p,q} T_p(\phi_{a,b}(x)) T_q(\phi_{c,d}(y)).$$

More specifically, the (i, j)th entry of the matrix can be represented to a prescribed accuracy by a short expansion of the form

$$f(x_i, x_j) \approx \sum_{p,q} \beta_{p,q} T_p(\phi_{a,b}(x_i)) T_q(\phi_{c,d}(y_j)).$$

We shall now show how these expansion coefficients can be used to compute an HSS representation for the matrix quickly.

We first need to specify the merge tree we are going to use. We do so as follows. We will assume that all the points x_i lie in the interval $[0, 1]$. Hence we will associate the interval $[0, 1]$ (and hence all the points x_i, and hence all indices) with the root node. With the left child of the root we associate the interval $[0, 0.5)$ and with the right child the interval $[0.5, 1]$. This means that we associate all points x_i in the interval $[0, 0.5)$ with the left child and hence all the corresponding indices with the left child. Similarly for the right child. To the left child of the left child of the root node, namely, Node$(2, 1)$, we associate the interval $[0, 0.25)$, to Node$(2, 2)$ we associate the interval $[0.25, 0.5)$, and so on. In this way we assign the indices to the merge tree.

Note that the number of indices in two different nodes at the same level can be different. Also note that we do not assume that the points x_i are equispaced.

Let us denote the set of points x_i that belong to Node(k, i) by $x_{k;i}$. Let us denote by $\Gamma_{k;i}$ the Chebyshev–Vandermonde matrix evaluated at the points $x_{k;i}$. We will assume that the number of columns in $\Gamma_{k;i}$ is fixed at p to ease the exposition.

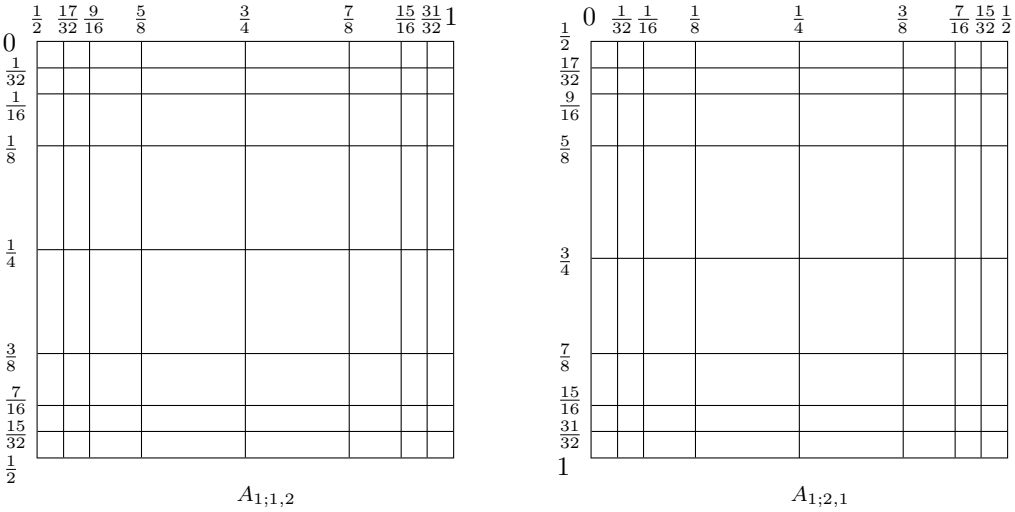


FIG. 5. *Block partitioning of $A_{1;1,2}$ and $A_{1;2,1}$ suitable for Chebyshev expansions. The vertical and horizontal lines are labeled according to the interval boundaries.*

Each node of the merge tree is associated with a particular interval of the real line. In particular Node(k,i) is associated with the interval $2^{-k}[i-1,i]$. Therefore it follows that $A_{k;i,j}$ is associated with the [illegible] $2^{-k}[i-1,i] \times 2^{-k}[j-1,j]$.

Figure 5 displays a partitioning of $A_{1;1,2}$ and $A_{1;2,1}$ that will prove useful. We observe that each off-diagonal block, except possibly the bottom-left and upper-right blocks, in the displayed partition is associated with a rectangle on which the function f has a short two-sided Chebyshev expansion. However, note that the blocks are sometimes specified by intervals on two different levels of the merge tree. Hence we will use the notation $A_{(k;i),(r;j)}$ to denote the submatrix whose row indices come from Node(k,i) and column indices come from Node(r,j). We shall also use the notation

$$A_{(k;i),(r;j)} = \Gamma_{k;i} C_{(k;i),(r;j)} \Gamma_{r;j}^H \tag{13}$$

for the corresponding two-sided Chebyshev expansion. Observe that $C_{(k;i),(r;j)}$ can be computed in time independent of the size of $A_{(k;i),(r;j)}$. Since the block $A_{K;i,i+1}$ does not necessarily have a short two-sided Chebyshev expansion, we will assume instead that it has fewer than p rows and columns, in which case it trivially has an expansion of the form (13).

To construct the HSS representation we remind the reader that it is the low-rank expansions of $H_{k;i}$ and $G_{k;i}$ that are crucial. Hence in Figure 6 we show the partitioning of $H_{2;3}$ that we will use. Now observe that we can construct a low-rank expansion for $A_{k;i,i+1}$ and $A_{k;i+1,i}$ for odd i, as follows. Let

$$\Delta_{k;i} = \operatorname{diag}\begin{pmatrix} \Gamma_{K;2^{K-k}(i-1)+1} \\ \Gamma_{K;2^{K-k}(i-1)+2} \\ \vdots \\ \Gamma_{k+2;2(2i-1)} \\ \Gamma_{k+2;2(2i)-1} \\ \vdots \\ \Gamma_{K;2^{K-k}i-1} \\ \Gamma_{K;2^{K-k}i} \end{pmatrix} \tag{14}$$

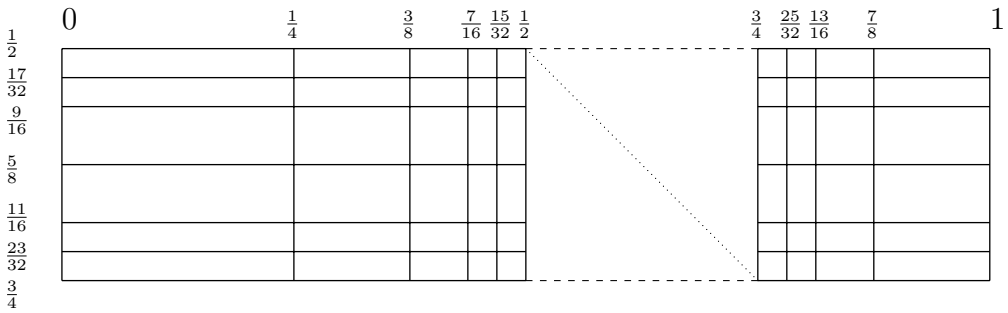


FIG. 6. *Block partitioning of $H_{2;3}$ suitable for Chebyshev expansions. The vertical and horizontal lines are labeled according to the associated interval boundaries. The missing block in the middle is the diagonal block. The dotted diagonal line shows the position of the diagonal.*

be a block-diagonal matrix. In the above notation we assume that $\Delta_{K;i} = \Gamma_{K;i}$ and

$$\Delta_{K-1;i} = \begin{pmatrix} \Gamma_{K;2i-1} \\ \Gamma_{K;2i} \end{pmatrix}.$$

Note that these formulas are consistent with (14). Then let

$$\begin{aligned} U_{k;i} &= \Delta_{k;i}, \\ V_{k;i} &= \Delta_{k;i}, \end{aligned}$$

and let $B_{k;i,i+1}$ be the block matrix with block entries

$$(B_{k;i,i+1})_{r,s} = C_{(k_r;i_r),(k_s;(i+1)_s)},$$

where $\mathrm{Node}(k_r, i_r)$ is the node corresponding to the rth diagonal block of $\Delta_{k;i}$. Similarly we define

$$(B_{k;i+1,i})_{r,s} = C_{(k_r;(i+1)_r),(k_s;i_s)}.$$

All we have to specify now is $R_{k;i}$ and $W_{k;i}$. First observe that $R_{k;i} = W_{k;i}$ since $U_{k;i} = V_{k;i}$. From the definition of $\Delta_{k;i}$ observe that

$$\Delta_{k;i} = \begin{pmatrix} \Delta_{k+1;2i-1}\Omega_{k+1;2i-1} & 0 \\ 0 & \Delta_{k+1;2i}\Omega_{k+1;2i} \end{pmatrix},$$

where

$$\begin{aligned} R_{k;2i-1} &= \begin{pmatrix} \Omega_{k;2i-1} & 0 \end{pmatrix}, \\ R_{k;2i} &= \begin{pmatrix} 0 & \Omega_{k;2i} \end{pmatrix}. \end{aligned}$$

Hence it is sufficient to specify the $\Omega_{k;i}$'s. To do that we first specify the two sets of auxiliary matrix-valued functions

$$\begin{aligned} \sigma_u(0) &= I, \\ \sigma_u(i+1) &= \begin{pmatrix} \sigma_u(i)C_L \\ C_R \end{pmatrix}, \end{aligned}$$

and

$$\sigma_l(0) = I,$$
$$\sigma_l(i+1) = \begin{pmatrix} C_L \\ \sigma_l(i) C_R \end{pmatrix}.$$

Then

$$\Omega_{k;2i} = \begin{pmatrix} \sigma_u(K-k-1) & 0 \\ 0 & I \end{pmatrix},$$
$$\Omega_{k;2i-1} = \begin{pmatrix} I & 0 \\ 0 & \sigma_l(K-k-1) \end{pmatrix},$$

with the understanding that $\sigma_u(-1)$ and $\sigma_l(-1)$ denote the empty matrices.

With this we have given a complete specification for computing the HSS representation (assuming a uniform tree) of a smooth matrix with a one-dimensional kernel function.

However, given the sparse structure of $U_{k;i}$ and $R_{k;i}$, the fast solvers and multipliers presented in this paper can, and should, be modified to exploit the extra structure. This is important, as the Chebyshev expansions are not optimal low-rank expansions.

5.2. Sparse matrices. In the previous subsection we showed how to construct rapidly the HSS representation of matrices whose entries are given by kernel functions that are smooth away from the diagonal. Such matrices are intimately associated with the fast multipole method and integral equations. In this subsection we consider sparse matrices. For sparse matrices we can quickly construct a possibly suboptimal HSS representation. For many sparse matrices this construction will actually lead to the optimal HSS representation.

We proceed as follows. First we must determine the row and column partition sizes. In this paper we will assume that these two partitions are identical. Suppose m_i denotes the size of the ith partition. We will again assume that the HSS tree is going to be uniform and that the number of partitions is 2^K for some K. The matrices D_i are straightforward to compute.

We form U_i as follows. Suppose the j_ith row in the ith partition is the first row in that partition to have a nonzero entry that is not in D_i; then the first column of U_i will be the zero column with a one in the j_ith position. Suppose g_i is the next row after the j_ith one in the ith partition that has a nonzero entry outside D_i; then the second column of U_i will be a zero column with a one in the g_ith position. We proceed until we have exhausted all the rows in the ith partition. Notice that we have constructed U_i such that it is guaranteed to be the column basis for $H_{K;i}$, as it must.

Next we form V_i. The construction is similar to that for U_i, except that we must deal with the columns of the ith partition, and in particular the nonzero entries in that partition that do not lie in D_i. Suppose the j_ith column in the ith partition is the first column in that partition to have a nonzero entry that is not in D_i; then the first column of V_i will be the zero column with a one in the j_ith position. Suppose g_i is the next column after the j_ith one in the ith partition that has a nonzero entry outside D_i; then the second column of V_i will be a zero column with a one in the g_ith position. We proceed until we have exhausted all the columns in the ith partition. Notice that we have constructed V_i such that it is guaranteed to be column basis for $G_{K;i}$, as it must.

Now we specify how to form $R_{k;i}$. First we observe that we could compute $U_{k;i}$ using the same ideas we used to compute $U_i = U_{K;i}$. From that we could then recover $R_{k;i}$. However, we can also do this in a direct fashion. As usual, let

$$U_{k;i} = \begin{pmatrix} (U_{k;i})_1 \\ (U_{k;i})_2 \end{pmatrix} = \begin{pmatrix} U_{k+1;2i-1} R_{k+1;2i-1} \\ U_{k+1;2i} R_{k+1;2i} \end{pmatrix}.$$

Then we observe that $R_{k+1;2i}$, for example, must drop the right columns in $U_{k+1;2i}$ so as to produce $(U_{k;i})_2$. Hence by looking at the nonzero entries of $A_{k+1;2i,2i-1}$ and $A_{k+1;2i;2i+1}$, we can determine potential columns of $U_{k+1;2i}$ that must be dropped. We pick $R_{k+1;2i}$ so that it drops just those columns. Note that not every nonzero row in $A_{k+1;2i,2i-1}$ and $A_{k+1;2i;2i+1}$ induces a drop in $U_{k+1;2i}$, since some other column in the same row might still have a nonzero entry.

We compute $W_{k;i}$ in a fashion similar to that for $R_{k;i}$ but with respect to $V_{k;i}$ rather than $U_{k;i}$.

All that is left to be specified is $B_{k;i,j}$. But this is easy now. $B_{k;i,j}$ is just the matrix obtained by dropping all zero rows and columns of $A_{k;i,j}$.

As can be seen, the HSS representation of a sparse matrix can be computed in time proportional to the number of nonzeros in the matrix, provided, of course, that the sparse matrix data structure supports efficient access for sequential reads of the nonzeros entries of any row or column. Many common sparse matrix data structures do exactly this, so we do not comment on it any further.

REFERENCES

[1] F. X. Canning and K. Rogovin, *Fast direct solution of moment-method matrices*, IEEE Antennas and Propagation Magazine, 40 (1998), pp. 15–26.

[2] J. Carrier, L. Greengard, and V. Rokhlin, *A fast adaptive multipole algorithm for particle simulations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 669–686.

[3] S. Chandrasekaran, P. Dewilde, M. Gu, T. Pals, and A. van der Veen, *Fast stable solver for sequentially semi-separable linear systems of equations*, in High Performance Computing—HiPC 2002: 9th International Conference, Lecture Notes in Comput. Sci. 2552, S. Sahni et al., eds., Springer-Verlag, Heidelberg, 2002, pp. 545–554.

[4] S. Chandrasekaran, P. Dewilde, M. Gu, T. Pals, and A. van der Veen, *Fast Stable Solvers for Sequentially Semi-separable Linear Systems of Equations*, Technical report, Mathematic Department, University of California, Berkeley, 2003.

[5] S. Chandrasekaran and M. Gu, *Fast and stable algorithms for banded plus semiseparable systems of linear equations*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 373–384.

[6] S. Chandrasekaran and M. Gu, *A fast and stable solver for recursively semi-separable systems of linear equations*, in Structured Matrices in Mathematics, Computer Science, and Engineering, II, Contemp. Math. 281, V. Olshevsky, ed., AMS, Providence, RI, 2001, pp. 39–53.

[7] S. Chandrasekaran and M. Gu, *Fast and stable eigendecomposition of symmetric banded plus semi-separable matrices*, Linear Algebra Appl., 313 (2000), pp. 107–114.

[8] 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.

[9] S. Chandrasekaran, M. Gu, and T. Pals, *A Fast and Stable Solver for Smooth Recursively Semi-separable Systems*, Paper presented at the SIAM Annual Conference, San Diego, CA, 2001, and SIAM Conference of Linear Algebra in Controls, Signals and Systems, Boston, MA, 2001.

[10] S. Chandrasekaran, M. Gu, and T. Pals, *Fast and Stable Algorithms for Hierarchically Semi-separable Representations*, Technical report, Department of Mathematics, University of California, Berkeley, 2004.

[11] S. Chandrasekaran and I. C. F. Ipsen, *On rank-revealing factorisations*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 592–622.

[12] Y. Chen, *Fast direct solver for the Lippmann–Schwinger equation*, Adv. Comput. Math., 16 (2002), pp. 175–190; also available online at http://www.math.nyu.edu/faculty/yuchen/onr/intro.htm.

[13] P. Dewilde and A. van der Veen, *Time-Varying Systems and Computations*, Kluwer Academic, Boston, MA, 1998.

[14] J. J. Dongarra, J. Du Croz, S. Hammarling, and I. Duff, *Algorithm* 679: *A set of level* 3 *basic linear algebra subprograms: Model implementation and test programs*, ACM Trans. Math. Softw., 16 (1990), pp. 18–28.

[15] Y. Eidelman and I. Gohberg, *On a new class of structured matrices*, Integral Equations Operator Theory, 34 (1999), pp. 293–324.

[16] Y. Eidelman and I. Gohberg, *A modification of the Dewilde van der Veen method for inversion of finite structured matrices,* Linear Algebra Appl., 343/344 (2001), pp. 419–450.

[17] I. Gohberg, T. Kailath, and I. Koltracht, *Linear complexity algorithms for semiseparable matrices*, Integral Equations Operator Theory, 8 (1985), pp. 780–804.

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

[19] D. Gope and V. Jandhyala, *An iteration-free fast multilevel solver for dense method of moment systems,* in Proceedings of the IEEE 10th Topical Meeting on Electrical Performance of Electronic Packaging, IEEE Press, Piscataway, NJ, 2001, pp. 177–180.

[20] L. Greengard and V. Rokhlin, *A fast algorithm for particle simulations*, J. Comput. Phys., 73 (1987), pp. 325–348.

[21] M. Gu and S. C. Eisenstat, *Efficient algorithms for computing a strong rank-revealing QR factorization*, SIAM J. Sci. Comput., 17 (1996), pp. 848–869.

[22] L. Gurel and W. C. Chew, *Fast direct (noniterative) solvers for integral-equation formulations of scattering problems,* in Antennas: Gateways to the Global Network, Vol. 1, IEEE Antennas and Propagation Society International Symposium, Vol. 1, IEEE Press, New York, 1998, pp. 298–301.

[23] W. Hackbusch, *A sparse arithmetic based on $\mathcal{H}$-matrices. Part* I: *Introduction to $\mathcal{H}$-matrices*, Computing, 62 (1999), pp. 89–108.

[24] W. Hackbusch, B. N. Khoromskij, and S. Sauter, *On H^2-Matrices*, preprint 50, MPI, Leipzig, 1999.

[25] W. Hackbusch and S. Borm, *Data-sparse approximation by adaptive H^2-matrices*, Computing, 69 (2002), pp. 1–35.

[26] P. Jones, J. Ma, and V. Rokhlin, *A fast direct algorithm for the solution of the Laplace equation on regions with fractal boundaries,* J. Comput. Phys., 113 (1994), pp. 35–51.

[27] N. Mastronardi, S. Chandrasekaran, and S. van Huffel, *Fast and stable two-way chasing algorithm for diagonal plus semi-separable systems of linear equations*, Numer. Linear Algebra Appl., 38 (2000), pp. 7–12.

[28] N. Mastronardi, S. Chandrasekaran, and S. van Huffel, *Fast and stable algorithms for reducing diagonal plus semi-separable matrices to tridiagonal and bidiagonal form*, BIT, 41 (2001), pp. 149–157.

[29] V. Rokhlin, *Applications of volume integrals to the solution of PDEs*, J. Comput. Phys., 86 (1990), pp. 414–439.

[30] P. Starr, *On the Numerical Solution of One-Dimensional Integral and Differential Equations*, Thesis advisor: V. Rokhlin, Research Report YALEU/DCS/RR-888, Department of Computer Science, Yale University, New Haven, CT, 1991.

[31] P. Starr and V. Rokhlin, *On the numerical solution of* 2*-point boundary value problem.* 2, Comm. Pure Appl. Math., 47 (1994), pp. 1117–1159.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 3, pp. 623–633

EFFICIENT IMPLEMENTATION OF THE MULTISHIFT QR ALGORITHM FOR THE UNITARY EIGENVALUE PROBLEM*

RODEN J. A. DAVID† AND DAVID S. WATKINS†

Abstract. We present an efficient implementation of the multishift QR algorithm for computing the eigenvalues of a unitary matrix. The algorithm can perform QR iterations of arbitrary degree, is conceptually simple, and is backward stable.

Key words. unitary matrix, eigenvalue, multishift QR algorithm

AMS subject classifications. 65F15, 15A18

DOI. 10.1137/050630787

1. Introduction. We consider the eigenvalue problem for a unitary matrix $U \in \mathbb{C}^{n\times n}$ that is upper Hessenberg; i.e., $u_{ij} = 0$ whenever $i > j + 1$. Without loss of generality we can assume that all of the subdiagonal entries $u_{j+1,j}$ are nonzero. Assuming this, then by a unitary diagonal similarity transformation we can make them real and positive. Thus we will assume that $u_{j+1,j} > 0$ for $j = 1, \ldots, n-1$. Then U can be expressed as a product of matrices of a very simple form [9]:

$$U = G_1 G_2 \cdots G_{n-1} G_n, \tag{1.1}$$

where $G_k = \operatorname{diag}\{I_{k-1}, \tilde{G}_k, I_{n-k-1}\}$,

$$\tilde{G}_k = \begin{bmatrix} \gamma_k & \sigma_k \\ \sigma_k & -\overline{\gamma}_k \end{bmatrix}, \quad \sigma_k > 0, \quad |\gamma_k|^2 + \sigma_k^2 = 1,$$

for $k = 1, \ldots, n-1$, and $G_n = \operatorname{diag}\{I_{n-1}, \gamma_n\}$ with $|\gamma_n| = 1$.

We will refer to the numbers $\gamma_1, \ldots, \gamma_n$, $\sigma_1, \ldots, \sigma_{n-1}$ collectively as [illegible]. Since these determine U completely, we see that we can store U, in the form of Schur parameters, in $O(n)$ storage space instead of the usual $O(n^2)$ for $n \times n$ matrices. This being the case, one might reasonably hope to compute the eigenvalues of U in $O(n^2)$ work instead of the usual $O(n^3)$. It turns out that this can be done, and a number of interesting methods have been proposed. The first was an ingenious method of Rutishauser [12], which is, however, unstable and can break down. It relies on LU decompositions that sometimes do not exist. Gragg [9] showed how to do an iteration of the shifted QR algorithm [8] in terms of Schur parameters. Supposing the QR iteration starts with a matrix U and ends with a matrix $\hat{U}$, Gragg derived formulas for computing the Schur parameters of $\hat{U}$ directly from those of U in $O(n)$ arithmetic. Making the reasonable practical assumption that all of the eigenvalues can be found in $O(n)$ QR iterations, we see that we can get the eigenvalues in $O(n^2)$ work. The formulas given in [9] turned out to be unstable, but they can be stabilized [13]. Other methods that have been proposed are described below.

This paper presents another scheme for performing a unitary QR iteration in $O(n)$ work. Our method has several virtues. For one, it can do multishift QR iterations of

*Received by the editors May 5, 2005; accepted for publication (in revised form) by D. Boley March 17, 2006; published electronically August 25, 2006.
http://www.siam.org/journals/simax/28-3/63078.html
†Department of Mathematics, Washington State University, Pullman, WA 99164-3113 (rdavid@math.wsu.edu, watkins@math.wsu.edu).

arbitrary degree. Furthermore, it is straightforward and easy to understand. Finally, it is backward stable. Numerical experiments confirm that the method works well.

2. Previous work. We have already mentioned the work of Rutishauser [12] and Gragg [9]. Bunse-Gerstner and He [6] proposed a bisection method based on a Sturm sequence. Gragg and Reichel [10] and Ammar, Reichel, and Sorensen [2, 3] developed divide-and-conquer algorithms, which were improved by Gu et al. [11].

Several methods make use of the [illegible] form. H is unitarily similar to

$$\tilde{H} = H_o H_e,$$

where H_o (resp., H_e) is the product of the G_i with odd (resp., even) subscripts. Thus, when n is even, for example,

$$H_o = \operatorname{diag}\{\tilde{G}_1, \tilde{G}_3, \ldots, \tilde{G}_{n-1}\} \quad \text{and} \quad H_e = \operatorname{diag}\{1, \tilde{G}_2, \tilde{G}_4, \ldots, \tilde{G}_n\}.$$

Ammar, Gragg, and Reichel [1] have used this form to develop an algorithm for real orthogonal matrices that reduces the problem to two half-sized bidiagonal singular value decompositions. The eigenvalue problem for $\tilde{H} = H_o H_e$ can also be formulated as a generalized eigenvalue problem for the [illegible]

$$H_0 - \lambda H_e^{-1}.$$

Bunse-Gerstner and Elsner [5] formulated variants of the QZ algorithm (single and double shift) for the odd-even pencil.

3. Multishift QR algorithm. Our method is an efficient implementation of the multishift QR algorithm [4, 17]. We will begin, therefore, with a brief review of how the multishift QR algorithm is implemented implicitly. Given a matrix $A \in \mathbb{C}^{n\times n}$ in unreduced upper Hessenberg form and shifts $\mu_i \in \mathbb{C}$ for $i = 1, 2, \ldots, m$, a multishift QR iteration of degree m carries out the steps

$$\begin{aligned} (A - \mu_i I) &= \check{Q}_i \check{R}_i, \\ \check{A}_i &:= \check{R}_i \check{Q}_i + \mu_i I \end{aligned}$$

for $i = 1, 2, \ldots, m$ implicitly. The final matrix $\hat{A} := \check{A}_m$ is produced directly from A and is unitarily similar to A by

$$\hat{A} = Q^* A Q, \tag{3.1}$$

where $Q = \check{Q}_1 \check{Q}_2 \cdots \check{Q}_m$. It can be shown [4, 17] that Q is also the unitary factor in the unitary-upper triangular decomposition

$$(A - \mu_m I)(A - \mu_{m-1} I) \cdots (A - \mu_1 I) = QR.$$

The transformation (3.1) from A to $\hat{A}$ is carried out implicitly as follows:

1. Construct a unitary matrix $V \in \mathbb{C}^{n\times n}$ that satisfies

$$V\mathbf{e}_1 = \frac{1}{\alpha}(A - \mu_m I)(A - \mu_{m-1} I) \cdots (A - \mu_1 I)\mathbf{e}_1,$$

   where $\alpha = \| (A - \mu_m I)(A - \mu_{m-1} I) \cdots (A - \mu_1 I)\mathbf{e}_1 \|_2$.
2. Reduce the matrix $V^* A V$ to upper Hessenberg form.

Since A is upper Hessenberg, the unitary matrix V has the block diagonal form $V = \text{diag}\{\tilde{V}_1, I_{n-m-1}\}$, where $\tilde{V}_1 \in \mathbb{C}^{(m+1)\times(m+1)}$ is unitary. In fact, the matrix V^* maps the vector

$$\mathbf{v} = (A - \mu_m I)(A - \mu_{m-1} I) \dots (A - \mu_1 I)\mathbf{e}_1$$

to $\mathbf{y} = (\alpha, 0, \dots, 0)^T \in \mathbb{C}^n$.

Because of the form of V, the transformation $A \mapsto V^*A$ acts only on the first $(m+1)$ rows, and the transformation $V^*A \mapsto (V^*A)V$ acts only on the first $(m+1)$ columns. Hence the unitary similarity transformation $A \mapsto \tilde{A} := V^*AV$ introduces an [illegible] of size $(m+1) \times (m+1)$ given by the submatrix

$$\begin{bmatrix} \tilde{a}_{2,1} & \cdots & \tilde{a}_{2,m+1} \\ \vdots & & \vdots \\ \tilde{a}_{m+2,1} & \cdots & \tilde{a}_{m+2,m+1} \end{bmatrix}. \tag{3.2}$$

To return $\tilde{A}$ to upper Hessenberg form, a unitary matrix P_1 is built such that the transformation $\tilde{A} \mapsto P_1^*\tilde{A}$ acts only on rows $2, \dots, m+2$ of $\tilde{A}$ to zero out the entries $\tilde{a}_{3,1}, \dots, \tilde{a}_{m+2,1}$. Matrix P_1 has the block diagonal form $\text{diag}\{I_1, \tilde{P}_1, I_{n-m-2}\}$. The transformation $P_1^*\tilde{A} \mapsto (P_1^*\tilde{A})P_1$ acts only on the columns $2, \dots, m+2$, leaving the newly created zeros unaffected, and creates a new row to the bulge. Hence the unitary similarity transformation $\tilde{A} \mapsto P_1^*\tilde{A}P_1$ returns the first column to upper Hessenberg form and moves the bulge one row and one column down. A second unitary matrix P_2 is built so that the transformation $P_1^*\tilde{A}P_1 \mapsto P_2^*(P_1^*\tilde{A}P_1)P_2$ returns the second column to Hessenberg form and moves the bulge one row and one column down. The process is repeated until the bulge is chased off the bottom of the matrix and $\tilde{A}$ is eventually returned to upper Hessenberg form. In all, unitary matrices $P_1, P_2, \dots, P_{n-2}$ are created to carry out this reduction. The kth unitary matrix has the block diagonal form

$$P_k = \begin{bmatrix} I_k & & \\ & \tilde{P}_k & \\ & & I_{n-m-k-1} \end{bmatrix} \tag{3.3}$$

for $k = 1, 2, \dots, n-m-2$ and

$$P_k = \begin{bmatrix} I_k & \\ & \tilde{P}_k \end{bmatrix} \tag{3.4}$$

for $k = n-m-1, \dots, n-2$. It can be shown [4, 17] that the upper Hessenberg matrix that is obtained at the end of this reduction of $\tilde{A}$ is the matrix $\hat{A}$ in (3.1). Hence matrices A and $\hat{A}$ are related by

$$\hat{A} = (P_{n-2}^* \cdots P_2^* P_1^* V^*) A (V P_1 P_2 \cdots P_{n-2}).$$

If A is unitary Hessenberg, we make the following modification in the scheme: The matrices $P_1, P_2, \dots, P_{n-2}$ are constructed such that we get real, positive subdiagonal entries in $\hat{A}$. We require this so that $\hat{A}$ has a factorization of the form (1.1). Matrix P_1, for instance, can be chosen as the unitary matrix that maps the first column of $\tilde{A}$ to the vector $(\tilde{a}_{11}, \xi, 0, 0, \dots, 0)^T \in \mathbb{C}^n$, where $\xi = \| (\tilde{a}_{21}, \dots, \tilde{a}_{m+2,1}) \|_2$.

4. Efficient unitary multishift QR iteration. We now show how to implement a multishift QR iteration efficiently on a unitary matrix in factored form. Let $U \in \mathbb{C}^{n\times n}$ be a unitary matrix in upper Hessenberg form with $u_{j+1,j} > 0$ for $j = 1, \ldots, n-1$. Then U has a factorization

$$U = G_1 G_2 \cdots G_{n-1} G_n, \tag{4.1}$$

where

$$G_k = \begin{bmatrix} I_{k-1} & & \\ & \tilde{G}_k & \\ & & I_{n-k-1} \end{bmatrix} \tag{4.2}$$

with

$$\tilde{G}_k = \begin{bmatrix} \gamma_k & \sigma_k \\ \sigma_k & -\overline{\gamma}_k \end{bmatrix}, \quad \sigma_k > 0, \quad |\gamma_k|^2 + \sigma_k^2 = 1,$$

for $k = 1, \ldots, n-1$, and

$$G_n = \begin{bmatrix} I_{n-1} & \\ & \gamma_n \end{bmatrix} \tag{4.3}$$

with $|\gamma_n| = 1$. The QR iteration will produce a new unitary matrix $\hat{U}$ in factored form:

$$\hat{U} = \hat{G}_1 \hat{G}_2 \cdots \hat{G}_{n-1} \hat{G}_n. \tag{4.4}$$

We define two vectors $\mathbf{g} = (\gamma_1, \ldots, \gamma_n) \in \mathbb{C}^n$ and $\mathbf{s} = (\sigma_1, \ldots, \sigma_{n-1}) \in \mathbb{R}^{n-1}$ to store matrix U. Let $\mu_i \in \mathbb{C}$ for $i = 1, \ldots, m$ be the shifts. The first part of the implicit algorithm is as follows. We construct a matrix $V = \operatorname{diag}\{\tilde{V}_1, I_{n-m-1}\}$ as described in the preceding section. If $\tilde{U} := V^* U V$, then $\tilde{U}$ contains the initial bulge given by the submatrix $\tilde{U}(2 : m+2, 1 : m+1)$. The second part of the algorithm is to return $\tilde{U}$ to upper Hessenberg form $\hat{U}$ by chasing this bulge. The idea behind our implementation is to multiply together the first few of the G_i factors to build a leading submatrix of U that is big enough to accommodate the bulge. We then build the bulge and begin to chase it downward. As we do so, we must multiply in additional G_i factors to accommodate the progressing bulge. However, we also get to factor out matrices $\hat{G}_1$, $\hat{G}_2, \ldots$ from the top since, as soon as the bulge begins to move downward, we can begin to refactor the top part of the matrix, for which the iteration is complete. At any given point in the algorithm, the part of the matrix that contains the bulge can be stored in a work area of dimension $(m+2) \times (m+2)$. On each forward step we must factor in one new G_i at the bottom of the work area, and we get to factor out a $\hat{G}_j$ at the top. The total storage space needed by our algorithm is thus $O(n + m^2)$.

Let

$$W_1 = \begin{bmatrix} \tilde{G}_1 & \\ & I_m \end{bmatrix} \begin{bmatrix} I_1 & & \\ & \tilde{G}_2 & \\ & & I_{m-1} \end{bmatrix} \cdots \begin{bmatrix} I_m & \\ & \tilde{G}_{m+1} \end{bmatrix}.$$

Thus W_1 is the $(m+2) \times (m+2)$ leading principal submatrix of $G_1 G_2 \cdots G_{m+1}$. This goes into the work area initially. Note that the submatrix $W_1(:, 1 : m+1)$, consisting of the first $(m+1)$ columns of W_1, is the submatrix $U(1 : m+2, 1 : m+1)$ of U.

It follows that the submatrix $\tilde{U}(1:m+2, 1:m+1)$, which contains the initial bulge, is the first $m+1$ columns of $W_2 := V_1^* W_1 V_1$, where $V_1 = \mathrm{diag}\{\tilde{V}_1, I_1\}$. In computing W_2, we have thus performed the transformation $U \mapsto V^*UV$ by working only with matrix W_1.

We now chase the bulge. The matrix $P_1 = \mathrm{diag}\{I_1, \tilde{P}_1, I_{n-m-2}\}$ is constructed such that the transformation $\tilde{U} \mapsto P_1^* \tilde{U}$ returns the first column of $\tilde{U}$ to upper Hessenberg form. In terms of the working matrix, we perform the transformation $W_2 \mapsto W_2^{(1)} := \tilde{P}_1^{(1)*} W_2$, where $\tilde{P}^{(1)} = \mathrm{diag}\{I_1, \tilde{P}_1\}$. Further, $\tilde{P}_1$ is constructed so that the entry $W_2^{(1)}(2,1) > 0$. Hence the first column of $W_2^{(1)}$ is $(\hat{\gamma}_1, \hat{\sigma}_1, 0, \ldots, 0)^T$. We can then perform the factorization

$$W_2^{(1)} = \begin{bmatrix} \tilde{G}_1^{(1)} & \\ & I_m \end{bmatrix} \begin{bmatrix} I_1 & \\ & \tilde{W}_2^{(1)} \end{bmatrix}, \tag{4.5}$$

where

$$\tilde{G}_1^{(1)} = \begin{bmatrix} \hat{\gamma}_1 & \hat{\sigma}_1 \\ \hat{\sigma}_1 & -\bar{\hat{\gamma}}_1 \end{bmatrix}.$$

The matrix $\hat{G}_1 = \mathrm{diag}\{\tilde{G}_1^{(1)}, I_{m-2}\}$ is the first matrix in the factorization (4.4). The first entries of the vectors $\mathbf{g}$ and $\mathbf{s}$ are replaced with the new Schur parameters $\hat{\gamma}_1$ and $\hat{\sigma}_1$. From (4.5), we see that $\tilde{W}_2^{(1)}$ is the trailing $(m+1)\times(m+1)$ principal submatrix of $\mathrm{diag}\{\tilde{G}_1^{(1)}, I_m\}^* W_2^{(1)}$. We extract $\tilde{W}_2^{(1)}$ and let

$$W_2^{(2)} := \begin{bmatrix} \tilde{W}_2^{(1)} & \\ & I_1 \end{bmatrix}. \tag{4.6}$$

This is our new working matrix. The next factor in (4.1) is multiplied in:

$$W_2^{(3)} := W_2^{(2)} \begin{bmatrix} I_m & \\ & \tilde{G}_{m+2} \end{bmatrix}.$$

Finally, to carry out the transformation $P_1^* \tilde{U} \mapsto \tilde{U}_1 := (P_1^* \tilde{U}) P_1$ we note from (3.3) that P_1 commutes with $G_{m+3}, \ldots, G_n$. Thus if

$$W_3 := W_2^{(3)} \begin{bmatrix} \tilde{P}_1 & \\ & I_1 \end{bmatrix},$$

then the first $(m+1)$ columns of W_3 form the submatrix $\tilde{U}_1(3:m+3, 2:m+2)$, which contains the new bulge. This completes the transformation $\tilde{U} \mapsto P_1^* \tilde{U} P_1$.

In general, for $k = 2, \ldots, n-m-2$, we have the working matrix W_{k+1}, whose first $(m+1)$ columns contain the bulge. The matrix P_k having the form (3.3) is built. In this block diagonal form, the unitary matrix $\tilde{P}_k$ is constructed such that the transformation

$$W_{k+1} \mapsto W_{k+1}^{(1)} := \tilde{P}_k^{(1)*} W_{k+1}, \tag{4.7}$$

where $\tilde{P}_k^{(1)} = \mathrm{diag}\{I_1, \tilde{P}_k\}$, returns the first column of W_{k+1} to upper Hessenberg form and makes the entry $W_{k+1}^{(1)}(2,1) > 0$. Next, the factorization

$$W_{k+1}^{(1)} = \begin{bmatrix} \tilde{G}_k^{(1)} & \\ & I_m \end{bmatrix} \begin{bmatrix} I_1 & \\ & \tilde{W}_{k+1}^{(1)} \end{bmatrix} \tag{4.8}$$

is performed, where

$$\tilde{G}_k^{(1)} = \begin{bmatrix} \hat{\gamma}_k & \hat{\sigma}_k \\ \hat{\sigma}_k & -\bar{\hat{\gamma}}_k \end{bmatrix}.$$

The kth entries in the vectors $\mathbf{g}$ and $\mathbf{s}$ are updated with $\hat{\gamma}_k$ and $\hat{\sigma}_k$, respectively. The submatrix $\tilde{W}_{k+1}^{(1)}$ is extracted, and the working matrix

$$W_{k+1}^{(2)} := \begin{bmatrix} \tilde{W}_{k+1}^{(1)} & \\ & I_1 \end{bmatrix} \tag{4.9}$$

is formed. The next factor in (4.1) is multiplied in,

$$W_{k+1}^{(3)} := W_{k+1}^{(2)} \begin{bmatrix} I_m & \\ & \tilde{G}_{m+k+1} \end{bmatrix},$$

and a full working matrix is formed by

$$W_{k+2} := W_{k+1}^{(3)} \begin{bmatrix} \tilde{P}_k & \\ & I_1 \end{bmatrix}. \tag{4.10}$$

When $k = n - m - 1$, the working matrix begins to shrink. After the operations (4.7) and (4.8), there is no need to make the extension indicated by (4.9), because $\tilde{G}_n = [\gamma_n]$ is only 1×1, not 2×2. On subsequent steps the working matrix continues to shrink, because there are no more factors to multiply in. By the time the bulge chase is complete, the working matrix has been reduced to 2×2 and can be factored to form

$$\begin{bmatrix} \hat{\gamma}_{n-1} & \hat{\sigma}_{n-1} \\ \hat{\sigma}_{n-1} & -\bar{\hat{\gamma}}_{n-1} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \hat{\gamma}_n \end{bmatrix}.$$

The new Schur parameters $\hat{\gamma}_{n-1}$, $\hat{\sigma}_{n-1}$, and $\hat{\gamma}_n$ replace the old ones in $\mathbf{g}$ and $\mathbf{s}$, and the iteration is complete.

Enforcement of unitarity. One other important detail needs to be mentioned. Each new pair of Schur parameters $\hat{\gamma}_k$, $\hat{\sigma}_k$ satisfies $|\hat{\gamma}_k|^2 + \hat{\sigma}_k^2 = 1$ in principle, but in practice roundoff errors will cause this equation to be violated by a tiny amount. Therefore the following normalization step is required:

$$\begin{aligned} &\nu \leftarrow \left(|\hat{\gamma}_k|^2 + \hat{\sigma}_k^2 \right)^{1/2}, \\ &\hat{\gamma}_k \leftarrow \hat{\gamma}_k / \nu, \\ &\hat{\sigma}_k \leftarrow \hat{\sigma}_k / \nu. \end{aligned} \tag{4.11}$$

This should be done even when $k = n$, taking $\hat{\sigma}_n = 0$. This enforcement of unitarity is essential to the stability of the algorithm. If it is not done, the matrix will (over the course of many iterations) drift away from being unitary, and the algorithm will fail.

Backward stability. If we could perform a QR iteration in exact arithmetic, we would have $\hat{U} = Q^* U Q$, where Q, U, and $\hat{U}$ are exactly unitary matrices. Now suppose we perform the iteration in floating-point arithmetic, at first supposing that U and $\hat{U}$ are fully assembled, i.e., not in the factored form $U = G_1 \cdots G_n$. Then it is

well established that the computed $\hat{U}$ satisfies $\hat{U} = Q^*(U + E)Q$, where Q is exactly unitary and $\| E \|_2$ is a modest multiple of the unit roundoff of the floating-point arithmetic [19, Chapter 3]. Thus the iteration is backward stable.

The additional complication that arises in our algorithm is that the matrix U is presented in factored form, and $\hat{U}$ is produced in factored form. In the course of the iteration, factors are multiplied together at some points and split apart at others. Errors occur during each of these operations, and we need to analyze their effect.

There is no problem with the phase in which factors are multiplied together. The multiplication of two unitary (or near unitary) matrices results in a product that has a tiny backward error.

The big question is what happens in the splitting-apart step shown in (4.5) and (4.8). The factorization (4.8) is effected by multiplying $W_{k+1}^{(1)}$ on the left by the conjugate transpose of $\mathrm{diag}\{\tilde{G}_k^{(1)}, I_m\}$ to obtain

$$\begin{bmatrix} I_1 & \\ & \tilde{W}_{k+1}^{(1)} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \tilde{W}_{k+1}^{(1)} & \\ 0 & & & \end{bmatrix}.$$

The generation of zeros in the first row of this matrix depends upon the fact that it is unitary and therefore has orthonormal columns. In fact, the matrix is not quite unitary, so those first-row entries will not be exactly zero in practice. We need to show that they are tiny enough that setting them to zero does not compromise the stability of the algorithm.

The success of the analysis hinges on the fact that we perform the normalization step (4.11) each time we produce a new pair of Schur parameters. This ensures that no matter how many steps we have taken in the algorithm, each of the G_i matrices is nearly unitary, having the form $G_i = \tilde{G}_i + E_i$, where $\tilde{G}_i$ is exactly unitary and $\| E_i \|_2$ is on the order of the unit roundoff u.

Now suppose we multiply together two matrices that are nearly unitary. Say we have $A = \tilde{A} + E$ and $B = \tilde{B} + F$, where $\tilde{A}$ and $\tilde{B}$ are unitary and $\| E \|_2$ and $\| F \|_2$ are on the order of the unit roundoff u. If we multiply them together, we obtain a computed product P that satisfies

$$P = AB + H,$$

where $\| H \|_2$ is on the order of u. Substituting in the forms of A and B, and letting $\tilde{P} = \tilde{A}\tilde{B}$, we find that

$$P = \tilde{P} + K,$$

where $K = E\tilde{B} + \tilde{A}F + EF + H$. Thus P is the sum of a unitary matrix and an error matrix K such that $\| K \|_2$ is on the order of u.

The first working array that gets split apart is $W_2^{(1)}$, which gets factored as shown in (4.5). $W_2^{(1)}$ was formed by multiplying together several matrices that are almost exactly unitary. Therefore, by applying the analysis of the previous paragraph repeatedly, we see that $W_2^{(1)} = \tilde{W} + E$, where $\tilde{W}$ is exactly unitary and $\| E \|_2$ is at most a modest multiple of u. The first column of $W_2^{(1)}$ has the form $(\check{\gamma}_1, \check{\sigma}_1, 0, \ldots, 0)^T$, where $\check{\gamma}_1^2 + \check{\sigma}_1^2 = 1 + \epsilon_1$, with ϵ_1 on the order of u. Let $G = \mathrm{diag}\{\tilde{G}_1^{(1)}, I_m\}$, the matrix

in (4.5) that has the property that G^* zeros out the entry $\check{\sigma}_1$ in the first column of $W_2^{(1)}$. The entries $\hat{\gamma}_1$ and $\hat{\sigma}_1$, which are used to build G, are obtained from $\check{\gamma}_1$ and $\check{\sigma}_1$ by carrying out the normalization step (4.11). Since roundoff errors are incurred in the normalization step, G^* transforms $(\check{\gamma}_1, \check{\sigma}_1, 0, \ldots, 0)^T$ to $(1+\epsilon_2, \epsilon_3, 0, \ldots, 0)^T$; that is, it does not exactly succeed in transforming $\check{\sigma}_1$ to zero. Let $\tilde{G}$ denote the theoretical G matrix built using the entries from the first column of $\tilde{W}$ rather than $W_2^{(1)}$. This matrix is exactly unitary, and the first column of $\tilde{G}^*\tilde{W}$ is exactly $e_1 = (1, 0, 0, \ldots, 0)^T$. Moreover, since $W_2^{(1)}$ differs only slightly from $\tilde{W}$ and the computation (4.11) has high relative accuracy in floating-point arithmetic, $G = \tilde{G} + F$, where $\| F \|_2$ is on the order of u.

Now consider the computation $G^*W_2^{(1)}$, which forms the factor $\mathrm{diag}\{I_1, \tilde{W}_2^{(1)}\}$ in (4.5). We have

$$G^*W_2^{(1)} = \tilde{G}^*\tilde{W} + H, \tag{4.12}$$

where $H = F^*\tilde{W} + \tilde{G}^*E + F^*E$. The matrix $\tilde{G}^*\tilde{W}$ is exactly unitary, its first column is exactly e_1, and its first row is exactly e_1^T. Since $\| H \|_2$ is at most a modest multiple of u, we conclude from (4.12) that the first row and column of the computed matrix $G^*W_2^{(1)}$ differ from e_1^T and e_1, respectively, by errors on the order of u. Therefore, the error we make in setting the first row and column to e_1^T and e_1, respectively, is tiny. Their contribution to the backward error is equally tiny.

We also deduce from (4.12) that the matrix $\tilde{W}_2^{(1)}$ that is created in (4.5) differs only negligibly from a unitary matrix. This is the part of the matrix that is carried forward to the working array (4.6). Now, proceeding inductively, we can conclude that at every step of the algorithm the matrix in the working array differs only negligibly from a unitary matrix. At each factorization (4.8) a tiny backward error is incurred, and the part of the working array that is moved forward differs only negligibly from a unitary matrix. Therefore the algorithm is backward stable. This argument depends critically on the normalization (4.11), which guarantees that each new factor that is brought into the working array is almost exactly unitary.

Operation count. The bulk of the arithmetic in our algorithm is contained in the steps (4.7) and (4.10). Each unitary transformation is taken to be the product of a reflector followed by a diagonal phase-correcting transformation to enforce the condition $\hat{u}_{k+1,k} > 0$. The latter costs $O(m)$ arithmetic; the real work is in applying the reflector. Each of these is at most $(m+1) \times (m+1)$ (smaller at the very end of the iteration), and the cost of applying it efficiently to the working matrix on left or right is about $4m^2$ flops [16, section 3.2]. Since the reflector is applied only to the small work area and not to the full Hessenberg matrix, the amount of arithmetic is $O(m^2)$ instead of $O(nm)$; this is where we realize our savings. Since $n-1$ reflectors are applied (on left and right) in the whole iteration, the arithmetic cost is about $8nm^2$ flops.

If m is fixed and small, then we can say that the cost of an iteration is $O(n)$, in the sense that the arithmetic is bounded by $C_m n$, where C_m is independent of n. However, the fact that C_m grows like m^2 as m is increased shows that it will be inefficient to take m too large.

There is another important reason for keeping m fairly small. If m is made much bigger than 8 or 10, roundoff errors interfere with the mechanism of shift transmission and render the QR iteration ineffective [15]. This phenomenon is known as

5. Shift strategies. Eberlein and Huang [7] presented a globally convergent shift strategy for the unitary QR algorithm, and they showed that it converges at least quadratically. Wang and Gragg [14] proposed a family of strategies that includes that of Eberlein and Huang. They demonstrated global convergence and showed that the convergence rate is always at least cubic. These strategies are for single QR iterations, the case $m = 1$.

Since we are taking multiple steps, we need a different strategy. The most common way to obtain m shifts is to take the eigenvalues of the trailing $m \times m$ submatrix of U. Watkins and Elsner [18] showed that this strategy is cubically convergent when it converges. However, it is not globally convergent, as the following well-known example shows. Let U be the unitary circulant shift matrix, which looks like

$$\begin{bmatrix} & & & 1 \\ 1 & & & \\ & 1 & & \\ & & 1 & \end{bmatrix}$$

in the 4×4 case. For any $m < n$, if we take the eigenvalues of the trailing submatrix as shifts, we get shifts $0, \ldots, 0$, which are equidistant from all of the eigenvalues. A QR iteration on U with these shifts goes nowhere.

Since the eigenvalues of a unitary matrix lie on the unit circle, it make sense to choose shifts that are on the unit circle. We tried two strategies. The first computes the eigenvalues of the trailing $m \times m$ submatrix and normalizes each of them by dividing it by its absolute value. If any of the tentative shifts happens to be zero, it is replaced by a random number on the unit circle. If we use this strategy on the circulant shift matrix, we get m random shifts.

A second strategy stems from the following observation. The last m rows of the unreduced Hessenberg matrix U are orthonormal. Since $u_{n-m+1,n-m} > 0$, the trailing $m \times m$ submatrix $U(n-m+1:n, n-m+1:n)$ is not unitary, but it is nearly unitary. Its rows are orthogonal, and they all have norm 1, except that the top row $U(n-m+1, n-m+1:n)$ has norm less than 1. Unitarity can be restored by dividing this row by its norm. In the rare case when the whole top row is zero, a suitable first row can be generated by orthonormalizing a random row against rows 2 through m. The m eigenvalues of the modified matrix then give m shifts on the unit circle.

When this strategy is used on the circulant shift matrix, the orthonormalization process will generate a first row of the form $(0, \ldots, 0, \gamma)$ with $|\gamma| = 1$. The shifts are then the roots of the equation $z^m - \gamma = 0$, which are equally spaced points on the unit circle.

We found that these two strategies work about equally well. Both are locally cubically convergent: As $u_{n-m+1,n-m} \to 0$, the trailing $m \times m$ submatrix becomes closer and closer to unitary. Its eigenvalues become ever closer to the unit circle, and normalizing them as in the first strategy moves them only slightly. On the other hand, if we modify the matrix as in the second strategy by normalizing its first row, that also moves the eigenvalues only slightly, because the rescaling factor is very close to 1. Thus both strategies behave asymptotically the same as the strategy that simply takes the eigenvalues of the trailing submatrix as shifts; that is, they converge cubically [18] when they converge.

We conjecture that both strategies converge globally.

6. Numerical results. To verify that our algorithm works as expected, we coded it in MATLAB and tried it out on numerous unitary matrices. Test problems

TABLE 6.1
Error in computed eigenvalues of a 1000×1000 *unitary matrix.*

2	Maximum error
$m = 1$	1.10×10^{-13}
$m = 2$	5.31×10^{-14}
$m = 3$	3.25×10^{-14}
$m = 4$	2.77×10^{-14}
$m = 5$	2.71×10^{-14}
$m = 6$	2.53×10^{-14}
$m = 7$	2.52×10^{-14}
$m = 8$	2.14×10^{-14}
MATLAB	1.03×10^{-14}

with known eigenvalues were generated as follows. A unitary diagonal matrix D was generated and its eigenvalues noted. A unitary matrix Q, random with respect to the Haar measure, was generated, and the random unitary matrix $B = QDQ^*$ formed. Then B was transformed to upper Hessenberg form to yield an upper Hessenberg unitary matrix A with known eigenvalues, which was then factored into the form (1.1).

The eigenvalues of unitary matrices are perfectly conditioned, so we always expect to be able to compute them to very high accuracy. We found that our algorithm was able to do this. The results in Table 6.1 are typical. These are for a matrix of order 1000×1000 with eigenvalues randomly distributed on the unit circle. We computed the eigenvalues with our code using $m = 1, 2, 3, \ldots, 8$ and obtained accurate results in all cases. It is interesting that increasing m increases the accuracy. At $m = 4$ the maximum error is only about one fourth what it is for $m = 1$. We already have at least two reasons for not taking m too large, but these numbers suggest that $m = 1$ may not be the best choice.

For real orthogonal matrices one should always take $m \geq 2$, and the complex shifts should be taken in conjugate pairs. Then the matrix $(A - \mu_m I) \cdots (A - \mu_1 I)$ is real, and all operations can be done in real arithmetic.

As Table 6.1 shows, we also had the standard MATLAB QR code compute the eigenvalues of the Hessenberg matrix, and we found that it was a bit more accurate than our codes, but the difference was not substantial.

The results in Table 6.1 are for a single matrix, but they are entirely typical of what we observed. The test matrices included matrices with many repeated eigenvalues and others with tight clusters of eigenvalues. The eigenvalues of smaller matrices are computed with slightly more accuracy than are those of large ones, but in all cases the results were qualitatively like those in Table 6.1. We conclude that our algorithm works as expected.

REFERENCES

[1] G. AMMAR, W. GRAGG, AND L. REICHEL, *On the eigenproblem for orthogonal matrices*, in Proceedings of the 25th Annual IEEE Conference on Decision and Control, Athens, Greece, 1986, IEEE Press, Piscataway, NJ, pp. 1963–1966.

[2] G. S. AMMAR, L. REICHEL, AND D. C. SORENSEN, *An implementation of a divide and conquer algorithm for the unitary eigenproblem*, ACM Trans. Math. Software, 18 (1992), pp. 292–307.

[3] G. S. AMMAR, L. REICHEL, AND D. C. SORENSEN, *Corrigendum: Algorithm* 730: *An implementation of a divide and conquer algorithm for the unitary eigenproblem*, ACM Trans. Math. Software, 20 (1994), p. 161.

[4] Z. Bai and J. Demmel, *On a block implementation of the Hessenberg multishift QR iteration*, Internat. J. High Speed Comput., 1 (1989), pp. 97–112.
[5] A. Bunse-Gerstner and L. Elsner, *Schur parameter pencils for the solution of the unitary eigenproblem*, Linear Algebra Appl., 154/156 (1991), pp. 741–778.
[6] A. Bunse-Gerstner and C. He, *On a Sturm sequence of polynomials for unitary Hessenberg matrices*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1043–1055.
[7] P. J. Eberlein and C. P. Huang, *Global convergence of the QR algorithm for unitary matrices with some results for normal matrices*, SIAM J. Numer. Anal., 12 (1975), pp. 97–104.
[8] J. G. F. Francis, *The QR transformation, Parts* I *and* II, Computer J., 4 (1961), pp. 265–272, 332–345.
[9] W. B. Gragg, *The QR algorithm for unitary Hessenberg*, J. Comput. Appl. Math., 16 (1986), pp. 1–8.
[10] W. B. Gragg and L. Reichel, *A divide and conquer algorithm for the unitary and orthogonal eigenproblems*, Numer. Math., 57 (1990), pp. 695–718.
[11] M. Gu, R. Guzzo, X.-B. Chi, and X.-Q. Cao, *A stable divide and conquer algorithm for the unitary eigenproblem*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 385–404.
[12] H. Rutishauser, *Bestimmung der Eigenwerte orthogonaler Matrizen*, Numer. Math., 9 (1966), pp. 104–108.
[13] M. Stewart, *An Error Analysis of a Unitary Hessenberg QR Algorithm*, Technical Report TR-CS-98-11, Department of Computer Science, Australian National University, Canberra, Australia, 1998; available online at http://eprints.anu.edu.au/archive/00001557/.
[14] T.-L. Wang and W. B. Gragg, *Convergence of the shifted QR algorithm for unitary Hessenberg matrices*, Math. Comp., 71 (2002), pp. 1473–1496.
[15] D. S. Watkins, *The transmission of shifts and shift blurring in the QR algorithm*, Linear Algebra Appl., 241/243 (1996), pp. 877–896.
[16] D. S. Watkins, *Fundamentals of Matrix Computations*, 2nd ed., John Wiley and Sons, New York, 2002.
[17] D. S. Watkins and L. Elsner, *Chasing algorithms for the eigenvalue problem*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 374–384.
[18] D. S. Watkins and L. Elsner, *Convergence of algorithms of decomposition type for the eigenvalue problem*, Linear Algebra Appl., 143 (1991), pp. 19–47.
[19] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, UK, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 3, pp. 634–641

ON THE CONVERGENCE OF ITERATIVE METHODS FOR SEMIDEFINITE LINEAR SYSTEMS*

YOUNG-JU LEE†, JINBIAO WU‡, JINCHAO XU§, AND LUDMIL ZIKATANOV§

Abstract. Necessary and sufficient conditions for the energy norm convergence of the classical iterative methods for semidefinite linear systems are obtained in this paper. These new conditions generalize the classic notion of the P-regularity introduced by Keller [*J. Soc. Indust. Appl. Math. Ser. B Numer. Anal.*, 2 (1965), pp. 281–290].

Key words. iterative methods, matrix splittings, P-regularity, weak regularity, energy norm convergence

AMS subject classifications. 65F10, 65N20, 65N30

DOI. 10.1137/050644197

1. Introduction.

Semidefinite linear systems arise in many applications, as a result of discretization of semidefinite partial differential equations or as an important part of a complex numerical model. Examples include the equations coming from discretizing the Poisson equation with Neumann boundary conditions (see Bochev and Lehoucq [3]), linear elasticity equations with traction free boundary conditions, systems arising in Markov processes [7], and also graph partitioning applications [8, 9]. Some more subtle examples are provided by the linear systems obtained from the generalized finite element method discretizations of partial differential equations (see [23, 21, 22] and [26]). For such problems, using iterative solvers is unavoidable, because direct methods such as Gaussian elimination or the Cholesky decomposition are hard to apply in a straightforward way (since the problem is singular).

We consider the problem of finding a solution $x \in \mathbb{R}^n$ to

$$Ax = b, \tag{1.1}$$

where $A \in \mathbb{R}^{n\times n}$ is a given symmetric and semidefinite matrix and $b \in \mathbb{R}^n$ is a given vector in the range of A. A stationary linear iterative method to solve (1.1) can be obtained using a splitting of the matrix $A = M - N$,

$$Mx^{\ell} = Nx^{\ell-1} + b. \tag{1.2}$$

Convergence properties of (1.2) for semidefinite problems have been studied by many authors. Classic as well as more recent results on this topic can be found in [1, 15, 4, 6, 19] (and also many references listed therein). We note that all of these convergence results require that the matrix M be an invertible matrix (see, e.g., [6]).

*Received by the editors November 3, 2005; accepted for publication (in revised form) by A. J. Wathen March 29, 2006; published electronically August 29, 2006. This work was supported in part by NSF grant DMS-0209497 and the Center for Computational Mathematics and Applications, Penn State University.

http://www.siam.org/journals/simax/28-3/64419.html

†Department of Mathematics, UCLA, 520 Portola Plaza, Los Angeles, CA 90095 (yjlee@math.ucla.edu).

‡Laboratory of Mathematics and Applied Mathematics, School of Mathematical Sciences, Peking University, Beijing 100871, China (jwu@math.pku.edu.cn).

§Department of Mathematics, The Pennsylvania State University, University Park, PA 16802 (xu@math.psu.edu, ludmil@psu.edu). The work of the fourth author was supported in part by Lawrence Livermore National Laboratory under contract B551021.

It is easy to see, however, that the iterations (1.2) are well defined even for a singular iterator M as long as the right-hand side falls in the range of M. Such a situation occurs, for example, in multigrid methods (when the coarsest grid problem is singular; see [13]) and balancing domain decomposition methods (nonoverlapping or overlapping; see [14, 12]). In this paper we will study the convergence of (1.2) without assuming that M is invertible. Our main result is given in Theorem 4.4 below. The conditions under which this convergence result holds have been applied to studying more general subspace correction algorithms for variational problems (multigrid and domain decomposition methods being classical examples for such methods). For nonsingular symmetric positive definite problems we refer to [25] and [27] for general theory of subspace correction methods, and for the relations between the conditions given here and the convergence of subspace correction methods for variational semidefinite problems in Hilbert spaces we refer to our recent work [13]. In any event, deriving necessary and sufficient conditions for energy norm convergence without assuming that M is invertible is of theoretical interest in its own right. We further point out that our result here is also new even when M is nonsingular. Related convergence results that use an algebraic framework for Schwarz methods (multiplicative and additive) can be found in Nepomnyaschikh [18], Chang and Sun [5], Marek and Szyld [16], and Nabben and Szyld [17].

The structure of the paper is as follows. In section 2, we give definitions and relevant notation related to matrix splittings. In section 3, we introduce the P-regularity and weak regularity of splittings and their relations to the energy norm convergence. We also provide a simple example of a matrix that is not P-regular but results in a convergent iteration (as defined in (2.3) below). In section 4, we give more refined necessary and sufficient conditions and prove energy norm convergence.

2. Notation and preliminaries. We first introduce standard notation. For a finite dimensional space V with an inner product $(\cdot,\cdot)$ and any subspace $W \subset V$, $W^{\perp}$ denotes the orthogonal complement of W with respect to the inner product $(\cdot,\cdot)$, and V/W denotes the quotient space; for a given matrix T, the range of T and the null space of T are denoted by $\mathcal{R}(T)$ and $\mathcal{N}(T)$, respectively. Further, following [25], we write $x_1 \quad y_1$ ($x_2 \quad y_2$) whenever there exist generic constants c_1 and c_2 such that $x_1 \le c_1 y_1$ ($x_2 \ge c_2 y_2$).

We consider again the semidefinite system (1.1). Given an initial guess x^0, we rewrite the iteration (1.2) for the solution of (1.1) as follows:

$$M(x^{\ell} - x^{\ell-1}) = b - Ax^{\ell-1}, \quad \ell = 1, 2, \ldots. \tag{2.1}$$

Clearly, if $\mathcal{N}(A) \neq \{0\}$, the solution of (1.1) is not unique (but unique in the quotient space $\mathbb{R}^n/\mathcal{N}(A)$). Similarly if $\mathcal{N}(M) \neq \{0\}$, the solution of (2.1) is not unique either. A special solution of (2.1) can be given by

$$x^{\ell} = x^{\ell-1} + M^{\dagger}(b - Ax^{\ell-1}), \tag{2.2}$$

where $M^{\dagger} \equiv (M^T M)^{-1} M^T$ is the Moore–Penrose generalized inverse of M. It is obvious that for an invertible M, $M^{\dagger} = M^{-1}$. The Moore–Penrose inverse for the splitting matrix M has been used, for example, by Joshi [10] in an attempt to generalize the result of Keller [11] for rectangular matrices A. However, since M is assumed to be of full rank in [10] when A is a square matrix, $M^{\dagger}$ is reduced to the usual inverse of M.

Probably the best known and quoted results in the theory of the convergence of iterative methods for singular linear systems are contained in work by Keller [11].

For an invertible iterator M, Keller defines a convergent linear iterative method as a method for which the error propagation matrix T satisfies

$$\lim_{k\to\infty} T^k = P_{\mathcal{N}}, \qquad \text{with} \qquad T := I - M^{-1}A = M^{-1}N, \tag{2.3}$$

where $P_{\mathcal{N}}$ is a projection (not necessarily orthogonal) onto the null space $\mathcal{N}(A)$. Other forms of abstract necessary and sufficient conditions for energy norm convergence are given in [1]. More refined analysis and also relations between various types of such conditions are identified and studied in [24]. In many instances, however, it is difficult to verify a relation such as (2.3). A more practical (but only sufficient) condition, known as the P-regularity condition, introduced in the aforementioned work by Keller, has been used as a criterion in many of the previous studies on the convergence of iterative methods [1, 15, 4] and [6].

3. On P-regular and weak regular splittings. We now recall the convergence concept for the iteration (2.2) given by Keller [11, Theorem 1, p. 282] (again we assume that A is semidefinite).

DEFINITION 3.1. [illegible] (2.2) [illegible] $x^0 \in R^n$ [illegible]

$$\lim_{\ell\to\infty} Ax^\ell = b. \tag{3.1}$$

Let x be a solution to (1.1); then it is easy to see that (3.1) is equivalent to

$$\lim_{\ell\to\infty} \|x - x^\ell\|_{\mathbb{R}^n/\mathcal{N}(A)} = 0$$

or

$$\lim_{\ell\to\infty} |x - x^\ell|_A = 0,$$

where $|x|_A = (Ax, x)^{1/2}$; see, e.g., [1, 11, 15] and references cited therein.

An obvious sufficient condition for the convergence (3.1) is

$$|I - M^\dagger A|_A < 1. \tag{3.2}$$

When this condition is satisfied, we will say that (2.2) is [illegible] or [illegible]

One main convergence result for (2.2), derived by Keller [11], can be summarized in the following theorem.

THEOREM 3.2 (see Keller [11]). [illegible] (2.2) [illegible] $A = M - N$ [illegible]

(K1) M [illegible]

(K2) $M^T + M - A$ [illegible]

For some special singular systems, considered by, e.g., Marek and Szyld [16], the convergence has been studied via the theory of nonnegative matrices, for which the [illegible] condition, proposed in Ortega and Rheinboldt [20], is often used as a sufficient condition. A version of the weak-regularity condition (see, e.g., Berman and Plemmons [1]) is as follows: A splitting $A = M - N$ is called weakly regular if M is invertible and, in addition, both M^{-1} and $M^{-1}N$ are nonnegative matrices. We shall now provide an example to show that neither P-regularity nor weak regularity of the matrix splitting is necessary for the convergence of (2.1).

1. Consider the symmetric positive semidefinite matrix A given below:

$$A = \begin{pmatrix} 1/2 & -1 \\ -1 & 2 \end{pmatrix}.$$

We introduce a splitting $A = M - N$, where

$$M = \begin{pmatrix} -1 & -4 \\ 0 & 4 \end{pmatrix} \qquad \text{and} \qquad M^{-1} = \begin{pmatrix} -1 & -1 \\ 0 & 1/4 \end{pmatrix}.$$

Then

$$\overline{M} = M + M^T - A = \begin{pmatrix} -5/2 & -3 \\ -3 & 6 \end{pmatrix}.$$

This splitting obviously is not P-regular, since $\overline{M}$ is apparently not positive definite. Moreover, the splitting above is also not weakly regular, since M^{-1} has two negative elements; that is, M^{-1} is not nonnegative.

However, it is also obvious that $E = I - M^{-1}A$ is convergent. More precisely,

$$E^\ell = E \quad \forall\, \ell \geq 1 \qquad \text{and} \qquad E = P_{\mathcal{N}(A)},$$

where $P_{\mathcal{N}(A)}$ is a projection onto $\mathcal{N}(A)$. Note that the projection $P_{\mathcal{N}(A)}$ is not orthogonal.

We would like to remark that in case $\mathcal{R}(M^\dagger A) \subset \mathcal{R}(A)$, a convergence result can easily be derived in a fashion similar to the positive definite case, since A becomes symmetric and positive definite on $\mathcal{R}(A)$ (see, e.g., [25]). An example for which such a relation holds is the Richardson iteration. However, this is not the case for other classical iterations, such as the Gauss–Seidel or Jacobi iterations and the multigrid and domain decomposition methods.

4. New conditions and analysis. In this section, we present new conditions that are necessary and sufficient for the energy norm convergence of the iteration (2.2). They are given as follows:

(A1) $\mathcal{R}(A) \subset \mathcal{R}(M)$, or equivalently, $\mathcal{N}(M^T) \subset \mathcal{N}(A)$.

(A2) $M^T + M - A$ is symmetric positive definite on $\mathcal{R}(M^\dagger A)$.

Note that when M is nonsingular, (A1) always holds. If the decomposition $A = M - N$ satisfies the assumption (A1) and additionally $\mathcal{N}(M) \subset \mathcal{N}(A)$, then the splitting is called a [illegible] (see, e.g., Berman and Neumann [2]).

The assumption (A1) is obviously necessary for (2.2) to be well-defined for any initial guess x^0, since $b - Ax^{\ell-1} \in \mathcal{R}(A)$. We note that both (A1) and (A2) are clearly weaker than (K1) and (K2) in Theorem 3.2. The identity (4.1), proven below, obviously holds for M being square and nonsingular. We also note that the relation given by (4.1) can also be found as an assumption on the iterator M in Joshi [10] for the study of iterative methods for problems with rectangular A.

LEMMA 4.1. [illegible] (A1) [illegible]

$$MM^\dagger A = A, \tag{4.1}$$

[illegible]

$$\|M^\dagger Ax\| \quad |x|_A \quad \forall x \in \mathbb{R}^n. \tag{4.2}$$

[illegible] $x^{\ell-1}$ [illegible] (2.1) [illegible] $x^\ell \in \mathbb{R}^n/\mathcal{N}(M)$ [illegible] (2.2)

[illegible]. By the definition of the Moore–Penrose inverse, it is obvious that $MM^\dagger M = M$. This then implies

$$MM^\dagger y = y \quad \forall y \in \mathcal{R}(M).$$

Now, from the assumption (A1), we obtain that

$$MM^\dagger y = y \quad \forall y \in \mathcal{R}(A) \subset \mathcal{R}(M).$$

This proves that $MM^\dagger A = A$, and hence (4.1) holds.

To prove (4.2) we observe that

$$\begin{aligned} |x|_A^2 = (Ax, x) &\le \|Ax\| \|x\|_{\mathbb{R}^n/\mathcal{N}(A)} \\ &\|MM^\dagger Ax\| |x|_A \quad \text{by (4.1)} \\ &\|M^\dagger Ax\| |x|_A, \quad \text{since } M \text{ is bounded.} \end{aligned}$$

This proves the inequality (4.2).

To complete the proof of the lemma, we first use the fact that (A1) implies that (2.1) is solvable; i.e., for any $x^{\ell-1}$ there exists x^ℓ such that $M(x^\ell - x^{\ell-1}) = b - Ax^{\ell-1}$. However, since $\mathcal{N}(M)$ is not empty, the x^ℓ is determined uniquely only in the space $\mathbb{R}^n/\mathcal{N}(M)$ for any $x^{\ell-1}$. Furthermore, the iterate x^ℓ that is obtained from (2.2) can be a solution to (2.1) for any given $x^{\ell-1}$. We observe that since $x^\ell - x^{\ell-1} = M^\dagger(b - Ax^{\ell-1})$,

$$M(x^\ell - x^{\ell-1}) = MM^\dagger A(x - x^{\ell-1}) = A(x - x^{\ell-1}), \quad \text{by (4.1).}$$

This completes the proof. ∎

The next lemma is related to assumption (A2) and has a well-known analogue when A is symmetric and positive definite. It implies that the P-regularity is necessary and sufficient for the energy norm convergence when A is nonsingular (see, e.g., Young [28]). For semidefinite A the result is as follows.

LEMMA 4.2. [illegible] (A1) [illegible] $x \in \mathbb{R}^n$

$$|x|_A^2 - |(I - M^\dagger A)x|_A^2 = ((M^T + M - A)M^\dagger Ax, M^\dagger Ax). \tag{4.3}$$

[illegible]. A direct calculation shows that for any $x \in \mathbb{R}^n$ we have

$$\begin{aligned} |x|_A^2 - |(I - M^\dagger A)x|_A^2 &= (Ax, M^\dagger Ax + (M^\dagger)^T Ax - (M^\dagger)^T AM^\dagger Ax) \\ &= (Ax, (M^\dagger + (M^\dagger)^T - (M^\dagger)^T AM^\dagger)Ax). \end{aligned} \tag{4.4}$$

The desired result then follows by transposition of both sides of (4.1), which is $A(M^\dagger)^T M^T = A$. This completes the proof. ∎

From the identity (4.4), we conclude that the assumption (A2) is sufficient for the energy norm convergence. To prove that it is also necessary, we now introduce a pair of conditions (A2a) and (A2b). Together they are equivalent to (A2), as seen in the next lemma.

LEMMA 4.3. [illegible] (A1) [illegible] (A2) [illegible]

(A2a) $\exists \omega \in (0,2)$ [illegible] $(M^\dagger Ax, M^\dagger Ax)_A \le \omega (M^\dagger Ax, Ax)\ \forall x \in \mathbb{R}^n$.
(A2b) $\exists \alpha > 0$ [illegible] $(M^\dagger Ax, M^\dagger Ax)_A \ge \alpha (M^\dagger Ax, M^\dagger Ax)\ \forall x \in \mathbb{R}^n$.

[illegible]. We first show that (A2a) and (A2b) imply (A2). We begin by rewriting (4.4) in the following form:

$$|x|_A^2 - |(I - M^\dagger A)x|_A^2 = \big((M^T + M - A)M^\dagger Ax, M^\dagger Ax\big) = 2(Ax, M^\dagger Ax) - (M^\dagger Ax, M^\dagger Ax)_A. \tag{4.5}$$

We then conclude that (A2) is equivalent to the existence of a constant $\delta > 0$ such that

$$2(y, M^\dagger y) \ge \delta (M^\dagger y, M^\dagger y) + (M^\dagger y, M^\dagger y)_A \quad \forall y \in \mathcal{R}(A). \tag{4.6}$$

From (A2a) and (A2b) we have that

$$\begin{aligned} 2(y, M^\dagger y) &\ge \frac{2}{\omega}(M^\dagger y, M^\dagger y)_A \\ &= (M^\dagger y, M^\dagger y)_A + \left(\frac{2}{\omega} - 1\right)(M^\dagger y, M^\dagger y)_A \\ &\ge (M^\dagger y, M^\dagger y)_A + \left(\frac{2}{\omega} - 1\right)\alpha (M^\dagger y, M^\dagger y). \end{aligned}$$

This means that (4.6) holds with $\delta = (2/\omega - 1)\alpha$, which proves (A2).

To prove the reverse implication, we assume that (A2) is satisfied. Then (A2a) can be obtained from (4.6),

$$2(y, M^\dagger y) \ge \left(\frac{\delta}{\|A\|} + 1\right)(M^\dagger y, M^\dagger y)_A \quad \forall y \in \mathcal{R}(A),$$

and (A2b) can be obtained by using (4.5), the Cauchy–Schwarz inequality, and (4.2), as follows:

$$\|M^\dagger Ax\|^2 \quad 2(Ax, M^\dagger Ax) \le 2|x|_A |M^\dagger Ax|_A \quad \|M^\dagger Ax\| |M^\dagger Ax|_A.$$

This completes the proof. ∎

We would also like to remark that (A2b) is equivalent to the following relation:

$$\mathcal{R}(M^\dagger A) \cap \mathcal{N}(A) = \{0\}. \tag{4.7}$$

When M is invertible, (4.7) has been considered as a part of necessary and sufficient condition for convergence in a work by Szyld [24].

The following theorem is the main result in this paper.

THEOREM 4.4. [illegible] (2.2) [illegible] (A1) [illegible] (A2) [illegible] (A1) (A2a) [illegible] (A2b) [illegible].

[illegible]. The identity (4.4), the assumption (A2), and (4.2) in Lemma 4.2 imply that

$$|x|_A^2 - |(I - M^\dagger A)x|_A^2 \quad \|M^\dagger Ax\|^2 \quad |x|_A^2.$$

This shows that $I - M^\dagger A$ is a contraction in the $|\cdot|_A$ seminorm; i.e.,

$$|(I - M^\dagger A)x|_A \le \delta |x|_A \quad \text{for some } \delta \in [0, 1).$$

We shall now prove that (A2) is also a necessary condition. Let us assume that (A2) does not hold. Then there exists $x \in \mathbb{R}^n$ such that

$$|x|_A^2 - |(I - M^\dagger A)x|_A^2 \leq 0. \tag{4.8}$$

This contradicts the fact that $|I - M^\dagger A|_A < 1$. This completes the proof. ∎

When A is symmetric and positive definite, the assumption (A2a) is the well-known necessary and sufficient condition for the energy norm convergence of the iterative method (2.2). However, as we shall see, for A being only positive semidefinite, (A2a) alone is not sufficient for convergence. For example, whenever $\mathcal{R}(M^\dagger A) = \mathcal{N}(A)$, (A2a) holds true, but it is not difficult to see that in general there will not be energy norm convergence unless (A2b) is added to the set of assumptions.

Example 2: (A2b) [illegible] Let

$$A = \begin{pmatrix} 1/2 & -1 \\ -1 & 2 \end{pmatrix}. \tag{4.9}$$

We introduce a splitting $A = M - N$, where M is given by

$$M^{-1} = \begin{pmatrix} 2 & 2 \\ -1 & 0 \end{pmatrix}.$$

A simple algebraic manipulation yields

$$\mathcal{R}(M^{-1}A) = \mathcal{N}(A).$$

It is then straightforward to see that

$$|E|_A^2 = |I - M^{-1}A|_A^2 = 1.$$

This means that the iteration is not convergent.

As we have seen before, the splitting given in Example 1 above is neither P-regular nor weak-regular. We now revisit the example to show that it in fact satisfies (A2a) and (A2b).

Example 1 [illegible] Consider the splitting given in Example 1. We have that

$$\mathcal{N}(A)^\perp = \operatorname{span}\left\{\begin{pmatrix} -1 \\ 2 \end{pmatrix}\right\}, \qquad \mathcal{N}(A) = \operatorname{span}\left\{\begin{pmatrix} 2 \\ 1 \end{pmatrix}\right\},$$

and

$$\left(\bar{M}\begin{pmatrix} -1 \\ 2 \end{pmatrix}, \begin{pmatrix} -1 \\ 2 \end{pmatrix}\right) = \frac{163}{2}.$$

This implies that the splitting $A = M - N$ satisfies (A2a). We note that since $\mathcal{R}(M^{-1}A) \cap \mathcal{N}(A) = \{0\}$, (A2b) holds true.

Acknowledgment. The authors wish to thank Professor Daniel Szyld for his helpful comments on the content of the paper, as well as for sharing with us many references on results related to the subject of this paper.

REFERENCES

[1] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Classics in Appl. Math. 9, SIAM, Philadelphia, 1994.

[2] A. Berman and M. Neumann, *Consistency and splittings*, SIAM J. Numer. Anal., 13 (1976), pp. 877–888.

[3] P. Bochev and R. B. Lehoucq, *On the finite element solution of the pure Neumann problem*, SIAM Rev., 47 (2005), pp. 50–66.
[4] Z. H. Cao, *A note on properties of splittings of singular symmetric positive semidefinite matrices*, Numer. Math., 88 (2001), pp. 603–606.
[5] Q.-S. Chang and W.-W. Sun, *On convergence of multigrid method for nonnegative definite systems*, J. Comput. Math., 23 (2005), pp. 177–184.
[6] A. Dax, *The convergence of linear stationary iterative processes for solving singular unstructured systems of linear equations*, SIAM Rev., 32 (1990), pp. 611–635.
[7] E. B. Dynkin, *Markov Processes*, Vols. I, II (translated with the authorization and assistance of the author by J. Fabius, V. Greenberg, A. Maitra, G. Majone), Grundlehren Math. Wiss., Bände 121, Academic Press, New York, 1965.
[8] S. Guattery and G. L. Miller, *On the quality of spectral separators*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 701–719.
[9] S. Guattery and G. L. Miller, *Graph embeddings and Laplacian eigenvalues*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 703–723.
[10] V. N. Joshi, *A note on the solution of rectangular linear systems by iteration*, SIAM Rev., 12 (1970), pp. 463–466.
[11] H. B. Keller, *On the solution of singular and semidefinite linear systems by iteration*, J. Soc. Indust. Appl. Math. Ser. B Numer. Anal., 2 (1965), pp. 281–290.
[12] J.-H. Kimn and M. Sarkis, *OBDD: Overlapping balancing domain decomposition methods and generalizations to the Helmholtz equation*, in Proceedings of the 16th International Conference on Domain Decompositions (New York, 2005), Springer-Verlag, New York, 2005, to appear.
[13] Y.-J. Lee, J. Wu, J. Xu, and L. Zikatanov, *A Sharp Convergence Estimate of the Method of Subspace Corrections for Singular System of Equations*, Technical report, The Pennsylvania State University, University Park, PA, 2002.
[14] J. Mandel, *Balancing domain decomposition*, Comm. Numer. Methods Engrg., 9 (1993), pp. 233–241.
[15] I. Marek and D. B. Szyld, *Comparison theorems for the convergence factor of iterative methods for singular matrices*, Linear Algebra Appl., 316 (2000), pp. 67–87.
[16] I. Marek and D. B. Szyld, *Algebraic Schwarz methods for the numerical solution of Markov chains*, Linear Algebra Appl., 386 (2004), pp. 67–81.
[17] R. Nabben and D. B. Szyld, *Schwarz iterations for symmetric positive semidefinite problems*, SIAM J. Matrix Anal., to appear.
[18] S. V. Nepomnyaschikh, *Schwartz alternating method for solving the singular Neumann problem*, Sov. J. Numer. Anal. Math. Modelling, 5 (1990), pp. 69–78.
[19] M. J. O'Carroll, *Inconsistencies and* S.O.R. *convergence for the discrete Neumann problem*, J. Inst. Math. Appl., 11 (1973), pp. 343–350.
[20] J. M. Ortega and W. C. Rheinboldt, *Monotone iterations for nonlinear equations with application to Gauss-Seidel methods*, SIAM J. Numer. Anal., 4 (1967), pp. 171–190.
[21] T. Strouboulis, I. Babuška, and K. Copps, *The design and analysis of the generalized finite element method*, Comput. Methods Appl. Mech. Engrg., 181 (2000), pp. 43–69.
[22] T. Strouboulis, K. Copps, and I. Babuška, *The generalized finite element method: An example of its implementation and illustration of its performance*, Internat. J. Numer. Methods Engrg., 47 (2000), pp. 1401–1417.
[23] T. Strouboulis, K. Copps, and I. Babuška, *The generalized finite element method*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 4081–4193.
[24] D. B. Szyld, *Equivalence of conditions for convergence of iterative methods for singular equations*, Numer. Linear Algebra Appl., 1 (1994), pp. 151–154.
[25] J. Xu, *Iterative methods by space decomposition and subspace correction*, SIAM Rev., 34 (1992), pp. 581–613.
[26] J. Xu and L. Zikatanov, *On multigrid methods for generalized finite element methods*, in Meshfree Methods for Partial Differential Equations (Bonn, 2001), Lecture Notes in Comput. Sci. Eng. 26, Springer, Berlin, 2003, pp. 401–418.
[27] J. Xu and L. Zikatanov, *The method of alternating projections and the method of subspace corrections in Hilbert space*, J. Amer. Math. Soc., 15 (2002) pp. 573–597.
[28] D. M. Young, *Iterative Solution of Large Linear Systems*, Dover, Mineola, NY, 2003 (unabridged republication of the 1971 Academic Press edition).

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 3, pp. 642–666

A LINK BETWEEN THE CANONICAL DECOMPOSITION IN MULTILINEAR ALGEBRA AND SIMULTANEOUS MATRIX DIAGONALIZATION*

LIEVEN DE LATHAUWER†

Abstract. Canonical decomposition is a key concept in multilinear algebra. In this paper we consider the decomposition of higher-order tensors which have the property that the rank is smaller than the greatest dimension. We derive a new and relatively weak deterministic sufficient condition for uniqueness. The proof is constructive. It shows that the canonical components can be obtained from a simultaneous matrix diagonalization by congruence, yielding a new algorithm. From the deterministic condition we derive an easy-to-check dimensionality condition that guarantees generic uniqueness.

Key words. multilinear algebra, higher-order tensor, canonical decomposition, parallel factors model, simultaneous matrix diagonalization

AMS subject classifications. 15A18, 15A69

DOI. 10.1137/040608830

1. Introduction.

An increasing number of problems in signal processing, data analysis, and scientific computing involves the manipulation of quantities whose elements are addressed by more than two indices. In the literature these higher-order analogues of vectors (first order) and matrices (second order) are called higher-order tensors, multidimensional matrices, or multiway arrays. The algebra of higher-order tensors is called multilinear algebra. This paper presents some new contributions concerning a tensor decomposition known as the canonical decomposition (CANDECOMP) [9] or parallel factors model (PARAFAC) [24, 41].

In the following subsection we first introduce some basic definitions. In section 1.2 we have a closer look at the CANDECOMP. In section 1.3 we set out the problem discussed in this paper and define our notation.

1.1. Basic definitions.

DEFINITION 1.1. n-mode vector $(I_1 \times I_2 \times \cdots \times I_N)$ $\mathcal{A}$ I_n i_n $\mathcal{A}$ [27]

DEFINITION 1.2. supersymmetric

DEFINITION 1.3. N $\mathcal{A}$ rank 1 N $U^{(1)}$ $U^{(2)}$ $\ldots$ $U^{(N)}$:

$$a_{i_1 i_2 \ldots i_N} = u^{(1)}_{i_1} u^{(2)}_{i_2} \ldots u^{(N)}_{i_N}$$

*Received by the editors May 23, 2005; accepted for publication (in revised form) by H. J. Woerdeman April 3, 2006; published electronically September 15, 2006. Parts of section 2 have appeared in the conference papers [15, 17]. This research was supported by several institutions: (1) the Flemish Government through (a) Research Council K.U.Leuven (GOA-Ambiorics, CoE EF/05/006 Optimization in Engineering), (b) F.W.O. project G.0321.06, (c) F.W.O. Research Communities ICCoS and ANMMM, and (d) Tournesol 2005; (2) the Belgian Federal Science Policy Office (IUAP P5/22).

http://www.siam.org/journals/simax/28-3/60883.html

†ETIS, UMR 8051, 6 avenue du Ponceau, BP 44, F 95014 Cergy-Pontoise Cedex, France (delathau@ensea.fr, http://www-etis.ensea.fr).

The outer product of $U^{(1)}, U^{(2)}, \ldots, U^{(N)}$ is denoted by $U^{(1)} \circ U^{(2)} \circ \cdots \circ U^{(N)}$.

[illegible] 1. Consider the $(2 \times 2 \times 2)$-tensor $\mathcal{A}$ defined by

$$a_{111} = -a_{121} = 3, \quad a_{211} = -a_{221} = 6, \quad a_{112} = -a_{122} = 1, \quad a_{212} = -a_{222} = 2.$$

The 1-mode, 2-mode, and 3-mode vectors are the columns of the matrices

$$\begin{pmatrix} 3 & -3 & 1 & -1 \\ 6 & -6 & 2 & -2 \end{pmatrix}, \qquad \begin{pmatrix} 3 & 6 & 1 & 2 \\ -3 & -6 & -1 & -2 \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} 3 & 6 & -3 & -6 \\ 1 & 2 & -1 & -2 \end{pmatrix},$$

respectively. The tensor is rank 1 because it can be decomposed as

$$\mathcal{A} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \circ \begin{pmatrix} 1 \\ -1 \end{pmatrix} \circ \begin{pmatrix} 3 \\ 1 \end{pmatrix}.$$

DEFINITION 1.4. [illegible] rank [illegible] $\mathcal{A}$ [illegible] 1 [illegible] $\mathcal{A}$ [illegible] [31]

DEFINITION 1.5. [illegible] scalar product $\langle \mathcal{A}, \mathcal{B} \rangle$ [illegible] $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ [illegible]

$$\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_N=1}^{I_N} a_{i_1 i_2 \ldots i_N} b_{i_1 i_2 \ldots i_N}.$$

This definition generalizes the standard scalar product of vectors ($\langle A, B \rangle = A^T B$) and the standard scalar product of matrices ($\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} a_{ij} b_{ij}$, with $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{I_1 \times I_2}$). Note that, for two $(I_1 \times I_2 \times \cdots \times I_N)$ rank-1 tensors $\mathcal{A} = U_1^{(1)} \circ U_2^{(1)} \circ \cdots \circ U_N^{(1)}$ and $\mathcal{B} = V_1^{(1)} \circ V_2^{(1)} \circ \cdots \circ V_N^{(1)}$, we have

$$\begin{aligned} \langle \mathcal{A}, \mathcal{B} \rangle &= \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_N=1}^{I_N} u_{i_1}^{(1)} u_{i_2}^{(2)} \ldots u_{i_N}^{(N)} v_{i_1}^{(1)} v_{i_2}^{(2)} \ldots v_{i_N}^{(N)} \\ &= (U_1^T V_1)(U_2^T V_2) \ldots (U_N^T V_N). \end{aligned} \tag{1.1}$$

DEFINITION 1.6. [illegible] Frobenius norm [illegible] $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ [illegible]

$$\|\mathcal{A}\| = \sqrt{\langle \mathcal{A}, \mathcal{A} \rangle} = \left(\sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_N=1}^{I_N} a_{i_1 i_2 \ldots i_N}^2 \right)^{\frac{1}{2}}.$$

DEFINITION 1.7. [illegible] Kruskal rank [illegible] k-rank [illegible] $\mathbf{A}$ [illegible] $\text{rank}_k(\mathbf{A})$ [illegible] r [illegible] r [illegible] $\mathbf{A}$ [illegible] [31]

By definition, we have that $\text{rank}_k(\mathbf{A}) \leqslant \text{rank}(\mathbf{A})$.

[illegible] 2. Consider the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 4 \\ 2 & 1 & 2 \end{pmatrix},$$

which has rank 2. The k-rank of $\mathbf{A}$ is 1, because its last two columns are proportional.

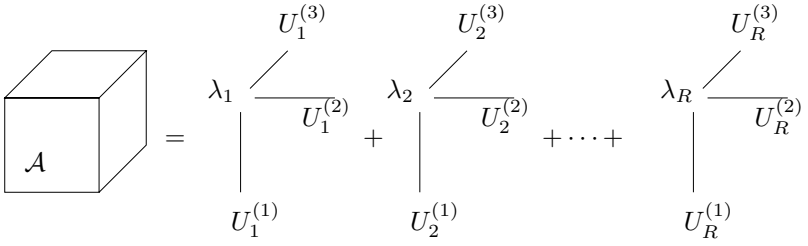


FIG. 1.1. *Visualization of the CANDECOMP for a third-order tensor.*

1.2. The canonical decomposition. We now introduce the decomposition that is dealt with in this paper.

DEFINITION 1.8. canonical decomposition $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ $\mathcal{A}$ 1 :

$$\mathcal{A} = \sum_{r=1}^{R} \lambda_r \, U_r^{(1)} \circ U_r^{(2)} \circ \cdots \circ U_r^{(N)}. \tag{1.2}$$

1.1

The supersymmetric variant in which $U_r^{(1)} = U_r^{(2)} = \cdots = U_r^{(N)}$, $r = 1, \ldots, R$, was already studied in the nineteenth century in the context of invariant theory [11]. Around 1970, the unsymmetric decomposition was independently introduced in psychometrics [9] and phonetics [24]. Later on, the decomposition was also applied in chemometrics and the food industry [1, 6, 41]. In these various disciplines the CANDECOMP is used for the purpose of multiway factor analysis. The term "canonical decomposition" is standard in psychometrics, while in chemometrics the decomposition is called a "parallel factors model." Recently, the CANDECOMP has found important applications in signal processing. In wireless telecommunications, it provides powerful means for the exploitation of different types of diversity [38, 39]. It also describes the basic structure of higher-order cumulants of multivariate data on which all algebraic methods for independent component analysis (ICA) are based [10, 14, 26]. Moreover, decomposition is finding its way into scientific computing, where it leads to a way around the curse of dimensionality [4, 5].

To a large extent, the practical importance of the CANDECOMP stems from its uniqueness properties. It is clear that one can arbitrarily permute the different rank-1 terms. Also, the factors of a same rank-1 term may be arbitrarily scaled, as long as their product remains the same. We call a CANDECOMP unique when it is only subject to these trivial indeterminacies. The following theorem establishes a condition under which uniqueness is guaranteed.

THEOREM 1.9. (1.2)

$$\sum_{n=1}^{N} \operatorname{rank}_k(\mathbf{U}^{(n)}) \geqslant 2R + N - 1, \tag{1.3}$$

R N

This theorem was first proved for real third-order tensors in [31]. A concise proof that also applies to complex tensors was given in [38]. The result was generalized to tensors of arbitrary order in [40].

Note that, contrary to singular value decomposition (SVD) in the matrix case, no orthogonality constraints are imposed on the matrices $\mathbf{U}^{(n)}$ to ensure uniqueness. Imposing orthogonality constraints yields a different decomposition that has different properties [20, 29, 30].

Contrary to matrices, there is no easy way to find the rank of higher-order tensors, except for some special cases [11, 16]. In addition, the rank of an $(I_1 \times I_2 \times \cdots \times I_N)$-tensor is not bounded by $\min(I_1, I_2, \ldots, I_N)$ [11, 31]. The determination of the rank of a given tensor is usually a matter of trial and error.

For a given R, it is common practice to look for the canonical components by straightforward minimization of the quadratic cost function

$$f(\hat{\mathcal{A}}) = \|\mathcal{A} - \hat{\mathcal{A}}\|^2 \tag{1.4}$$

over all rank-R tensors $\hat{\mathcal{A}}$, which we will parametrize as

$$\hat{\mathcal{A}} = \sum_{r=1}^{R} \hat{\lambda}_r \, \hat{U}_r^{(1)} \circ \hat{U}_r^{(2)} \circ \cdots \circ \hat{U}_r^{(N)}. \tag{1.5}$$

It is possible to resort to an alternating least-squares (ALS) algorithm, in which the vector estimates are updated mode per mode [6, 9, 38]. The idea is as follows. Define

$$\begin{aligned} \hat{\mathbf{U}}^{(n)} &= [\hat{U}_1^{(n)} \ \hat{U}_2^{(n)} \ \ldots \ \hat{U}_R^{(n)}], \\ \hat{\Lambda} &= \operatorname{diag}([\hat{\lambda}_1 \ \hat{\lambda}_2 \ \ldots \ \hat{\lambda}_R]). \end{aligned}$$

Now let us imagine that the matrices $\hat{\mathbf{U}}^{(m)}$, $m \neq n$, are fixed and that the only unknowns are the components of the matrix $\hat{\mathbf{U}}^{(n)} \cdot \hat{\Lambda}$. Because of the multilinearity of the CANDECOMP, the estimation of these components is a classical linear least squares problem. An ALS iteration consists of repeating this procedure for different mode numbers: in each step the estimate of one of the matrices $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(N)}$ is optimized, while the other matrix estimates are kept constant. In [34] a Gauss–Newton method is described, in which all the factors of the CANDECOMP are updated simultaneously; in addition, the inherent indeterminacy of the decomposition has been fixed by adding a quadratic regularization constraint on the component entries. We also mention that the canonical components can in principle not be obtained by means of a deflation algorithm. The reason is that the best rank-1 approximation of $\mathcal{A}$ generally does not correspond to one of the terms in (1.2), and that the residue is in general not of rank $R-1$ [13, 28, 45].

In [16] we studied the special case of an $(I_1 \times I_2 \times I_3)$-tensor $\mathcal{A}$ of which (i) the rank $R \leqslant \min(I_1, I_2)$, (ii) the set $\{U_r^{(1)}\}_{(1 \leqslant r \leqslant R)}$ is linearly independent, (iii) the set $\{U_r^{(2)}\}_{(1 \leqslant r \leqslant R)}$ is linearly independent, and (iv) the set $\{U_r^{(3)}\}_{(1 \leqslant r \leqslant R)}$ does not contain collinear vectors. In this case, the canonical components can be obtained from a simultaneous matrix decomposition. Simultaneous matrix decompositions have become an important tool for signal processing and data analysis in the last decade [2, 3, 8, 19, 23, 35, 36, 42, 43, 44]. Let us, for instance, consider a simultaneous diagonalization by congruence:

$$\begin{aligned} \mathbf{M}_1 &= \mathbf{W} \cdot \Lambda_1 \cdot \mathbf{W}^T \\ &\vdots \\ \mathbf{M}_N &= \mathbf{W} \cdot \Lambda_N \cdot \mathbf{W}^T, \end{aligned} \tag{1.6}$$

in which $\mathbf{M}_1, \ldots, \mathbf{M}_N \in \mathbb{R}^{P \times P}$ are given symmetric matrices, $\mathbf{W} \in \mathbb{R}^{P \times P}$ is an unknown nonsingular matrix, and $\Lambda_1, \ldots, \Lambda_N \in \mathbb{R}^{P \times P}$ are unknown diagonal matrices. Theoretically, $\mathbf{W}$ can already be obtained from two of these decompositions. Let us assume for convenience that $\mathbf{M}_n$ is nonsingular and that all the diagonal entries of $\Lambda_m \cdot \Lambda_n^{-1}$ are mutually different. Then $\mathbf{W}$ follows from the eigenvalue decomposition (EVD) $\mathbf{M}_m \cdot \mathbf{M}_n^{-1} = \mathbf{W} \cdot \Lambda_m \cdot \Lambda_n^{-1} \cdot \mathbf{W}^{-1}$ [32]. From a numerical point of view, it is preferable to take all the equations in (2.13) into account when the matrices $\mathbf{M}_1, \ldots, \mathbf{M}_N$ are only known with limited precision. Equation (2.13) then has to be solved in some optimal way—for instance, by minimizing

$$g(\hat{\mathbf{W}}, \hat{\Lambda}_1, \ldots, \hat{\Lambda}_N) = \sum_{n=1}^{N} \|\mathbf{M}_n - \hat{\mathbf{W}} \cdot \hat{\Lambda}_n \cdot \hat{\mathbf{W}}^T\|^2.$$

1.3. This paper. In this paper we consider the special case of tensors that are tall in one mode. More precisely we assume that $I_N \geqslant R$. This case occurs very often in practice. The tall mode may, for instance, be formed by different samples over time or different samples from a population. Note that in this case condition (1.3) generically reduces to

$$\sum_{n=1}^{N-1} \min(I_n, R) \geqslant R + N - 1. \tag{1.7}$$

(We call a property generic when it holds everywhere, except for a set of Lebesgue measure 0.) Hence, the maximum value R for which uniqueness of the CANDECOMP is guaranteed is bounded by $\sum_{n=1}^{N-1} I_n - N + 1$.

In this paper we derive a new sufficient condition for uniqueness in the case that $I_N \geqslant R$. The proof is constructive. It shows that the canonical components can be obtained from a simultaneous matrix diagonalization by congruence. The case of third-order tensors is treated in section 2. Fourth-order tensors are discussed in section 3. Along these lines, the approach can be generalized to tensors of arbitrary order. In section 4 some numerical results are shown. The presentation is in terms of real tensors. Complex tensors can be dealt with in the same way.

The derivation in section 2.1 was inspired by the ICA algorithm presented in [7]. In the latter paper, a "rank-1 detecting device" was proposed that is similar to mapping Φ in Theorem 2.1. It was subsequently shown that this device could be used to find the ICA solution from the fourth-order cumulant tensor of the data via an EVD of a real symmetric matrix. In the derivation the symmetries of the cumulant tensor were exploited. Here we only make use of the algebraic structure of the CANDECOMP. The canonical components are computed by means of the (approximate) simultaneous decomposition of a set of matrices instead of the decomposition of a single matrix. The ICA application is worked out in more detail in [18].

1.4. Notation. In this paper scalars are denoted by lowercase italic letters (a, b, ...), vectors are written as italic capitals (A, B, ...), matrices correspond to boldface capitals ($\mathbf{A}$, $\mathbf{B}$, ...), and tensors are written as calligraphic letters ($\mathcal{A}$, $\mathcal{B}$, ...). This notation is consistently used for lower-order parts of a given structure. For instance, the entry with row index i and column index j in a matrix $\mathbf{A}$, i.e., $(\mathbf{A})_{ij}$, is symbolized by a_{ij} (also $(A)_i = a_i$ and $(\mathcal{A})_{i_1 i_2 \ldots i_N} = a_{i_1 i_2 \ldots i_N}$). The ith column vector of a matrix $\mathbf{A}$ is denoted as A_i, i.e., $\mathbf{A} = [A_1 A_2 \ldots]$. Italic capitals are also used to denote index upper bounds (e.g., $i = 1, 2, \ldots, I$). The zero tensor is denoted by $\mathcal{O}$. The symbol $\otimes$

denotes the Kronecker product,

$$\mathbf{A} \otimes \mathbf{H} \stackrel{\text{def}}{=} \begin{pmatrix} a_{11}\mathbf{H} & a_{12}\mathbf{H} & \dots \\ a_{21}\mathbf{H} & a_{22}\mathbf{H} & \dots \\ \vdots & \vdots & \end{pmatrix},$$

and $\odot$ represents the Khatri–Rao or columnwise Kronecker product [37]:

$$\mathbf{A} \odot \mathbf{H} \stackrel{\text{def}}{=} (A_1 \otimes H_1 \dots A_R \otimes H_R).$$

The operator diag$(\cdot)$ stacks its vector argument in a square diagonal matrix. We denote the 2-norm condition number of a matrix, i.e., the ratio of its largest to its smallest singular value, by cond$(\cdot)$. The $(N \times N)$ identity matrix is represented by $\mathbf{I}_{N\times N}$. The $(I \times J)$ zero matrix is denoted by $\mathbf{0}_{I\times J}$. Finally, $\mathbf{P}_{J\cdot I\times I\cdot J}$ is the $(IJ \times IJ)$ permutation matrix of which the entries at positions $((j-1)I+i, (i-1)J+j)$, $i = 1, 2, \dots, I$, $j = 1, 2, \dots, J$, are equal to one, the other entries being equal to zero.

2. The third-order case.

2.1. Deterministic uniqueness condition and algorithm. Consider an $(I \times J \times K)$-tensor $\mathcal{T}$ of which the CANDECOMP is given by

$$t_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} \qquad \forall i, j, k \tag{2.1}$$

in which $\mathbf{A} \in \mathbb{R}^{I\times R}$, $\mathbf{B} \in \mathbb{R}^{J\times R}$, $\mathbf{C} \in \mathbb{R}^{K\times R}$. We assume that $\min(IJ, K) \geqslant R$. Consider a matrix $\mathbf{T} \in \mathbb{R}^{IJ\times K}$ in which the entries of $\mathcal{T}$ are stacked as follows:

$$(\mathbf{T})_{(i-1)J+j,k} = t_{ijk}.$$

We have

$$\mathbf{T} = (\mathbf{A} \odot \mathbf{B}) \cdot \mathbf{C}^T. \tag{2.2}$$

We assume that both $\mathbf{A} \odot \mathbf{B}$ and $\mathbf{C}$ are full column rank. Both conditions are generically satisfied if $R \leqslant \min(IJ, K)$, as will be explained in section 2.2. Note that, in this case, the rank of the tensor $\mathcal{T}$ is equal to the rank of its matrix representation $\mathbf{T}$. We notice that if $\mathbf{A} \odot \mathbf{B}$ is not full column rank, (2.1) is not unique [33]. As a matter of fact, in this case a decomposition with a smaller number of terms is possible. (If, for instance, $A_R \otimes B_R = \sum_{r=1}^{R-1} \alpha_r A_r \otimes B_r$, then $\mathcal{T} = \sum_{r=1}^{R-1} A_r \circ B_r \circ (C_r + \alpha_r C_R)$.) On the other hand, if $\mathbf{C}$ is not full column rank, then the rank of $\mathcal{T}$ may nevertheless be equal to R and the CANDECOMP may still be unique (e.g., (1.3) may be satisfied). In that case, the rank of $\mathcal{T}$ cannot be estimated as the rank of $\mathbf{T}$, and the algorithm below will fail.

Consider a factorization of $\mathbf{T}$ of the form

$$\mathbf{T} = \mathbf{E} \cdot \mathbf{F}^T, \tag{2.3}$$

with $\mathbf{E} \in \mathbb{R}^{IJ\times R}$ and $\mathbf{F} \in \mathbb{R}^{K\times R}$ full column rank. Because of (2.2) and (2.3), we have

$$\mathbf{A} \odot \mathbf{B} = \mathbf{E} \cdot \mathbf{W} \tag{2.4}$$

for some nonsingular $\mathbf{W} \in \mathbb{R}^{R\times R}$. The task is now to find $\mathbf{W}$ such that the columns of $\mathbf{E}\cdot\mathbf{W}$ are Kronecker products. A vector that is equal to the Kronecker product of a vector $A \in \mathbb{R}^I$ and a vector $B \in \mathbb{R}^J$ can be represented as an $(I \times J)$ rank-1 matrix; cf. (2.14)–(2.15) below. Matrices with rank at most 1 and matrices of which the rank is strictly greater than 1 can be distinguished by means of the bilinear mapping introduced in the following theorem.

THEOREM 2.1. [illegible] $\Phi : (\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^{I\times J} \times \mathbb{R}^{I\times J} \to \Phi(\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^{I\times I\times J\times J}$ [illegible]

$$(\Phi(\mathbf{X}, \mathbf{Y}))_{ijkl} = x_{ik}y_{jl} + y_{ik}x_{jl} - x_{il}y_{jk} - y_{il}x_{jk}. \tag{2.5}$$

[illegible] $\Phi(\mathbf{X}, \mathbf{X}) = \mathcal{O}$ [illegible] $\mathbf{X}$ [illegible] 1 [illegible]. The "if" part is obvious. For the "only if" part, let the SVD of $\mathbf{X}$ be given by $\mathbf{U}\cdot\Sigma\cdot\mathbf{V}^T$, with $\Sigma = \mathrm{diag}([\sigma_1 \ \ldots \ \sigma_M])$, where $M = \min(I, J)$. We have

$$\begin{aligned} x_{ik}x_{jl} &= \sum_{r,s=1}^{M} \sigma_r\sigma_s u_{ir}v_{kr}u_{js}v_{ls}, \\ x_{il}x_{jk} &= \sum_{r,s=1}^{M} \sigma_r\sigma_s u_{ir}v_{lr}u_{js}v_{ks}, \\ \Phi(\mathbf{X}, \mathbf{X}) &= 2\sum_{r,s=1}^{M} \sigma_r\sigma_s(U_r \circ U_s \circ V_r \circ V_s - U_r \circ U_s \circ V_s \circ V_r). \end{aligned} \tag{2.6}$$

Rank-1 terms corresponding to the same $r = s$ cancel out in (2.6). Due to the orthogonality of $\mathbf{U}$ and $\mathbf{V}$, the other terms are mutually orthogonal, as can be verified by means of (1.1). Because of the linear independence of these terms, we must have that $\sigma_r\sigma_s = 0$ whenever $r \neq s$. Hence, Σ is at most rank 1.

Another way to see this is by observing that the entries of $\Phi(\mathbf{X}, \mathbf{X})/2$ correspond to the determinants of the different (2×2) submatrices of $\mathbf{X}$. A necessary and sufficient condition for $\mathbf{X}$ to be at most rank 1 is that all these determinants vanish. □

Define matrices $\mathbf{E}_1, \ldots, \mathbf{E}_R \in \mathbb{R}^{I\times J}$ corresponding to each column of $\mathbf{E}$ in (2.3) so that

$$(\mathbf{E}_r)_{ij} = e_{(i-1)J+j,r} \qquad \forall i, j, r$$

and let $\mathcal{P}_{rs} = \Phi(\mathbf{E}_r, \mathbf{E}_s)$. Note that Φ is symmetric in its arguments; hence

$$\mathcal{P}_{rs} = \mathcal{P}_{sr} \qquad \forall r, s. \tag{2.7}$$

Since Φ is bilinear, we have from (2.4)

$$\mathcal{P}_{rs} = \sum_{t,u=1}^{R} (\mathbf{W}^{-1})_{tr}(\mathbf{W}^{-1})_{us}\Phi\left(A_tB_t^T,\ A_uB_u^T\right). \tag{2.8}$$

Assume at this point that there exists a symmetric matrix $\mathbf{M}$ of which the entries satisfy the following set of homogeneous linear equations (we will justify this assumption below):

$$\sum_{r,s=1}^{R} m_{rs}\mathcal{P}_{rs} = \mathcal{O}. \tag{2.9}$$

Substitution of (2.8) in (2.9) yields

$$\sum_{r,s=1}^{R}\sum_{t,u=1}^{R}(\mathbf{W}^{-1})_{tr}(\mathbf{W}^{-1})_{us}m_{rs}\Phi\left(A_tB_t^T,\ A_uB_u^T\right)=\mathcal{O}.$$

According to Theorem 2.1, we have $\Phi\left(A_tB_t^T,\ A_tB_t^T\right)=\mathcal{O}$, $1\leqslant t\leqslant R$. Hence

$$\sum_{r,s=1}^{R}\sum_{\substack{t,u=1\\ t\neq u}}^{R}(\mathbf{W}^{-1})_{tr}(\mathbf{W}^{-1})_{us}m_{rs}\Phi\left(A_tB_t^T,\ A_uB_u^T\right)=\mathcal{O}.$$

Furthermore, due to (2.7) and the symmetry of $\mathbf{M}$ we have

$$\sum_{r,s=1}^{R}\sum_{\substack{t,u=1\\ t<u}}^{R}(\mathbf{W}^{-1})_{tr}(\mathbf{W}^{-1})_{us}m_{rs}\Phi\left(A_tB_t^T,\ A_uB_u^T\right)=\mathcal{O}. \tag{2.10}$$

Denote

$$\lambda_{tu}=\sum_{r,s=1}^{R}(\mathbf{W}^{-1})_{tr}(\mathbf{W}^{-1})_{us}m_{rs}. \tag{2.11}$$

Let us now make the crucial assumption that the tensors $\Phi\left(A_tB_t^T,\ A_uB_u^T\right)$, $1\leqslant t<u\leqslant R$, are linearly independent. Then (2.10) implies that $\lambda_{tu}=0$ when $t\neq u$. As a consequence, (2.11) can be written in a matrix format as

$$\mathbf{M}=\mathbf{W}\cdot\Lambda\cdot\mathbf{W}^T, \tag{2.12}$$

in which Λ is diagonal. Actually, one can see that any diagonal matrix Λ generates a matrix $\mathbf{M}$ that satisfies (2.9). Hence, if the tensors $\{\Phi(A_tB_t^T,A_uB_u^T)\}_{t<u}$ are linearly independent, these matrices form an R-dimensional subspace of the symmetric $(R\times R)$ matrices. Let $\{\mathbf{M}_r\}$ represent a basis of this subspace. We have

$$\begin{aligned}\mathbf{M}_1&=\mathbf{W}\cdot\Lambda_1\cdot\mathbf{W}^T\\ &\vdots\\ \mathbf{M}_R&=\mathbf{W}\cdot\Lambda_R\cdot\mathbf{W}^T,\end{aligned} \tag{2.13}$$

in which $\Lambda_1,\ldots,\Lambda_R$ are diagonal. Equation (2.13) is of the form (1.6). The matrix $\mathbf{W}$ can be determined from this simultaneous matrix decomposition by means of the algorithms presented in [6, 9, 16, 19, 34, 42, 43, 44]. Comparing these algorithms is outside the scope of this paper.

Once $\mathbf{W}$ is known, $\mathbf{A}\odot\mathbf{B}$ can be obtained from (2.4). Let the columns of $\mathbf{A}\odot\mathbf{B}$ be mapped to $(I\times J)$ matrices $\mathbf{G}_r$ as follows:

$$(\mathbf{G}_r)_{ij}=(\mathbf{A}\odot\mathbf{B})_{(i-1)J+j,r},\quad r=1,\ldots,R. \tag{2.14}$$

Then we have

$$\mathbf{G}_r=A_rB_r^T,\quad r=1,\ldots,R, \tag{2.15}$$

from which $\mathbf{A}$ and $\mathbf{B}$ can be obtained. On the other hand, from (2.2), (2.3), and (2.4) it follows that

$$\mathbf{C} = \mathbf{F} \cdot \mathbf{W}^{-T}. \tag{2.16}$$

Equation (2.13) can also be interpreted as the CANDECOMP of a cubic $(R \times R \times R)$-tensor $\mathcal{M}$ of rank R. In $\mathcal{M}$, the matrices $\mathbf{M}_1, \ldots, \mathbf{M}_R$ are stacked as follows:

$$m_{ijk} = (\mathbf{M}_k)_{ij} \qquad \forall i, j, k.$$

Define a matrix $\mathbf{L} \in \mathbb{R}^{R \times R}$ as follows:

$$\mathbf{L} = \begin{pmatrix} (\Lambda_1)_{11} & \cdots & (\Lambda_1)_{RR} \\ \vdots & & \vdots \\ (\Lambda_R)_{11} & \cdots & (\Lambda_R)_{RR} \end{pmatrix}.$$

Then (2.13) can be written as

$$\mathcal{M} = \sum_{r=1}^{R} W_r \circ W_r \circ L_r,$$

which is indeed a CANDECOMP of $\mathcal{M}$. Hence the computation of the CANDECOMP (2.1), with possibly $R < I$ and/or $R < J$, has been reformulated as a problem of the type discussed in [16].

We conclude that the CANDECOMP in (2.1) is unique if $\mathbf{C}$ is full column rank and if the tensors $\{\Phi(A_t B_t^T, A_u B_u^T)\}_{1 \leqslant t < u \leqslant R}$ are linearly independent. This is an easy-to-check deterministic sufficient (but not necessary) condition for uniqueness. If it is satisfied, the canonical components may be computed from the equations derived above. Algorithm 2.1 summarizes the procedure.

If $\mathbf{C}$ is column rank deficient, and $\operatorname{rank}(\mathcal{T}) = R$, then the algorithm fails, as already explained above. If $\{\Phi(A_t B_t^T, A_u B_u^T)\}_{1 \leqslant t < u \leqslant R}$ are linearly dependent, then (2.9) has solutions that cannot be decomposed as in (2.12), and the algorithm fails as well.

In practice, tensor $\mathcal{T}$ may only be known with limited precision. In this respect, some comments concerning the practical implementation of Algorithm 2.1 are in order:

, , 2. The rank R may be obtained as the number of significant singular values of $\mathbf{T}$.

, , 3. This factorization may, for instance, be obtained as follows: Let the SVD of $\mathbf{T}$ be given by $\mathbf{T} = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^T$. Let $\tilde{\mathbf{U}} \in \mathbb{R}^{I \times R}$, $\tilde{\mathbf{S}} \in \mathbb{R}^{R \times R}$, $\tilde{\mathbf{V}} \in \mathbb{R}^{J \times R}$ denote the dominant part of $\mathbf{U}$, $\mathbf{S}$, $\mathbf{V}$, respectively. Then we may take $\mathbf{E}$ and $\mathbf{F}$ equal to

$$\mathbf{E} = \tilde{\mathbf{U}} \cdot \tilde{\mathbf{S}}, \qquad \mathbf{F} = \tilde{\mathbf{V}}.$$

, , 4. Actually only $\mathcal{P}_{rs}$, $r \leqslant s$, have to be computed, because of (2.7).

, , 6. Because of (2.7) and the symmetry of $\mathbf{M}$, the equation can be written as

$$\sum_{s=1}^{R} m_{ss} \mathcal{P}_{ss} + 2 \sum_{\substack{s,t=1 \\ s<t}}^{R} m_{st} \mathcal{P}_{st} = \mathcal{O}. \tag{2.17}$$

This equation has to be solved in the least-squares sense. Stack $\mathcal{P}_{st}$ in a vector $P_{st} \in \mathbb{R}^{I^2 J^2}$, $1 \leqslant r \leqslant s \leqslant R$. Let the R singular vectors of the coefficient matrix

ALGORITHM 2.1

$\mathcal{T} \in \mathbb{R}^{I\times J\times K}$ satisfying

$$\mathcal{T} = \sum_{r=1}^{R} A_r \circ B_r \circ C_r,$$

with both $\{C_r\}_{1\leqslant r\leqslant R}$ and $\{\Phi(A_t B_t^T, A_u B_u^T)\}_{1\leqslant t<u\leqslant R}$ linearly independent.

rank R and CANDECOMP factor matrices $\mathbf{A} \in \mathbb{R}^{I\times R}$, $\mathbf{B} \in \mathbb{R}^{J\times R}$, $\mathbf{C} \in \mathbb{R}^{K\times R}$.

1. Stack $\mathcal{T}$ in $\mathbf{T} \in \mathbb{R}^{IJ\times K}$ as follows:

$$(\mathbf{T})_{(i-1)J+j,k} = (\mathcal{T})_{ijk} \qquad \forall i, j, k.$$

2. $R = \operatorname{rank}(\mathbf{T})$.
3. Compute factorization

$$\mathbf{T} = \mathbf{E} \cdot \mathbf{F}^T,$$

with $\mathbf{E} \in \mathbb{R}^{IJ\times R}$ and $\mathbf{F} \in \mathbb{R}^{K\times R}$ full column rank.

4. Stack $\mathbf{E}$ in $\mathcal{E} \in \mathbb{R}^{I\times J\times R}$ as follows:

$$(\mathcal{E})_{ijr} = (\mathbf{E})_{(i-1)J+j,r} \qquad \forall i, j, r.$$

5. Compute $\mathcal{P}_{rs} \in \mathbb{R}^{I\times I\times J\times J}$, $1 \leqslant r, s \leqslant R$, as follows:

$$(\mathcal{P}_{rs})_{ijkl} = e_{ikr}e_{jls} + e_{iks}e_{jlr} - e_{ilr}e_{jks} - e_{ils}e_{jkr} \qquad \forall i, j, k, l.$$

6. Compute the kernel of

$$\sum_{s,t=1}^{R} m_{st}\mathcal{P}_{st} = \mathcal{O}$$

under the constraint $m_{st} = m_{ts}$ $\forall s, t$. Stack R linearly independent solutions in symmetric matrices $\mathbf{M}_1, \ldots, \mathbf{M}_R \in \mathbb{R}^{R\times R}$.

7. Determine $\mathbf{W} \in \mathbb{R}^{R\times R}$ that simultaneously diagonalizes $\mathbf{M}_1, \ldots, \mathbf{M}_R$:

$$\begin{aligned} \mathbf{M}_1 &= \mathbf{W} \cdot \Lambda_1 \cdot \mathbf{W}^T \\ &\vdots \\ \mathbf{M}_R &= \mathbf{W} \cdot \Lambda_R \cdot \mathbf{W}^T. \end{aligned}$$

8. $\mathbf{A} \odot \mathbf{B} = \mathbf{E} \cdot \mathbf{W}$ and $\mathbf{C} = \mathbf{F} \cdot \mathbf{W}^{-T}$.
9. Stack $\mathbf{A} \odot \mathbf{B}$ in $\mathbf{G}_1, \ldots, \mathbf{G}_R \in \mathbb{R}^{I\times J}$ as follows:

$$(\mathbf{G}_r)_{ij} = (\mathbf{A} \odot \mathbf{B})_{(i-1)J+j,r} \qquad \forall i, j.$$

10. Obtain A_r, B_r from

$$\mathbf{G}_r = A_r B_r^T \qquad \forall r.$$

$[P_{11}, \ldots, P_{RR}, 2P_{12}, 2P_{13}, \ldots, 2P_{R-1,R}]$, corresponding to the smallest singular values, be denoted by $(w_{1,1,r}, \ldots, w_{R,R,r}, w_{1,2,r}, w_{1,3,r}, \ldots, w_{R-1,R,r})^T$, $1 \leqslant r \leqslant R$. Then we may take $\mathbf{M}_r$ equal to

$$\mathbf{M}_r = \begin{pmatrix} w_{1,1,r} & w_{1,2,r} & \cdots & w_{1,R,r} \\ w_{1,2,r} & w_{2,2,r} & \cdots & w_{2,R,r} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1,R,r} & w_{2,R,r} & \cdots & w_{R,R,r} \end{pmatrix} \quad \forall r.$$

7. The matrices $\mathbf{M}_r$ may be weighted according to their expected relative precision. The singular values of the coefficient matrix in step 6 give an indication of this precision.

10. A_r and B_r are obtained from the best rank-1 approximation of $\mathbf{G}_r$.

1. It turns out that our deterministic sufficient condition for uniqueness has also, in an entirely different manner, been derived in [22]. In that paper, a matrix $\mathbf{U} \in \mathbb{C}^{I^2J^2 \times R(R-1)/2}$ is defined as follows:

(2.18)

$$(\mathbf{U})_{(i_1-1)(IJ^2)+(i_2-1)J^2+(j_1-1)J+j_2, \frac{(u-2)(u-1)}{2}+t} = \begin{vmatrix} a_{it} & a_{iu} \\ a_{kt} & a_{ku} \end{vmatrix} \cdot \begin{vmatrix} a_{jt} & a_{ju} \\ a_{lt} & a_{lu} \end{vmatrix},$$

$$1 \leqslant i_1, i_2 \leqslant I, \qquad 1 \leqslant j_1, j_2 \leqslant J, \qquad 1 \leqslant t < u \leqslant R.$$

It is shown that the CANDECOMP is unique if $\mathbf{U}$ and $\mathbf{C}$ are full column rank. It is easy to verify that

$$(\mathbf{U})_{(i_1-1)(IJ^2)+(i_2-1)J^2+(j_1-1)J+j_2, \frac{(u-2)(u-1)}{2}+t} = \left(\Phi(A_tB_t^T, A_uB_u^T)\right)_{i_1i_2j_1j_2}. \tag{2.19}$$

In other words, the columns of $\mathbf{U}$ are vector representations of the tensors $\{\Phi(A_tB_t^T, A_uB_u^T)\}_{1 \leqslant t < u \leqslant R}$. Hence, the uniqueness conditions in this paper and in [22] are the same.

2.2. Generic uniqueness condition. In this section we examine under which conditions on R both $\{C_r\}_{1 \leqslant r \leqslant R}$ and $\{\Phi(A_tB_t^T, A_uB_u^T)\}_{1 \leqslant t < u \leqslant R}$ are generically linearly independent. We will derive bounds on R that depend only on the dimensions of the tensor. A generic tensor whose rank and dimensions satisfy these constraints has a CANDECOMP that is unique and comprises components that can be computed by means of Algorithm 2.1. We start from the following lemma.

LEMMA 2.2. $\mathbf{A} \in \mathbb{R}^{I \times R}$ $\mathbf{B} \in \mathbb{R}^{J \times R}$

$$\operatorname{rank}(\mathbf{A} \odot \mathbf{B}) = \min(IJ, R).$$

. Denote $\tilde{R} = \operatorname{rank}(\mathbf{A} \odot \mathbf{B})$. Let us assume that $\tilde{R} < \min(IJ, R)$. The theorem follows from the observation that a generic perturbation of the vectors $A_r \odot B_r$ makes the set linearly independent. Let us map $A_r \odot B_r$ to the $(I \times J)$ matrix $A_rB_r^T$, $r = 1, \ldots, R$. Assume, without loss of generality, that $A_1B_1^T$ lies in the vector space V generated by $A_rB_r^T$, $r = 2, 3, \ldots, R$. It suffices to prove that a generic perturbation of $A_1B_1^T$ does not lie in V. Let $\mathbf{V}^\perp \in \mathbb{R}^{I \times J}$ be orthogonal to V, i.e., the scalar product of $\mathbf{V}^\perp$ and any matrix in V is zero. We have $\langle A_1B_1^T, \mathbf{V}^\perp \rangle = A_1^T\mathbf{V}^\perp B_1 = 0$. Let the perturbed version of $A_1B_1^T$ be denoted by $\tilde{A}_1\tilde{B}_1^T$. Generically we have $\langle \tilde{A}_1\tilde{B}_1^T, \mathbf{V}^\perp \rangle = \tilde{A}_1^T\mathbf{V}^\perp\tilde{B}_1 \neq 0$, i.e., the perturbation has a component orthogonal

to V. As a consequence, $\tilde{A}_1 \odot \tilde{B}_1$ has a component orthogonal to $A_r \odot B_r$, $r = 2, 3, \ldots, R$. □

2. That matrices **A** and **B** are full rank or full k-rank does not guarantee their Khatri–Rao product will be full rank. Consider, for instance,

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & -1 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 2 & -2 \end{pmatrix}.$$

We have $\text{rank}(\mathbf{A}) = \text{rank}_k(\mathbf{A}) = \text{rank}(\mathbf{B}) = \text{rank}_k(\mathbf{B}) = 2$. However,

$$\mathbf{A} \odot \mathbf{B} = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 0 & 2 & -2 \\ 0 & 0 & 1 & -1 \\ 0 & 1 & 2 & 2 \end{pmatrix},$$

such that $\text{rank}(\mathbf{A} \odot \mathbf{B}) = \text{rank}_k(\mathbf{A} \odot \mathbf{B}) = 3 < 4$.

Before continuing with Lemmas 2.3 and 2.4, we explain the intuition behind Lemma 2.2. We start with a geometric description of the surface formed by the matrices of which the rank is at most 1; cf. [12, 25] and the references therein.

Let S_N be the sphere consisting of the unit-norm vectors in $\mathbb{R}^N$. Define the outer product $\mathsf{S}_I \times \mathsf{S}_J$ as the set formed by the outer products of any vector on S_I and any vector on S_J. This corresponds to the set of unit-norm rank-1 matrices in $\mathbb{R}^{I \times J}$. It consists of two disjoint parts, consisting of the positive and negative semidefinite rank-1 matrices, respectively. Each of these parts corresponds to a highly symmetric surface in $\mathbb{R}^{I \times J}$. Namely, each part is mapped onto itself by any transformation of the form

$$f : \mathbb{R}^{I \times J} \to \mathbb{R}^{I \times J} : \mathbf{X} \to f(\mathbf{X}) = \mathbf{Q}_I \cdot \mathbf{X} \cdot \mathbf{Q}_J,$$

in which $\mathbf{Q}_I$ and $\mathbf{Q}_J$ are orthogonal matrices in $\mathbb{R}^{I \times I}$ and $\mathbb{R}^{J \times J}$, respectively, representing rotations and/or reflections. The full set of $(I \times J)$ matrices of which the rank is at most 1, represented by $\mathbb{R}^{I \times J}_{R \leqslant 1}$, is obtained by allowing arbitrary scalings of the elements of $\mathsf{S}_I \times \mathsf{S}_J$. Hence $\mathbb{R}^{I \times J}_{R \leqslant 1}$ corresponds to a double cone built on $\mathsf{S}_I \times \mathsf{S}_J$.

Let us focus on the case of symmetric (2×2) matrices, which form a vector space of dimension 3, and hence allow for a visual representation (see Figure 2.1). The symmetric positive semidefinite unit-norm rank-1 matrices form a circle. Reflection around the origin yields a second circle, corresponding to the symmetric negative semidefinite unit-norm rank-1 matrices. Arbitrary symmetric rank-1 matrices are obtained by scaling, i.e., they form a double cone built on the two circles. It is now clear that, with probability one, three arbitrarily chosen points on the double cone are not confined to a common two-dimensional plane. This is equivalent to saying that the rank of $\mathbf{A} \odot \mathbf{A}$ for $\mathbf{A} \in \mathbb{R}^{2 \times 3}$ is generically equal to 3, since the columns $A_r \otimes A_r$ of $\mathbf{A} \odot \mathbf{A}$ can be interpreted as a vector representation of the rank-1 matrices $A_r A_r^T$.

The situation for $\mathbb{R}^{I \times J}_{R \leqslant 1}$ is completely similar. Randomly sampling points on the double cone yields a set that is maximally linearly independent. This has been formalized in Lemma 2.2.

We now have the following two lemmas.

Lemma 2.3. $\mathfrak{V} = \{V_m | 1 \leqslant m \leqslant M\}$ $\mathbb{R}^{N^2}$ W_1 … W_R $\mathbb{R}^N$ $\mathfrak{W}_R = \{W_p \otimes W_q | 1 \leqslant p < q \leqslant R\}$

$$R \leqslant N + 1 \qquad M + \frac{R(R-1)}{2} \leqslant N^2, \tag{2.20}$$

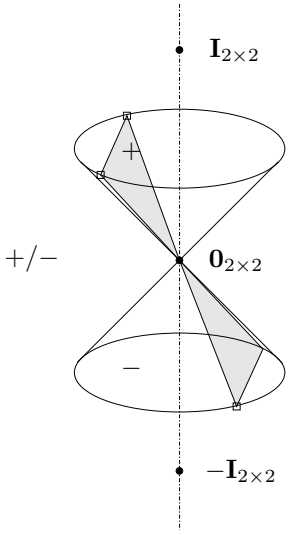


FIG. 2.1. *Visualization of the vector space of symmetric* (2×2) *matrices. The double cone is formed by the rank-*1 *matrices. The upper cone contains the positive definite matrices. The lower cone contains the negative definite matrices. The surrounding space contains the indefinite matrices. Taking at random three points on the double cone yields a linearly independent set. The three points indicated by a little square belong to the same subspace, represented by the dashed plane. After a (generic) small displacement of these points on the double cone, they are no longer constrained to a two-dimensional subspace.*

$\mathfrak{V} \cup \mathfrak{W}_R$ W_r $1 \leqslant r \leqslant R$

. Let $\mathfrak{W}_{R-1} = \{W_p \otimes W_q | 1 \leqslant p < q \leqslant R-1\}$. The proof is by induction. We first show that the lemma holds for $M \leqslant N^2 - 1$ and $R = 2$. Then we show that, assuming that the lemma holds for $(M, R-1)$, it still holds for (M, R) if (2.20) is satisfied.

Let $V^\perp \in \mathbb{R}^{N^2}$ be orthogonal to the vectors in $\mathfrak{V}$. To initialize the induction, it suffices to show that $W_1 \otimes W_2$ generically has a component in the direction of $V^\perp$. Define a matrix $\mathbf{V}^\perp \in \mathbb{R}^{N\times N}$ by $(\mathbf{V}^\perp)_{n_1 n_2} = (V^\perp)_{(n_1-1)N+n_2}$. Then we have $(W_1 \otimes W_2)^T V^\perp = W_1^T \mathbf{V}^\perp W_2$, which is indeed generically different from zero.

Now we prove the induction step. The matrices $[W_1 \dots W_{R-1}], [W_2 \dots W_R] \in \mathbb{R}^{N\times R}$ are generically full column rank if $R \leqslant N+1$. By a property of the Kronecker product, $[W_1 \dots W_{R-1}] \otimes [W_2 \dots W_R]$ is also full column rank. The set $\mathfrak{W}_R$, consisting of columns of the latter matrix, is thus linearly independent. Now suppose that the set $\mathfrak{V} \cup \mathfrak{W}_R$ is linearly dependent. We prove that the set becomes linearly independent by a generic perturbation of the vector W_R. We prove this by contradiction. Let W_R be replaced by a vector $\tilde{W}_R$ that is not proportional to W_R. The set $\mathfrak{W}_R$ is consistently replaced by $\tilde{\mathfrak{W}}_R$. Suppose that $\mathfrak{V} \cup \tilde{\mathfrak{W}}_R$ is still linearly dependent. Generically, we may assume that V_1 can be written as a linear combination of the vectors in $(\mathfrak{V} \setminus \{V_1\}) \cup \mathfrak{W}_R$. We may also assume that V_1 is a linear combination of the vectors in $(\mathfrak{V} \setminus \{V_1\}) \cup \tilde{\mathfrak{W}}_R$. In other words, V_1 is in the intersection of the subspaces U and $\tilde{\mathsf{U}}$ generated by $(\mathfrak{V} \setminus \{V_1\}) \cup \mathfrak{W}_R$ and $(\mathfrak{V} \setminus \{V_1\}) \cup \tilde{\mathfrak{W}}_R$, respectively. U equals the sum of the subspace generated by $(\mathfrak{V} \setminus \{V_1\}) \cup \mathfrak{W}_{R-1}$ and the subspace generated by $\{W_1 \otimes W_R, \dots, W_{R-1} \otimes W_R\}$. $\tilde{\mathsf{U}}$ equals the sum of the subspace generated by

$(\mathfrak{V} \setminus \{V_1\}) \cup \mathfrak{W}_{R-1}$ and the subspace generated by $\{W_1 \otimes \tilde{W}_R, \ldots, W_{R-1} \otimes \tilde{W}_R\}$. U cannot be equal to $\mathbb{R}^{N^2}$ since $\dim(\mathsf{U}) \leqslant M - 1 + R(R-1)/2 < N^2$; neither can $\tilde{\mathsf{U}}$ be equal to $\mathbb{R}^{N^2}$. Taking into account that the vectors $\{W_1 \otimes W_R, \ldots, W_{R-1} \otimes W_R, W_1 \otimes \tilde{W}_R, \ldots, W_{R-1} \otimes \tilde{W}_R\}$, being the columns of $[W_1 \ldots W_{R-1}] \otimes [W_R\, \tilde{W}_R]$, are linearly independent, we conclude that the intersection of U and $\tilde{\mathsf{U}}$ is equal to the subspace generated by $(\mathfrak{V} \setminus \{V_1\}) \cup \mathfrak{W}_{R-1}$. Since V_1 is in the intersection of U and $\tilde{\mathsf{U}}$, it is a linear combination of the vectors in $(\mathfrak{V} \setminus \{V_1\}) \cup \mathfrak{W}_{R-1}$. This means that the set $\mathfrak{V} \cup \mathfrak{W}_{R-1}$ is linearly dependent, which is in contradiction to the induction hypothesis. □

LEMMA 2.4. $\mathfrak{V} = \{V_m | 1 \leqslant m \leqslant M\}$ $\mathbb{R}^{I^2 J^2}$ $A_1 \ldots A_R$ $\mathbb{R}^I$ $B_1 \ldots B_R$ $\mathbb{R}^J$

$$R \leqslant IJ + 1 \qquad M + \frac{R(R-1)}{2} \leqslant I^2 J^2, \tag{2.21}$$

$\mathfrak{V} \cup \{A_p \otimes B_p \otimes A_q \otimes B_q | 1 \leqslant p < q \leqslant R\}$ A_r B_r $1 \leqslant r \leqslant R$

. The proof is analogous to the proof of Lemma 2.3. The role of $[W_1 \ldots W_{R-1}]$, $[W_2 \ldots W_R]$ is now played by $[A_1 \otimes B_1 \ldots A_{R-1} \otimes B_{R-1}], [A_2 \otimes B_2 \ldots A_R \otimes B_R] \in \mathbb{R}^{IJ \times R}$. The latter matrices are generically full column rank if $R \leqslant IJ + 1$ because of Lemma 2.2. □

We now have the following theorem.

THEOREM 2.5. (2.1) $R \leqslant K$ $R(R-1) \leqslant I(I-1)J(J-1)/2$

. The second inequality implies that $R \leqslant IJ$. According to Lemma 2.2, $\mathbf{A} \odot \mathbf{B}$ is generically full column rank, which is a necessary requirement for (2.1) to be a CANDECOMP (cf. above). We will prove the theorem by checking that the deterministic conditions for uniqueness derived in section 2.1 are generically satisfied. According to the first inequality of the theorem, $\mathbf{C}$ is tall. Hence, it is generically full column rank. We will now show that the second inequality generically guarantees linear independence of $\{\Phi(A_p B_p^T, A_q B_q^T)\}_{p<q}$.

Consider the following bijective mapping of vectors in $\mathbb{R}^{I^2 J^2}$ to tensors in $\mathbb{R}^{I \times I \times J \times J}$:

$$(\mathcal{F}_1(X))_{ijkl} = x_{(i-1)IJ^2 + (j-1)J^2 + (k-1)J + l}.$$

The image vector of $\Phi(A_p B_p^T, A_q B_q^T)$ under the inverse mapping $\mathcal{F}_1^{-1}$ is given by

$$\begin{aligned}
&A_p \otimes A_q \otimes (B_p \otimes B_q - B_q \otimes B_p) \\
&+ A_q \otimes A_p \otimes (B_q \otimes B_p - B_p \otimes B_q) \\
&= (A_p \otimes A_q - A_q \otimes A_p) \otimes (B_p \otimes B_q - B_q \otimes B_p) \\
&= [(\mathbf{I}_{I^2 \times I^2} - \mathbf{P}_{I^2 \times I^2}) \cdot (A_p \otimes A_q)] \otimes [(\mathbf{I}_{J^2 \times J^2} - \mathbf{P}_{J^2 \times J^2}) \cdot (B_p \otimes B_q)] \\
&= [(\mathbf{I}_{I^2 \times I^2} - \mathbf{P}_{I^2 \times I^2}) \otimes (\mathbf{I}_{J^2 \times J^2} - \mathbf{P}_{J^2 \times J^2})] \cdot [A_p \otimes A_q \otimes B_p \otimes B_q] \\
&\stackrel{\text{def}}{=} \mathbf{G} \cdot [A_p \otimes A_q \otimes B_p \otimes B_q].
\end{aligned} \tag{2.22}$$

Linear independence of $\{\Phi(A_p B_p^T, A_q B_q^T)\}_{p<q}$ is equivalent to linear independence of the image vectors. The latter are linearly independent if and only if the intersection of the kernel of $\mathbf{G}$ and the subspace generated by $\{A_p \otimes A_q \otimes B_p \otimes B_q\}_{p<q}$ contains only the null vector. In other words, a basis of the kernel of $\mathbf{G}$ and the vectors $A_p \otimes A_q \otimes B_p \otimes B_q$, $p < q$, have to form a linearly independent set. The dimension

of the kernel of $\mathbf{G}$ is $I^2J^2 - \text{rank}(\mathbf{G})$. According to Lemma 2.4, the set formed by a basis of the kernel and the $R(R-1)/2$ vectors $A_p \otimes A_q \otimes B_p \otimes B_q$ is generically linearly independent if

$$\frac{R(R-1)}{2} \leqslant \text{rank}\,(\mathbf{G})\,. \tag{2.23}$$

We now compute rank($\mathbf{G}$). Consider $Y' \in \mathbb{R}^{I^2}$ and $\mathbf{Y} \in \mathbb{R}^{I \times I}$, linked by $y'_{(i_1-1)I+i_2} = y_{i_1 i_2}$, $i_1, i_2 = 1, \ldots, I$. The matrix $\mathbf{P}_{I^2 \times I^2}$ is such that $\mathbf{P}_{I^2 \times I^2} Y'$ and $\mathbf{Y}^T$ are linked in the same way, i.e., $(\mathbf{P}_{I^2 \times I^2} Y')_{(i_1-1)I+i_2} = y_{i_2 i_1}$. Hence the kernel of $\mathbf{I}_{I^2 \times I^2} - \mathbf{P}_{I^2 \times I^2}$ corresponds to the $I(I+1)/2$-dimensional space of symmetric $(I \times I)$ matrices. Therefore $\text{rank}(\mathbf{I}_{I^2 \times I^2} - \mathbf{P}_{I^2 \times I^2}) = I(I-1)/2$ and $\text{rank}(\mathbf{I}_{J^2 \times J^2} - \mathbf{P}_{J^2 \times J^2}) = J(J-1)/2$. By a property of the Kronecker product we obtain

$$\text{rank}\,(\mathbf{G}) = I(I-1)J(J-1)/4. \tag{2.24}$$

Combining (2.23) and (2.24) yields that the set $\{\Phi(A_p B_p^T, A_q B_q^T)\}_{p<q}$ is generically linearly independent if and only if

$$\frac{R(R-1)}{2} \leqslant \frac{I(I-1)J(J-1)}{4}. \qquad \square$$

3. The fourth-order case.

3.1. Deterministic uniqueness condition and algorithm. Now consider an $(I \times J \times K \times L)$-tensor $\mathcal{T}$ of which the CANDECOMP is given by

$$t_{ijkl} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} d_{lr}, \tag{3.1}$$

in which $\mathbf{A} \in \mathbb{R}^{I \times R}$, $\mathbf{B} \in \mathbb{R}^{J \times R}$, $\mathbf{C} \in \mathbb{R}^{K \times R}$, $\mathbf{D} \in \mathbb{R}^{L \times R}$.

Consider a matrix $\mathbf{T} \in \mathbb{R}^{IJK \times L}$ in which the entries of $\mathcal{T}$ are stacked as follows:

$$(\mathbf{T})_{(i-1)JK+(j-1)K+k,l} = t_{ijkl} \qquad \forall i, j, k, l.$$

We have

$$\mathbf{T} = (\mathbf{A} \odot \mathbf{B} \odot \mathbf{C}) \cdot \mathbf{D}^T. \tag{3.2}$$

We assume that both $\mathbf{A} \odot \mathbf{B} \odot \mathbf{C}$ and $\mathbf{D}$ are full column rank. Both conditions are generically satisfied if $R \leqslant \min(IJK, L)$ (generic properties will be examined in detail in section 3.2). In this case, the rank of $\mathcal{T}$ is equal to the rank of $\mathbf{T}$.

Consider a factorization of $\mathbf{T}$ of the form

$$\mathbf{T} = \mathbf{E} \cdot \mathbf{F}^T, \tag{3.3}$$

with $\mathbf{E} \in \mathbb{R}^{IJK \times R}$ and $\mathbf{F} \in \mathbb{R}^{L \times R}$ full column rank. Because of (3.2) and (3.3), we have

$$\mathbf{A} \odot \mathbf{B} \odot \mathbf{C} = \mathbf{E} \cdot \mathbf{W} \tag{3.4}$$

for some nonsingular $\mathbf{W} \in \mathbb{R}^{R \times R}$. The task is now to find $\mathbf{W}$ such that the columns of $\mathbf{E} \cdot \mathbf{W}$ correspond to third-order rank-1 tensors. Therefore, we will make use of the following third-order variant of Theorem 2.1.

THEOREM 3.1. $\Psi_1 : (\mathcal{X}, \mathcal{Y}) \in \mathbb{R}^{I\times J\times K} \times \mathbb{R}^{I\times J\times K} \to \Psi_1(\mathcal{X}, \mathcal{Y}) \in \mathbb{R}^{I\times I\times J\times J\times K\times K}$ $\Psi_2 : (\mathcal{X}, \mathcal{Y}) \in \mathbb{R}^{I\times J\times K} \times \mathbb{R}^{I\times J\times K} \to \Psi_2(\mathcal{X}, \mathcal{Y}) \in \mathbb{R}^{I\times I\times J\times J\times K\times K}$ $\Psi : (\mathcal{X}, \mathcal{Y}) \in \mathbb{R}^{I\times J\times K} \times \mathbb{R}^{I\times J\times K} \to \Psi(\mathcal{X}, \mathcal{Y}) \in \mathbb{R}^{I\times I\times J\times J\times K\times K\times 2}$

$$(\Psi(\mathcal{X}, \mathcal{Y}))_{ijklmn1} = (\Psi_1(\mathcal{X}, \mathcal{Y}))_{ijklmn}$$
$$= x_{ikm}y_{jln} + y_{ikm}x_{jln} - x_{jkm}y_{iln} - y_{jkm}x_{iln}, \tag{3.5}$$
$$(\Psi(\mathcal{X}, \mathcal{Y}))_{ijklmn2} = (\Psi_2(\mathcal{X}, \mathcal{Y}))_{ijklmn}$$
$$= x_{ikm}y_{jln} + y_{ikm}x_{jln} - x_{ilm}y_{jkn} - y_{ilm}x_{jkn}. \tag{3.6}$$

$\Psi(\mathcal{X}, \mathcal{X}) = 0$ $\mathcal{X}$ 1 . The "if" part is obvious. For the "only if" part, let us first consider the condition

$$x_{ikm}x_{jln} - x_{jkm}x_{iln} = 0, \tag{3.7}$$

following from (3.5). Define a matrix $\mathbf{X}_{(1)} \in \mathbb{R}^{I\times JK}$ by the elementwise equation

$$(\mathbf{X}_{(1)})_{i,(j-1)K+k} = x_{ijk} \qquad \forall i, j, k.$$

The columns of $\mathbf{X}_{(1)}$ correspond to the different mode-1 vectors of $\mathcal{X}$. Equation (3.7) is equivalent to

$$\det\left(\begin{pmatrix} x_{ikm} & x_{iln} \\ x_{jkm} & x_{jln} \end{pmatrix}\right) = 0.$$

Imposing constraint (3.7) for all indices is equivalent to claiming that the determinant of any (2×2) submatrix of $\mathbf{X}_{(1)}$ vanishes. This is satisfied if and only if $\mathbf{X}_{(1)}$ is at most rank 1. In other words, (3.7) holds for all index combinations if and only if all the mode-1 vectors of $\mathcal{X}$ are proportional. Similarly, the condition

$$x_{ikm}x_{jln} - x_{ilm}x_{jkn} = 0, \tag{3.8}$$

following from (3.6), is satisfied for all indices if and only if all the mode-2 vectors are proportional. Consider the matrices $\mathbf{X}_k \in \mathbb{R}^{I\times J}$, $1 \leqslant l \leqslant K$, defined by the elementwise equation $(\mathbf{X}_k)_{ij} = x_{ijk}$. If all mode-1 vectors are proportional to a vector A and if all mode-2 vectors are proportional to a vector B, then all matrices $\mathbf{X}_k$ are proportional to AB^T:

$$\mathbf{X}_k = c_k AB^T$$

or

$$x_{ijk} = a_i b_j c_k$$

for all indices. Hence, (3.7) and (3.8) guarantee that $\mathcal{X}$ is at most rank 1, and vice-versa. □

3. One could add a third tensor slice to $\Psi(\mathcal{X}, \mathcal{Y})$, as follows:

$$(\Psi(\mathcal{X}, \mathcal{Y}))_{ijklmn3} = (\Psi_3(\mathcal{X}, \mathcal{Y}))_{ijklmn}$$
$$= x_{ikm}y_{jln} + y_{ikm}x_{jln} - x_{ikn}y_{jlm} - y_{ikn}x_{jlm}.$$

However, as the proof of Theorem 3.1 demonstrates, this brings in no additional information. On the other hand, one may arbitrarily choose which two of the tensor slices $\Psi_1(\mathcal{X},\mathcal{Y})$, $\Psi_2(\mathcal{X},\mathcal{Y})$, $\Psi_3(\mathcal{X},\mathcal{Y})$ are retained. In what follows, we will work with $\Psi(\mathcal{X},\mathcal{Y})$ as defined in Theorem 3.1.

Define tensors $\mathcal{E}_1,\ldots,\mathcal{E}_R \in \mathbb{R}^{I\times J\times K}$ by

$$(\mathcal{E}_r)_{ijk} = e_{(i-1)JK+(j-1)K+k,r} \quad \forall i,j,k,r$$

and let $\mathcal{Q}_{rs} = \Psi(\mathcal{E}_r,\mathcal{E}_s)$. Due to the bilinearity of Ψ, we have

$$\mathcal{Q}_{rs} = \sum_{t,u=1}^{R} (\mathbf{W}^{-1})_{tr}(\mathbf{W}^{-1})_{us}\Psi(A_t\circ B_t\circ C_t, A_u\circ B_u\circ C_u). \tag{3.9}$$

In analogy with section 2.1, we have that linear independence of $\{\Psi(A_t\circ B_t\circ C_t, A_u\circ B_u\circ C_u)\}_{1\leqslant t<u\leqslant R}$ guarantees that any symmetric matrix $\mathbf{M}$ of which the entries satisfy the following set of homogeneous linear equations

$$\sum_{s,t=1}^{R} m_{rs}\mathcal{Q}_{rs} = \mathcal{O}, \tag{3.10}$$

can be decomposed as

$$\mathbf{M} = \mathbf{W}\cdot\Lambda\cdot\mathbf{W}^T, \tag{3.11}$$

in which Λ is diagonal. Equation (3.10) has R linearly independent solutions, which lead to a simultaneous matrix diagonalization as in (2.13), from which $\mathbf{W}$ can be obtained. Once $\mathbf{W}$ is known, $\mathbf{A}\odot\mathbf{B}\odot\mathbf{C}$ can be obtained from (3.4). On the other hand, from (3.2), (3.3), and (3.4) we have

$$\mathbf{D} = \mathbf{F}\cdot\mathbf{W}^{-T}. \tag{3.12}$$

We conclude that the CANDECOMP in (3.1) is unique if $\mathbf{D}$ is full column rank and if the tensors $\{\Psi(A_t\circ B_t\circ C_t, A_u\circ B_u\circ C_u)\}_{1\leqslant t<u\leqslant R}$ are linearly independent. In that case, the canonical components may be computed using Algorithm 3.1. Similar comments to those regarding Algorithm 2.1 are in order. With respect to step 10, we mention that A_r, B_r, C_r are obtained from the best rank-1 approximation of $\mathcal{G}_r$.

3.2. Generic uniqueness condition. In this section, we check under which conditions on R both $\{D_r\}_{1\leqslant r\leqslant R}$ and $\{\Psi(A_t\circ B_t\circ C_t, A_u\circ B_u\circ C_u)\}_{1\leqslant t<u\leqslant R}$ are generically linearly independent. Under these conditions, a generic tensor has a unique CANDECOMP, the components of which can be computed by means of Algorithm 3.1.

We have the following theorem.

THEOREM 3.2. [illegible] (3.1) [illegible] $R \leqslant L$ [illegible] $R(R-1) \leqslant IJK(3IJK - IJ - IK - JK - I - J - K + 3)/4$ [illegible]. In analogy with the proof of Theorem 2.5, we have that $\mathbf{A}\odot\mathbf{B}\odot\mathbf{C}$ and $\mathbf{D}$ are generically full column rank. We will now show that the second inequality of the theorem generically guarantees linear independence of $\{\Psi(A_p\circ B_p\circ C_p, A_q\circ B_q\circ C_q)\}_{p<q}$.

Consider the following mapping of vectors in $\mathbb{R}^{I^2J^2K^2}$ to tensors in $\mathbb{R}^{I\times I\times J\times J\times K\times K}$:

$$(\mathcal{F}_1(X))_{ijklmn} = x_{(i-1)IJ^2K^2+(j-1)J^2K^2+(k-1)JK^2+(l-1)K^2+(m-1)K+n}.$$

ALGORITHM 3.1

$\mathcal{T} \in \mathbb{R}^{I\times J\times K\times L}$ satisfying

$$\mathcal{T} = \sum_{r=1}^{R} A_r \circ B_r \circ C_r \circ D_r,$$

with both $\{D_r\}_{1\leqslant r\leqslant R}$ and $\{\Psi(A_t \circ B_t \circ C_t, A_u \circ B_u \circ C_u)\}_{1\leqslant t<u\leqslant R}$ linearly independent.

rank R and CANDECOMP factor matrices $\mathbf{A} \in \mathbb{R}^{I\times R}$, $\mathbf{B} \in \mathbb{R}^{J\times R}$, $\mathbf{C} \in \mathbb{R}^{K\times R}$, $\mathbf{D} \in \mathbb{R}^{L\times R}$.

1. Stack $\mathcal{T}$ in $\mathbf{T} \in \mathbb{R}^{IJK\times L}$ as follows:

$$(\mathbf{T})_{(i-1)JK+(j-1)K+k,l} = (\mathcal{T})_{ijkl} \qquad \forall i,j,k,l.$$

2. $R = \operatorname{rank}(\mathbf{T})$.
3. Compute factorization

$$\mathbf{T} = \mathbf{E}\cdot\mathbf{F}^T,$$

with $\mathbf{E} \in \mathbb{R}^{IJK\times R}$ and $\mathbf{F} \in \mathbb{R}^{L\times R}$ full column rank.

4. Stack $\mathbf{E}$ in $\mathcal{E} \in \mathbb{R}^{I\times J\times K\times R}$ as follows:

$$(\mathcal{E})_{ijkr} = (\mathbf{E})_{(i-1)JK+(j-1)K+k,r} \quad \forall i,j,k,r.$$

5. Compute $\mathcal{Q}_{rs} \in \mathbb{R}^{I\times I\times J\times J\times K\times K\times 2}$, $1 \leqslant r,s \leqslant R$, as follows:

$$(\mathcal{Q}_{rs})_{ijklmn1} = e_{ikmr}e_{jlns} + e_{ikms}e_{jlnr} - e_{jkmr}e_{ilns} - e_{jkms}e_{ilnr},$$
$$(\mathcal{Q}_{rs})_{ijklmn2} = e_{ikmr}e_{jlns} + e_{ikms}e_{jlnr} - e_{ilmr}e_{jkns} - e_{ilms}e_{jknr}.$$

6. Compute the kernel of

$$\sum_{s,t=1}^{R} m_{st}\mathcal{Q}_{st} = \mathcal{O}$$

under the constraint $m_{st} = m_{ts}$ $\forall s,t$. Stack R linearly independent solutions in symmetric matrices $\mathbf{M}_1, \ldots, \mathbf{M}_R \in \mathbb{R}^{R\times R}$.

7. Determine $\mathbf{W} \in \mathbb{R}^{R\times R}$ that simultaneously diagonalizes $\mathbf{M}_1, \ldots, \mathbf{M}_R$:

$$\mathbf{M}_1 = \mathbf{W}\cdot\Lambda_1\cdot\mathbf{W}^T$$
$$\vdots$$
$$\mathbf{M}_R = \mathbf{W}\cdot\Lambda_R\cdot\mathbf{W}^T.$$

8. $\mathbf{A}\odot\mathbf{B}\odot\mathbf{C} = \mathbf{E}\cdot\mathbf{W}$ and $\mathbf{D} = \mathbf{F}\cdot\mathbf{W}^{-T}$.
9. Stack $\mathbf{A}\odot\mathbf{B}\odot\mathbf{C}$ in $\mathcal{G}_1, \ldots, \mathcal{G}_R \in \mathbb{R}^{I\times J\times K}$ as follows:

$$(\mathcal{G}_r)_{ijk} = (\mathbf{A}\odot\mathbf{B}\odot\mathbf{C})_{(i-1)JK+(j-1)K+k,r} \qquad \forall i,j,k.$$

10. Obtain A_r, B_r, C_r from

$$\mathcal{G}_r = A_r \circ B_r \circ C_r \qquad \forall r.$$

The image vector of $\Psi_1(A_p \circ B_p \circ C_p, A_q \circ B_q \circ C_q)$ under the inverse mapping $\mathcal{F}_1^{-1}$ is given by

$$\begin{aligned}
&(A_p \otimes A_q - A_q \otimes A_p) \otimes (B_p \otimes B_q \otimes C_p \otimes C_q - B_q \otimes B_p \otimes C_q \otimes C_p) \\
&= [(\mathbf{I}_{I^2 \times I^2} - \mathbf{P}_{I^2 \times I^2}) \cdot (A_p \otimes A_q)] \\
&\quad \otimes [(\mathbf{I}_{J^2K^2 \times J^2K^2} - \mathbf{P}_{J^2 \times J^2} \otimes \mathbf{P}_{K^2 \times K^2}) \cdot (B_p \otimes B_q \otimes C_p \otimes C_q)] \\
&= [(\mathbf{I}_{I^2 \times I^2} - \mathbf{P}_{I^2 \times I^2}) \otimes (\mathbf{I}_{J^2K^2 \times J^2K^2} - \mathbf{P}_{J^2 \times J^2} \otimes \mathbf{P}_{K^2 \times K^2})] \\
&\quad \cdot [A_p \otimes A_q \otimes B_p \otimes B_q \otimes C_p \otimes C_q] \\
&\stackrel{\text{def}}{=} \mathbf{G}_1 \cdot [A_p \otimes A_q \otimes B_p \otimes B_q \otimes C_p \otimes C_q]. \qquad (3.13)
\end{aligned}$$

Similarly, the image vector of $\Psi_2(A_p \circ B_p \circ C_p, A_q \circ B_q \circ C_q)$ under $\mathcal{F}_1^{-1}$ is given by

$$\begin{aligned}
&[(\mathbf{I}_{I^2K^2 \times I^2K^2} - \mathbf{P}_{I^2 \times I^2} \otimes \mathbf{P}_{K^2 \times K^2}) \otimes (\mathbf{I}_{J^2 \times J^2} - \mathbf{P}_{J^2 \times J^2})] \\
&\cdot [\mathbf{I}_{I^2 \times I^2} \otimes \mathbf{P}_{K^2.J^2 \times J^2.K^2}] \cdot [A_p \otimes A_q \otimes B_p \otimes B_q \otimes C_p \otimes C_q] \\
&\stackrel{\text{def}}{=} \mathbf{G}_2 \cdot [A_p \otimes A_q \otimes B_p \otimes B_q \otimes C_p \otimes C_q]. \qquad (3.14)
\end{aligned}$$

Define $\mathbf{G} = (\mathbf{G}_1^T \; \mathbf{G}_2^T)^T$. The tensors $\{\Psi(A_p \circ B_p \circ C_p, A_q \circ B_q \circ C_q)\}_{p<q}$ are linearly independent if and only if a basis of the kernel of $\mathbf{G}$ and the vectors $A_p \otimes A_q \otimes B_p \otimes B_q \otimes C_p \otimes C_q$, $p < q$, form a linearly independent set. By reasoning as in the proofs of Lemma 2.4 and Theorem 2.5, we obtain that this is generically guaranteed if

$$R(R-1)/2 \leqslant \text{rank}(\mathbf{G}). \qquad (3.15)$$

We will now compute $\text{rank}(\mathbf{G})$. Define

$$\begin{aligned}
\mathsf{K}_{1,1} &= \{\mathcal{H} \in \mathbb{R}^{I \times I \times J \times J \times K \times K} | h_{ijklmn} = h_{jiklmn}\}, \\
\mathsf{K}_{1,2} &= \{\mathcal{H} \in \mathbb{R}^{I \times I \times J \times J \times K \times K} | h_{ijklmn} = h_{ijlknm}\}, \\
\mathsf{K}_{2,1} &= \{\mathcal{H} \in \mathbb{R}^{I \times I \times J \times J \times K \times K} | h_{ijklmn} = h_{ijlkmn}\}, \\
\mathsf{K}_{2,2} &= \{\mathcal{H} \in \mathbb{R}^{I \times I \times J \times J \times K \times K} | h_{ijklmn} = h_{jiklnm}\}.
\end{aligned}$$

$\mathsf{K}_1 = \mathsf{K}_{1,1} \cap \mathsf{K}_{1,2}$, $\mathsf{K}_2 = \mathsf{K}_{2,1} \cap \mathsf{K}_{2,2}$, and $\mathsf{K} = \mathsf{K}_1 \cap \mathsf{K}_2$ are the kernels of $\mathbf{G}_1$, $\mathbf{G}_2$, and $\mathbf{G}$, respectively. We have

$$\text{rank}\,(\mathbf{G}) = I^2J^2K^2 - \dim(\mathsf{K}_1 \cap \mathsf{K}_2), \qquad (3.16)$$

$$\dim(\mathsf{K}_1 \cap \mathsf{K}_2) = \dim(\mathsf{K}_1) + \dim(\mathsf{K}_2) - \dim(\mathsf{K}_1 + \mathsf{K}_2), \qquad (3.17)$$

$$\dim(\mathsf{K}_1) = \dim(\mathsf{K}_{1,1}) + \dim(\mathsf{K}_{1,2}) - \dim(\mathsf{K}_{1,1} + \mathsf{K}_{1,2}), \qquad (3.18)$$

$$\dim(\mathsf{K}_2) = \dim(\mathsf{K}_{2,1}) + \dim(\mathsf{K}_{2,2}) - \dim(\mathsf{K}_{2,1} + \mathsf{K}_{2,2}), \qquad (3.19)$$

$$\begin{aligned}
\dim(\mathsf{K}_1 + \mathsf{K}_2) = &\dim(\mathsf{K}_{1,1} + \mathsf{K}_{1,2} + \mathsf{K}_{2,1} + \mathsf{K}_{2,2}) \\
&+ \dim(\mathsf{K}_{1,1}) + \dim(\mathsf{K}_{1,2}) + \dim(\mathsf{K}_{2,1}) + \dim(\mathsf{K}_{2,2}) \\
&- \dim(\mathsf{K}_{1,1} \cap \mathsf{K}_{1,2}) - \dim(\mathsf{K}_{1,1} \cap \mathsf{K}_{2,1}) - \dim(\mathsf{K}_{1,1} \cap \mathsf{K}_{2,2}) \\
&- \dim(\mathsf{K}_{1,2} \cap \mathsf{K}_{2,1}) - \dim(\mathsf{K}_{1,2} \cap \mathsf{K}_{2,2}) - \dim(\mathsf{K}_{2,1} \cap \mathsf{K}_{2,2}) \\
&+ \dim(\mathsf{K}_{1,2} \cap \mathsf{K}_{2,1} \cap \mathsf{K}_{2,2}) + \dim(\mathsf{K}_{1,1} \cap \mathsf{K}_{2,1} \cap \mathsf{K}_{2,2}) \\
&+ \dim(\mathsf{K}_{1,1} \cap \mathsf{K}_{1,2} \cap \mathsf{K}_{2,2}) + \dim(\mathsf{K}_{1,1} \cap \mathsf{K}_{1,2} \cap \mathsf{K}_{2,1}) \\
&- \dim(\mathsf{K}_{1,1} \cap \mathsf{K}_{1,2} \cap \mathsf{K}_{2,1} \cap \mathsf{K}_{2,2}). \qquad (3.20)
\end{aligned}$$

By counting degrees of freedom we obtain

$$\dim(\mathsf{K}_{1,1} \cap \mathsf{K}_{2,1}) = \frac{I(I+1)J(J+1)K^2}{4}, \tag{3.21}$$

$$\dim(\mathsf{K}_{1,1} \cap \mathsf{K}_{2,2}) = \frac{I(I+1)K(K+1)J^2}{4}, \tag{3.22}$$

$$\dim(\mathsf{K}_{1,2} \cap \mathsf{K}_{2,1}) = \frac{J(J+1)K(K+1)I^2}{4}, \tag{3.23}$$

$$\dim(\mathsf{K}_{1,2} \cap \mathsf{K}_{2,2}) = \frac{IJK(IJK+I+J+K)}{4}, \tag{3.24}$$

$$\dim(\mathsf{K}_{1,1} \cap \mathsf{K}_{1,2} \cap \mathsf{K}_{2,1} \cap \mathsf{K}_{2,2}) = \dim(\mathsf{K}_{1,2} \cap \mathsf{K}_{2,1} \cap \mathsf{K}_{2,2}) \tag{3.25}$$

$$= \dim(\mathsf{K}_{1,1} \cap \mathsf{K}_{2,1} \cap \mathsf{K}_{2,2}) \tag{3.26}$$

$$= \dim(\mathsf{K}_{1,1} \cap \mathsf{K}_{1,2} \cap \mathsf{K}_{2,2}) \tag{3.27}$$

$$= \dim(\mathsf{K}_{1,1} \cap \mathsf{K}_{1,2} \cap \mathsf{K}_{2,1}) \tag{3.28}$$

$$= \frac{IJK(I+1)(J+1)(K+1)}{8}. \tag{3.29}$$

The second condition in the theorem follows from combining (3.15)–(3.29). ∎

4. Numerical experiments. In the numerical experiments in this section, tensors are generated in the following way:

$$\tilde{\mathcal{T}} = \frac{\mathcal{T}}{\|\mathcal{T}\|} + \sigma_N \frac{\mathcal{N}}{\|\mathcal{N}\|}, \tag{4.1}$$

in which $\mathcal{T}$ is exactly rank R and can be decomposed as in (1.2). The second term in (4.1) is a noise term. The entries of $\mathcal{N}$ are drawn from a zero-mean unit-variance Gaussian distribution and σ_N controls the noise level.

Monte Carlo experiments consisting of 100 runs are carried out. The canonical components are estimated in three different ways. First, Algorithm 2.1 is applied. The simultaneous matrix diagonalization in step 4 is realized by means of the extended QZ-iteration proposed in [42]. This iteration is initialized by means of the generalized Schur decomposition [21] of $\mathbf{M}_1$ and $\mathbf{M}_2$ in (2.13). Denote subsequent estimates of $\mathbf{Q}$ by $\hat{\mathbf{Q}}_k$ and $\hat{\mathbf{Q}}_{k+1}$. Then the algorithm is stopped when the Frobenius norm of the off-diagonal part of $\hat{\mathbf{Q}}_{k+1}^H \hat{\mathbf{Q}}_k$ drops below $1e-4$. Second, the ALS algorithm described in [6, 9, 38, 41] is applied. It is initialized with 10 different random starting values. Let $\underline{\mathbf{U}}^{(N)}$ be obtained from $\mathbf{U}^{(N)}$ by dividing all columns by their Frobenius norm. The ALS algorithm is stopped when the Frobenius norm of the difference of two subsequent estimates $\underline{\hat{\mathbf{U}}}_k^{(N)}$ and $\underline{\hat{\mathbf{U}}}_{k+1}^{(N)}$ of $\underline{\mathbf{U}}^{(N)}$, optimally ordered and scaled, drops below a certain threshold ϵ_{ALS}; at most 500 iteration steps are carried out. Of the 10 results that are obtained, the best is retained. Third, we used the extended QZ-result to initialize the ALS algorithm.

A condition number for $\mathcal{T}$ is defined as follows:

$$\kappa(\mathcal{A}) = \mathrm{cond}([\lambda_1 U_1^{(1)} \otimes U_1^{(2)} \otimes \cdots \otimes U_1^{(N)}, \ldots, \lambda_R U_R^{(1)} \otimes U_R^{(2)} \otimes \cdots \otimes U_R^{(N)}]).$$

This definition generalizes the standard 2-norm condition number of a matrix, which is obtained by taking λ_r, $U_r^{(1)}$, and $U_r^{(2)}$ equal to the singular values, left singular vectors, and right singular vectors, respectively. The value of $\kappa(\mathcal{A})$ indicates how close the canonical rank-1 components are to an $(R-1)$-dimensional subspace. For

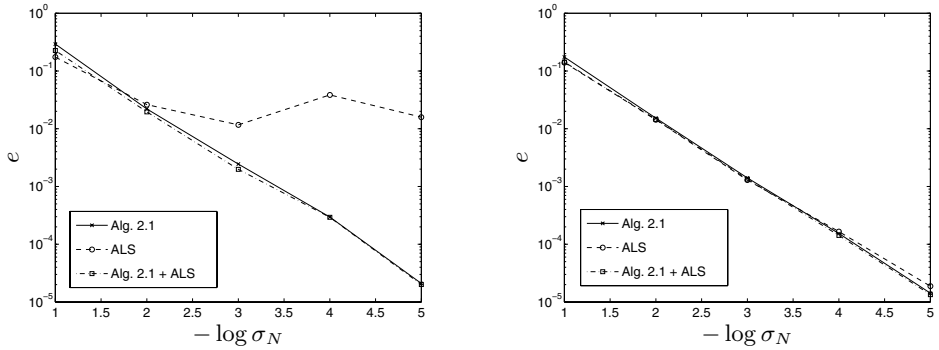


FIG. 4.1. *Relative error obtained in the first experiment* ($R = 4$). *Left: mean. Right: median.*

instance, if two rank-1 terms are close, the condition number will be large. The condition number will also be large if the norm of one of the rank-1 terms is small.

The accuracy is measured in terms of the relative error $e = \|\underline{\mathbf{U}}^{(N)} - \hat{\underline{\mathbf{U}}}^{(N)}\|/\|\underline{\mathbf{U}}^{(N)}\|$, in which $\hat{\underline{\mathbf{U}}}^{(N)}$ is the estimate of $\underline{\mathbf{U}}^{(N)}$, optimally ordered and scaled.

In a first experiment, we consider $\tilde{\mathcal{T}} \in \mathbb{R}^{3\times 4\times 12}$. The entries of $\mathbf{U}^{(1)}$, $\mathbf{U}^{(2)}$, and $\mathbf{U}^{(3)}$ are drawn from a zero-mean unit-variance Gaussian distribution. We look at the effect of varying the noise level σ_N on the error e. We consider two cases: $R = 4$ and $R = 6$. Note that uniqueness of the decomposition in the case $R = 6$ is not covered by Theorem 1.9. In the case $R = 4$, we set $\epsilon_{ALS} = 1e-7$, and in the case $R = 6$, we choose $\epsilon_{ALS} = 1e-8$.

The results for $R = 4$ are plotted in Figure 4.1. The mean ALS accuracy is much worse than the mean accuracy obtained for Algorithm 2.1. To a large extent, this is due to the fact that in a number of cases, in particular those in which $\mathcal{T}$ was ill-conditioned, the ALS algorithm did not find the global optimum or did not converge in 500 steps. The mean and the standard deviations of $\kappa(\mathcal{T})$, over all tensor realizations, were both equal to 5. Exceptionally bad results do not influence the median curves. By choosing the threshold ϵ_{ALS} as small as $1e-7$, a median accuracy similar to that of the extended QZ-iteration was obtained. However, this made the computational cost of the best ALS iteration (out of 10) typically a factor 500 greater than that of Algorithm 2.1. We conclude that Algorithm 2.1 was more robust and less computationally demanding than the ALS algorithm. An additional ALS stage, after the extended QZ-iteration, did not allow for a significant improvement of the accuracy.

The results for $R = 6$ are plotted in Figure 4.2. Clearly, the ALS algorithm did not find the solution. Moreover, the computational cost of the best ALS iteration (out of 10), was typically three orders of magnitude higher than that of Algorithm 2.1. We conclude that this problem was too hard for the ALS approach, while Algorithm 2.1 performed well. An additional ALS stage, after the extended QZ-iteration, allowed us to marginally improve the accuracy.

In a second experiment, we specifically consider the influence of the condition number. Tensors $\tilde{\mathcal{T}} \in \mathbb{R}^{3\times 3\times 9}$ are generated as in (4.1), in which now $\mathbf{U}^{(1)}$, $\mathbf{U}^{(2)}$, $\mathbf{U}^{(3)}$ are given by

$$\mathbf{U}^{(1)} = \mathbf{U}^{(2)} = \begin{pmatrix} 1 & 0 & 0 & 1/\sqrt{3} \\ 0 & 1 & 0 & 1/\sqrt{3} \\ 0 & 0 & 1 & 1/\sqrt{3} \end{pmatrix},$$

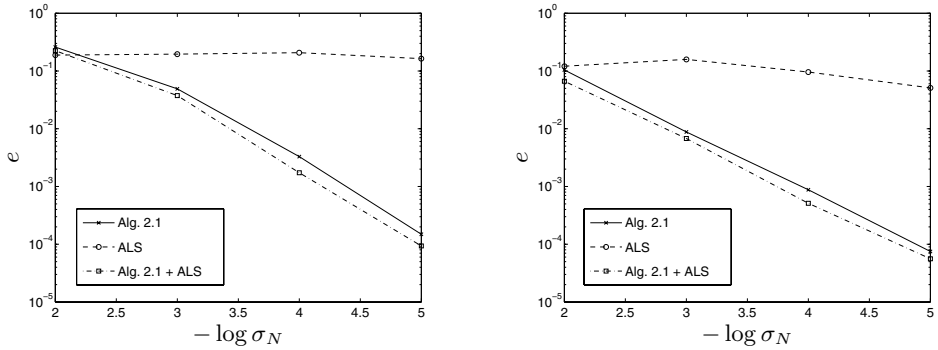


FIG. 4.2. *Relative error obtained in the first experiment* ($R = 6$). *Left: mean. Right: median.*

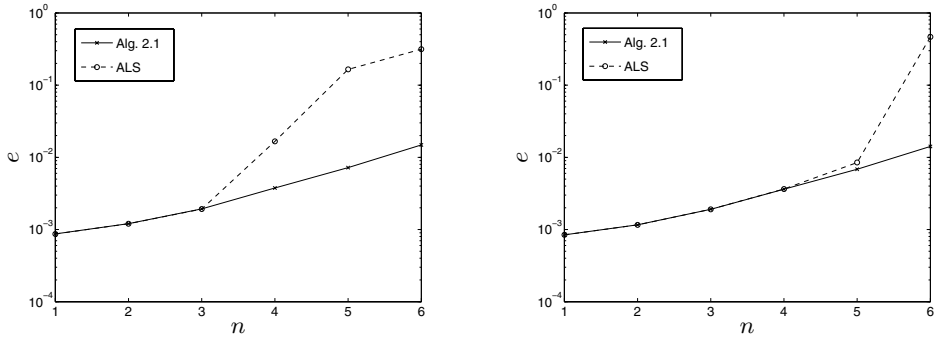


FIG. 4.3. *Relative error obtained in the second experiment. Left: mean. Right: median.*

$$\mathbf{U}^{(3)} = \begin{pmatrix} -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \end{pmatrix}^T.$$

Furthermore, we have $\lambda_1 = \lambda_2 = \lambda_3 = 1$ and $\lambda_4 = 2^{1-n}$. The condition number $\kappa(\mathcal{T})$ is then approximately equal to 2^{n-1}. By varying n between 1 and 6, we make $\kappa(\mathcal{T})$ vary between about 1 and 32. The noise amplitude σ_N is fixed to $1e-3$. We set $\epsilon_{ALS} = 1e-7$; increasing this tolerance decreases the accuracy obtained by the ALS algorithm. The results are shown in Figure 4.3. We see that, for increasing values of $\kappa(\mathcal{T})$, ALS breaks down while Algorithm 2.1 continues to work properly. Moreover, the computational cost of the best ALS iteration (out of 10) was typically a factor 500 greater than that of Algorithm 2.1. An additional ALS stage after the extended QZ-iteration did not improve the accuracy.

In a third experiment we consider fourth-order tensors $\tilde{\mathcal{T}} \in \mathbb{R}^{3\times 2\times 2\times 12}$. The entries of $\mathbf{U}^{(1)}$, $\mathbf{U}^{(2)}$, $\mathbf{U}^{(3)}$, and $\mathbf{U}^{(4)}$ are drawn from a zero-mean unit-variance Gaussian distribution. We look at the effect of varying the noise level σ_N on the error e. We consider the case $R = 4$. We set $\epsilon_{ALS} = 1e-7$; increasing this tolerance decreases the accuracy obtained by the ALS algorithm. The results are shown in Figure 4.4. Similar conclusions can be drawn to those in the first experiment. The computational cost of the best ALS iteration (out of 10) was typically two orders of magnitude greater than that of Algorithm 2.1. An additional ALS stage after the extended QZ-iteration did not improve the accuracy.

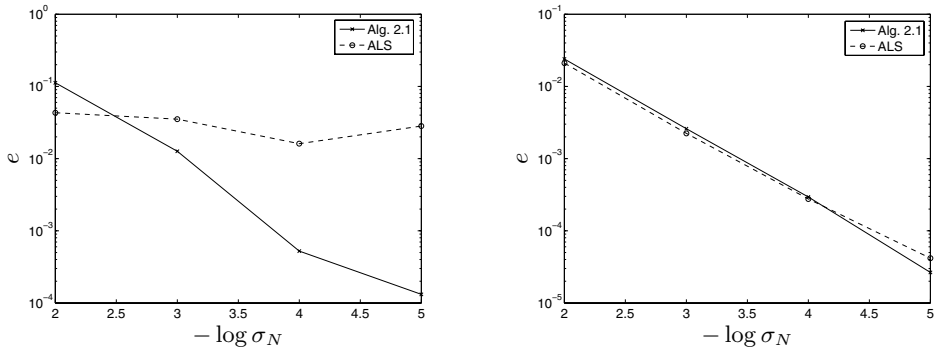


FIG. 4.4. *Relative error obtained in the third experiment. Left: mean. Right: median.*

5. Conclusion. In this paper we have considered the CANDECOMP of higher-order tensors of which at least one dimension is not smaller than the rank. This problem is key to many applications. We have explicitly addressed the case of third- and fourth-order tensors. Along these lines, the approach can be generalized to tensors of arbitrary order.

Under the working assumptions of this paper, the rank of a tensor is equal to the rank of a matrix representation of that tensor. Hence, it does not have to be estimated by means of trial and error.

We have derived a new deterministic condition that guarantees uniqueness of the CANDECOMP. The proof leads to a new algorithm in which the canonical components are obtained from a simultaneous matrix diagonalization by congruence. Numerical experiments showed that this algorithm is superior to the standard ALS algorithm with random initializations, especially when the problem is not well-conditioned or involves a high number of terms.

From the deterministic condition a simple bound on the tensor rank has been derived under which the CANDECOMP is generically unique. Assuming an $(I_1 \times I_2 \times \cdots \times I_N)$-tensor $\mathcal{A}$ of which $\text{rank}(\mathcal{A}) \leqslant I_N$, the bound is roughly proportional to the product of $I_1, \ldots, I_{N-1}$. This is a much weaker constraint than (1.7), in which the bound is up to a constant equal to the sum of $I_1, \ldots, I_{N-1}$.

Acknowledgment. We would like to thank Dr. J. Dehaene (K.U.Leuven) for discussions that helped to improve the presentation of section 2.2.

REFERENCES

[1] C.J. APPELLOF AND E.R. DAVIDSON, *Strategies for analyzing data from video fluoromatric monitoring of liquid chromatographic effluents*, Analytical Chemistry, 53 (1981), pp. 2053–2056.

[2] A. BELOUCHRANI, K. ABED-MERAIM, J.-F. CARDOSO, AND E. MOULINES, *A blind source separation technique using second-order statistics*, IEEE Trans. Signal Process., 45 (1997), pp. 434–444.

[3] A. BELOUCHRANI AND M.G. AMIN, *Blind source separation based on time-frequency signal representations*, IEEE Trans. Signal Process., 46 (1998), pp. 2888–2897.

[4] G. BEYLKIN AND M.J. MOHLENKAMP, *Numerical operator calculus in higher dimensions*, Proc. Natl. Acad. Sci. USA, 99 (2002), pp. 10246–10251.

[5] G. BEYLKIN AND M.J. MOHLENKAMP, *Algorithms for numerical analysis in high dimensions*, SIAM J. Sci. Comput., 26 (2005), pp. 2133–2159.

[6] R. BRO, *PARAFAC. Tutorial & applications*, Chemom. Intell. Lab. Syst., Special Issue 2nd Internet Conf. in Chemometrics (INCINC'96), 38 (1997), pp. 149–171.

[7] J.-F. Cardoso, *Super-symmetric decomposition of the fourth-order cumulant tensor. Blind identification of more sources than sensors*, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Toronto, Canada, 1991, pp. 3109–3112.

[8] J.-F. Cardoso and A. Souloumiac, *Blind beamforming for non-Gaussian signals*, IEE Proc.-F, 140 (1994), pp. 362–370.

[9] J. Carroll and J. Chang, *Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition*, Psychometrika, 9 (1970), pp. 267–283.

[10] P. Comon, *Independent component analysis, a new concept?*, Signal Process., Special Issue on Higher Order Statistics, 36 (1994), pp. 287–314.

[11] P. Comon and B. Mourrain, *Decomposition of quantics in sums of powers of linear forms*, Signal Process., Special Issue on Higher Order Statistics, 53 (1996), pp. 93–108.

[12] J. Dehaene, *Continuous-Time Matrix Algorithms, Systolic Algorithms and Adaptive Neural Networks*, Ph.D. Thesis, Electrical Engineering Department (ESAT), K.U.Leuven, Belgium, 1995.

[13] 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.

[14] L. De Lathauwer, B. De Moor, and J. Vandewalle, *An introduction to independent component analysis*, J. Chemometrics, Special Issue Multi-way Analysis, 14 (2000), pp. 123–149.

[15] L. De Lathauwer, *The canonical decomposition and blind identification with more inputs than outputs: Some algebraic results*, in Proceedings of the Fourth International Symposium on Independent Component Analysis and Blind Signal Separation (ICA 2003), Nara, Japan, 2003, pp. 781–784.

[16] L. De Lathauwer, B. De Moor, and J. Vandewalle, *Computation of the canonical decomposition by means of a simultaneous generalized Schur decomposition*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 295–327.

[17] L. De Lathauwer, *Parallel factor analysis by means of simultaneous matrix decompositions*, in Proceedings of the First IEEE International Workshop on Computational Advances in Multi-sensor Adaptive Processing (CAMSAP 2005), Puerto Vallarta, Jalisco State, Mexico, 2005, pp. 125–128.

[18] L. De Lathauwer, J. Castaing, and J.-F. Cardoso, *Fourth-Order Cumulant Based Under determined Independent Component Analysis*, Tech. Rep., Lab. ETIS, Cergy-Pontoise, France, 2006, submitted.

[19] B. Flury, *Common Principal Components & Related Multivariate Models*, John Wiley, New York, 1988.

[20] A. Franc, *Etude Algébrique des Multitableaux: Apports de l'Algèbre Tensorielle*, Thèse de Doctorat, Spécialité Statistiques, Univ. de Montpellier II, Montpellier, France, 1992.

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

[22] T. Jiang and N.D. Sidiropoulos, *Kruskal's permutation lemma and the identification of CANDECOMP/PARAFAC and bilinear models with constant modulus constraints*, IEEE Trans. Signal Process., 52 (2004), pp. 2625–2636.

[23] M. Haardt and J. A. Nossek, *Simultaneous Schur decomposition of several non-symmetric matrices to achieve automatic pairing in multidimensional harmonic retrieval problems*, IEEE Trans. Signal Process., 46 (1998), pp. 161–169.

[24] R. Harshman, *Foundations of the PARAFAC procedure: Model and conditions for an "explanatory" multi-mode factor analysis*, UCLA Working Papers in Phonetics, 16 (1970), pp. 1–84.

[25] R.D. Hill and S.R. Waters, *On the cone of positive semidefinite matrices*, Linear Algebra Appl., 90 (1987), pp. 81–88.

[26] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley, New York, 2001.

[27] H. Kiers, *Towards a standardized notation and terminology in multiway analysis*, J. Chemometrics, 14 (2000), pp. 105–122.

[28] E. Kofidis and P.A. Regalia, *On the best rank-1 approximation of higher-order supersymmetric tensors*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 863–884.

[29] T. G. Kolda, *Orthogonal tensor decompositions*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 243–255.

[30] T. Kolda, *A counterexample to the possibility of an extension of the Eckart–Young low-rank approximation theorem for the orthogonal rank tensor decomposition*, SIAM J. Matrix

Anal. Appl., 24 (2003), pp. 762–767.
[31] J.B. Kruskal, *Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics*, Linear Algebra Appl., 18 (1977), pp. 95–138.
[32] S.E. Leurgans, R.T. Ross, and R.B. Abel, *A decomposition for three-way arrays*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1064–1083.
[33] X. Liu and N. Sidiropoulos, *Cramér-Rao lower bounds for low-rank decomposition of multidimensional arrays*, IEEE Trans. Signal Process., 49 (2001), pp. 2074–2086.
[34] P. Paatero, *The multilinear engine—A table-driven, least squares program for solving multilinear problems, including the n-way parallel factor analysis model*, J. Comput. Graphical Statist., 8 (1999), pp. 854–888.
[35] D.-T. Pham and J.-F. Cardoso, *Blind separation of instantaneous mixtures of non-stationary sources*, IEEE Trans. Signal Process., 49 (2001), pp. 1837–1848.
[36] D.T. Pham, *Joint approximate diagonalization of positive definite Hermitian matrices*, SIAM J. Matrix Anal. Appl., 22 (2001), pp. 1136–1152.
[37] C.R. Rao and S. Mitra, *Generalized Inverse of Matrices and Its Applications*, John Wiley, New York, 1971.
[38] N. Sidiropoulos, G. Giannakis, and R. Bro, *Blind PARAFAC receivers for DS-CANDE-COMPMA systems*, IEEE Trans. Signal Process., 48 (2000), pp. 810–823.
[39] N. Sidiropoulos, R. Bro, and G. Giannakis, *Parallel factor analysis in sensor array processing*, IEEE Trans. Signal Process., 48 (2000), pp. 2377–2388.
[40] N. Sidiropoulos and R. Bro, *On the uniqueness of multilinear decomposition of N-way arrays*, J. Chemometrics, 14 (2000), pp. 229–239.
[41] A. Smilde, R. Bro, and P. Geladi, *Multi-way Analysis. Applications in the Chemical Sciences*, John Wiley, Chichester, UK, 2004.
[42] A.-J. van der Veen and A. Paulraj, *An analytical constant modulus algorithm*, IEEE Trans. Signal Process., 44 (1996), pp. 1136–1155.
[43] A.-J. van der Veen, *Joint diagonalization via subspace fitting techniques*, in Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Salt Lake City, UT, 2001, pp. 2773–2776.
[44] 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.
[45] T. Zhang and G.H. Golub, *Rank-one approximation to high order tensors*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 534–550.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 3, pp. 667–681

GENERALIZED HADAMARD PRODUCT AND THE DERIVATIVES OF SPECTRAL FUNCTIONS*

HRISTO S. SENDOV†

Abstract. Real valued functions, $F(X)$, on a symmetric matrix argument are called spectral if $F(U^TXU) = F(X)$ for every orthogonal matrix U and $X \in \operatorname{dom} F$. We are interested in a description of the higher order derivatives (when they exist) of F with respect to X. Formulae for the gradient and the Hessian of F are given in [A. S. Lewis, *Math. Oper. Res.*, 21 (1996), pp. 576–588] and [A. S. Lewis and H. S. Sendov, *SIAM Matrix Anal. Appl.*, 23 (2001), pp. 368–386]. In this work we present common features of these two formulae that are preserved in the higher order derivatives.

Key words. spectral function, twice differentiable, higher order derivatives, eigenvalues, symmetric function, perturbation theory, multilinear algebra

AMS subject classifications. Primary, 49R50, 47A75; Secondary, 15A18, 15A69

DOI. 10.1137/050623206

1. Introduction.

Spectral functions are real valued functions on a symmetric matrix argument invariant under conjugation by orthogonal matrices. More precisely, $F : S^n \to \mathbb{R}$ is spectral if

$$F(U^TXU) = F(X)$$

for all $X \in \operatorname{dom} F$ and $U \in O^n$—the orthogonal group on $\mathbb{R}^n$. By restricting F to the subspace of diagonal matrices, it is not difficult to see that spectral functions can be represented by the composition

$$F = f \circ \lambda,$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a symmetric function ($f(Px) = f(x)$ for any permutation matrix P and vector x), and $\lambda : S^n \to \mathbb{R}^n$ is the eigenvalue map $\lambda(X) = (\lambda_1(X), \ldots, \lambda_n(X))$—the vector of eigenvalues of X. We assume throughout that

$$\lambda_1(X) \geq \cdots \geq \lambda_n(X).$$

The study of spectral functions generalizes the study of the individual eigenvalues of a symmetric matrix since if we let

$$\begin{aligned} &\phi_k(x) : \mathbb{R}^n \to \mathbb{R}, \\ &\phi_k(x) := \text{ the } k\text{th largest element of } \{x_1, \ldots, x_n\}, \end{aligned}$$

then $\phi_k(x)$ is symmetric and

$$\lambda_k(X) = (\phi_k \circ \lambda)(X).$$

Various differential properties of eigenvalues have been studied for a long time. They find a lot of applications in areas ranging from matrix perturbation theory [17]

*Received by the editors January 24, 2005; accepted for publication (in revised form) by M. L. Overton March 6, 2006; published electronically September 19, 2006. This research was supported by NSERC.

http://www.siam.org/journals/simax/28-3/62320.html

†Department of Mathematics and Statistics, University of Guelph, Guelph, Ontario, N1G 2W1, Canada (hssendov@uoguelph.ca).

and eigenvalue optimization [10], [9] to quantum mechanics [4]. The Taylor directional expansion (when it exists) of the eigenvalues of symmetric matrices depending on one scalar parameter is described in the monograph by Kato [3]. This naturally raises questions about the differentiability properties of the more general spectral functions. Many such questions have already been investigated in the literature, and the answers to most of them follow the same pattern: $f \circ \lambda$ has a property at the matrix X if and only if f has the same property at the vector $\lambda(X)$. In this way, properties of the function $f \circ \lambda$ on S^n are reduced to properties of the simpler function f on $\mathbb{R}^n$.

Some of the properties of $f \circ \lambda$ at (or around) a matrix X that hold if and only if f has the same property at (or around) the vector $\lambda(X)$ are as follows:

(i) F is lower semicontinuous at X if and only if f is at $\lambda(X)$ [6].
(ii) F is lower semicontinuous and convex if and only if f is [2], [6].
(iii) The symmetric function corresponding to the Fenchel conjugate of F is the Fenchel conjugate of f [14], [6]. (A similar statement holds for the recession function of F [14].)
(iv) F is pointed, has good asymptotic behavior, or is a barrier function on the set $\lambda^{-1}(C)$ if and only if f is such on C [14].
(v) F is Lipschitz around X if and only if f is such around $\lambda(X)$ [7].
(vi) F is (continuously) differentiable at X if and only if f is at $\lambda(X)$ [7].
(vii) F is strictly differentiable at X if and only if f is at $\lambda(X)$ [7], [8].
(viii) $\nabla(f \circ \lambda)$ is semismooth at X if and only if ∇f is at $\lambda(X)$ [13].
(ix) If f is lower semicontinuous and convex, then F is twice epidifferentiable at X relative to Ω if and only if f is twice epidifferentiable at $\lambda(X)$ relative to $\lambda(\Omega)$ [18], where Ω is an arbitrary epigradient.
(x) F has a quadratic expansion at X if and only if f has a quadratic expansion at $\lambda(X)$ [12].
(xi) F is twice (continuously) differentiable at X if and only if f is twice (continuously) differentiable at $\lambda(X)$ [11].
(xii) $F \in \mathcal{C}^\infty$ at $X \Leftrightarrow f \in \mathcal{C}^\infty$ at $\lambda(X)$ [1].
(xiii) F is analytic at X if and only if f is at $\lambda(X)$ [19].
(xiv) F is a polynomial of the entries of X if and only if f is a polynomial. This is a consequence of the Chevalley restriction theorem [20, p. 143].

There are of course exceptions to that pattern. For example, f being directionally differentiable at $\lambda(X)$ does not imply that $f \circ \lambda$ is such at X; see [7].

Formulae for the gradient and the Hessian of the spectral function F given in terms of the derivatives of the symmetric function f were derived in [7] and [11]. In order to reproduce them here we need a bit more notation. For any vector x in $\mathbb{R}^n$, denote by $\operatorname{Diag} x$ the diagonal matrix with vector x on the main diagonal, and denote by $\operatorname{diag}\colon M^n \to \mathbb{R}^n$ its dual operator defined by $\operatorname{diag}(X) = (x_{11}, \ldots, x_{nn})$. Recall that the Hadamard product of two matrices $A = [A^{ij}]$ and $B = [B^{ij}]$ of the same dimension is the matrix $A \circ B = [A^{ij} B^{ij}]$. Thus we have

$$\nabla(f \circ \lambda)(X) = V\big(\operatorname{Diag} \nabla f(\lambda(X))\big)V^T, \tag{1}$$

$$\nabla^2(f \circ \lambda)(X)[H_1, H_2] = \nabla^2 f(\lambda(X))[\operatorname{diag} \tilde{H}_1, \operatorname{diag} \tilde{H}_2] + \langle \mathcal{A}(\lambda(X)), \tilde{H}_1 \circ \tilde{H}_2 \rangle, \tag{2}$$

where V is any orthogonal matrix such that $X = V\big(\operatorname{Diag} \lambda(X)\big)V^T$ is the spectral decomposition of X; $\tilde{H}_i = V^T H_i V$ for $i = 1, 2$, and $x \in \mathbb{R}^n \to \mathcal{A}(x)$ is a matrix valued map that is continuous if $\nabla^2 f(x)$ is.

In [11] a conjecture was made that F is k-times (continuously) differentiable at X if and only if f is such at $\lambda(X)$. When that happens, a natural issue is to find a

practical description of the kth derivative of F and an efficient way to compute it. In addition, explicit formulae for the first kth derivatives of F generalize the terms in the kth order Taylor directional expansion (when it exists) of the individual eigenvalues, given in [3].

This work aims to generalize some common features in formulae (1) and (2) that are preserved in the higher order derivatives of $f \circ \lambda$. The language we present simplifies the description of the higher order derivatives of spectral functions and offers a systematic way for evaluating them when those derivatives are viewed as multilinear functions on the space of symmetric matrices. In section 2 we introduce a multilinear map on the space of square matrices, which generalizes the Hadamard product between two matrices. In section 4 we present its multilinear dual operator, which generalizes the Diag operator. The connections with the derivatives of spectral functions are indicated throughout.

The current paper is the first of three. In [15] we formulate calculus-type rules for the interaction between that generalization of the Hadamard product and the eigenvalues of symmetric matrices. Then in [16] we describe how to compute the higher order derivatives of spectral functions in two general cases. For example, we show that Conjecture 4.1 holds for the derivatives of any function (not necessarily symmetric) of the eigenvalues of symmetric matrices at a matrix X with distinct eigenvalues. And second, we show that it holds for the derivatives of separable spectral functions at an arbitrary symmetric matrix X. (Separable spectral functions are those arising from symmetric functions $f(x) = g(x_1) + \cdots + g(x_n)$ for some function g on a scalar argument.) The computation of the maps $\mathcal{A}_\sigma(x)$ (see (16) below) in these two cases is particularly simple.

2. Generalizations of the Hadamard product. By $\{H_{pq} : 1 \le p, q \le n\}$ we denote the standard basis of the space M^n of all $n \times n$ real matrices. That is, the matrices H_{pq} are such that $(H_{pq})^{ij}$ is 1 if $(i, j) = (p, q)$, and 0 otherwise.

The Hadamard product, $H_1 \circ H_2$, between two matrices H_1 and H_2 from M^n is a matrix valued function on two matrix arguments, linear in each argument separately. Thus, it is uniquely determined by its values on the pairs of basic matrices $(H_{p_1q_1}, H_{p_2q_2})$. On such basic pairs the Hadamard product is defined as

$$(H_{p_1q_1} \circ H_{p_2q_2})^{ij} = \begin{cases} 1 & \text{if } i = p_1 = p_2 \text{ and } j = q_1 = q_2, \\ 0 & \text{otherwise.} \end{cases}$$

An analogous object is obtained if a[illegible] is defined as follows:

$$(H_{p_1q_1} \circ_{(12)} H_{p_2q_2})^{ij} := \begin{cases} 1 & \text{if } i = p_1 = q_2 \text{ and } j = p_2 = q_1, \\ 0 & \text{otherwise,} \end{cases}$$

and then extended to a bilinear function on $M^n \times M^n$. The Hadamard product and the cross Hadamard product are essentially the same:

$$H_{p_1q_1} \circ_{(12)} H_{p_2q_2} = H_{p_1q_1} \circ H^T_{p_2q_2} = H_{p_1q_1} \circ H_{q_2p_2}.$$

These observations can be generalized in the following way. Denote by $\mathbb{N}$ the set of all natural numbers and by $\mathbb{N}_k$ the set $\{1, 2, \ldots, k\}$. A k[illegible] $\mathbb{R}^n$ is a real valued map on $\mathbb{R}^n \times \cdots \times \mathbb{R}^n$ (k-times) linear in each argument separately. When a basis in $\mathbb{R}^n$ is fixed, a k-tensor can be viewed as an $n \times \cdots \times n$ (k-times) "block" of numbers. We index the elements of a tensor in a similar way to the entries of a matrix; thus by $T^{i_1 \ldots i_k}$ we denote the $(i_1, \ldots, i_k)$th entry of T. The space of all k-tensors on

$\mathbb{R}^n$ will be denoted by $T^{k,n}$. The set of all permutations on $\mathbb{N}_k$ as well as the set of all $n \times n$ permutation matrices will be denoted by P^k. (It will be clear from the context which one we mean.)

DEFINITION 2.1. [illegible] σ [illegible] $\mathbb{N}_k$ [illegible] σ [illegible] k [illegible] $H_{p_1q_1}\ H_{p_2q_2}, \ldots, H_{p_kq_k}$ [illegible] k [illegible] $\mathbb{R}^n$ [illegible]

$$(H_{p_1q_1} \circ_\sigma H_{p_2q_2} \circ_\sigma \cdots \circ_\sigma H_{p_kq_k})^{i_1i_2\ldots i_k} = \begin{cases} 1 & \text{[illegible] } i_s = p_s = q_{\sigma(s)}\ \forall s = 1, \ldots, k, \\ 0 & \text{[illegible]} \end{cases}$$

[illegible] k [illegible] k [illegible]

Another way to write the above definition is using the Kronecker delta symbol. Recall that δ_{ij} is equal to 1 if $i = j$, and 0 otherwise. Thus,

$$\begin{aligned}(H_{p_1q_1} \circ_\sigma H_{p_2q_2} \circ_\sigma \cdots \circ_\sigma H_{p_kq_k})^{i_1i_2\ldots i_k} &= \delta_{i_1p_1}\delta_{i_1q_{\sigma(1)}} \cdots \delta_{i_kp_k}\delta_{i_kq_{\sigma(k)}} \\ &= \delta_{i_1p_1}\delta_{p_1q_{\sigma(1)}} \cdots \delta_{i_kp_k}\delta_{p_kq_{\sigma(k)}}.\end{aligned} \tag{3}$$

The next lemma gives the formula for the general entry of the σ-Hadamard product between arbitrary matrices.

LEMMA 2.2. [illegible] σ [illegible]

$$\begin{aligned}(H_1 \circ_\sigma H_2 \circ_\sigma \cdots \circ_\sigma H_k)^{i_1i_2\ldots i_k} &= H_1^{i_1i_{\sigma^{-1}(1)}} \cdots H_k^{i_ki_{\sigma^{-1}(k)}} \\ &= H_{\sigma(1)}^{i_{\sigma(1)}i_1} \cdots H_{\sigma(k)}^{i_{\sigma(k)}i_k}.\end{aligned}$$

[illegible] Let σ be a permutation on $\mathbb{N}_k$ and let $H_1, \ldots, H_k$ be arbitrary matrices. Since the product is linear in each argument separately, we compute

$$\begin{aligned}&(H_1 \circ_\sigma H_2 \circ_\sigma \cdots \circ_\sigma H_k)^{i_1i_2\ldots i_k} \\ &\quad = \sum_{p_1,q_1=1}^{n,n} \cdots \sum_{p_k,q_k=1}^{n,n} H_1^{p_1q_1} \cdots H_k^{p_kq_k} (H_{p_1q_1} \circ_\sigma H_{p_2q_2} \circ_\sigma \cdots \circ_\sigma H_{p_kq_k})^{i_1i_2\ldots i_k} \\ &\quad = \sum_{p_1,q_1=1}^{n,n} \cdots \sum_{p_k,q_k=1}^{n,n} H_1^{p_1q_1} \cdots H_k^{p_kq_k} \delta_{i_1p_1}\delta_{i_1q_{\sigma(1)}} \cdots \delta_{i_kp_k}\delta_{i_kq_{\sigma(k)}} \\ &\quad = \sum_{p_1,q_1=1}^{n,n} \cdots \sum_{p_k,q_k=1}^{n,n} H_1^{p_1q_1} \cdots H_k^{p_kq_k} \delta_{i_1p_1}\delta_{i_{\sigma^{-1}(1)}q_1} \cdots \delta_{i_kp_k}\delta_{i_{\sigma^{-1}(k)}q_k} \\ &\quad = H_1^{i_1i_{\sigma^{-1}(1)}} \cdots H_k^{i_ki_{\sigma^{-1}(k)}} \\ &\quad = H_{\sigma(1)}^{i_{\sigma(1)}i_1} \cdots H_{\sigma(k)}^{i_{\sigma(k)}i_k}. \qquad \square\end{aligned}$$

COROLLARY 2.3. [illegible] $k-1$ [illegible]

$$\begin{aligned}&(H_{p_1q_1} \circ_\sigma \cdots \circ_\sigma H_{p_{k-1}q_{k-1}} \circ_\sigma H)^{i_1i_2\ldots i_k} \\ &\quad = \begin{cases} H^{i_ki_k}\Big(\displaystyle\prod_{s=1}^{k-1} \delta_{i_sp_s}\delta_{i_sq_{\sigma(s)}}\Big) & \text{if } k = \sigma^{-1}(k), \\ H^{i_{\sigma(l)}i_l}\big(\delta_{i_lp_l}\delta_{i_kq_{\sigma(k)}}\big)\Big(\displaystyle\prod_{\substack{s=1 \\ s\neq l}}^{k-1} \delta_{i_sp_s}\delta_{i_sq_{\sigma(s)}}\Big) & \text{if } l := \sigma^{-1}(k) \neq k. \end{cases}\end{aligned}$$

. Suppose first that $l := \sigma^{-1}(k) \neq k$. Using the result of the previous lemma we calculate

$$\begin{aligned}(H_{p_1q_1} \circ_\sigma \cdots \circ_\sigma H_{p_{k-1}q_{k-1}} \circ_\sigma H)^{i_1 i_2 \dots i_k} &= H_{p_1q_1}^{i_1 i_{\sigma^{-1}(1)}} \cdots H_{p_{k-1}q_{k-1}}^{i_{k-1} i_{\sigma^{-1}(k-1)}} H^{i_k i_{\sigma^{-1}(k)}} \\ &= \delta_{i_1 p_1} \delta_{i_{\sigma^{-1}(1)} q_1} \cdots \delta_{i_{k-1} p_{k-1}} \delta_{i_{\sigma^{-1}(k-1)} q_{k-1}} H^{i_k i_{\sigma^{-1}(k)}} \\ &= \delta_{i_1 p_1} \delta_{i_1 q_{\sigma(1)}} \cdots \delta_{i_{l-1} p_{l-1}} \delta_{i_{l-1} q_{\sigma(l-1)}} H^{i_{\sigma(l)} i_l} \delta_{i_{l+1} p_{l+1}} \delta_{i_{l+1} q_{\sigma(l+1)}} \\ &\qquad \cdots \delta_{i_{k-1} p_{k-1}} \delta_{i_{k-1} q_{\sigma(k-1)}} \big(\delta_{i_l p_l} \delta_{i_k q_{\sigma(k)}}\big).\end{aligned}$$

The case $l = k$ follows as well. □

The above corollary can be easily modified when the matrix H is in an arbitrary position in the product.

We often represent a permutation by its cycle decomposition. For example, (123)(45) is the permutation on $\mathbb{N}_5$ that maps 1 to 2, 2 to 3, 3 to 1, in addition to 4 to 5 and 5 to 4.

2.4. We already saw that when $k = 2$ and $\sigma = (12)$ the σ-Hadamard product is essentially the ordinary Hadamard product:

$$H_1 \circ_{(12)} H_2 = H_1 \circ H_2^T.$$

If we restrict our attention to the space of symmetric matrices, then the two products coincide. In the case when $\sigma = (1)(2)$ we get

$$H_1 \circ_{(1)(2)} H_2 = (\operatorname{diag} H_1)(\operatorname{diag} H_2)^T.$$

2.5. In the case $k = 1$, there is one permutation, $\sigma = (1)$, on the elements of the set $\mathbb{N}_1$, and the σ-Hadamard product corresponding to it is a vector valued linear map:

$$\begin{aligned}(\circ_\sigma H_{p_1q_1})^{i_1} &= \begin{cases} 1 & \text{if } i_1 = p_1 = q_1, \\ 0 & \text{otherwise} \end{cases} \\ &= (\operatorname{diag} H_{q_1p_1})^{i_1}.\end{aligned}$$

Extending by linearity we get

$$\circ_\sigma H = \operatorname{diag} H.$$

The standard scalar product between any two k-tensors T_1 and T_2 is given by

$$\langle T_1, T_2 \rangle = \sum_{i_1,\dots,i_k=1}^{n,\dots,n} T_1^{i_1 \dots i_k} T_2^{i_1 \dots i_k}.$$

LEMMA 2.6. T k $\mathbb{R}^n$ H M^n $H_{p_1q_1}, \dots, H_{p_{k-1}q_{k-1}}$ M^n σ $\mathbb{N}_k$

(i) If $\sigma^{-1}(k) = k$, then

$$\langle T, H_{p_1q_1} \circ_\sigma \cdots \circ_\sigma H_{p_{k-1}q_{k-1}} \circ_\sigma H \rangle = \Big(\prod_{t=1}^{k-1} \delta_{p_t q_{\sigma(t)}} \Big) \sum_{t=1}^{n} T^{p_1 \dots p_{k-1} t} H^{tt}.$$

(ii) If $\sigma^{-1}(k) = l$, where $l \neq k$, then

$$\langle T, H_{p_1q_1} \circ_\sigma \cdots \circ_\sigma H_{p_{k-1}q_{k-1}} \circ_\sigma H \rangle = \Big(\prod_{\substack{t=1 \\ t \neq l}}^{k-1} \delta_{p_t q_{\sigma(t)}} \Big) T^{p_1 \cdots p_{k-1} q_{\sigma(k)}} H^{q_{\sigma(k)} p_{\sigma^{-1}(k)}}.$$

[illegible]. Using the definitions and observation (3), we calculate

$$\begin{aligned}
&\langle T, H_{p_1q_1} \circ_\sigma H_{p_2q_2} \circ_\sigma \cdots \circ_\sigma H_{p_{k-1}q_{k-1}} \circ_\sigma H \rangle \\
&= \sum_{p_k,q_k=1}^{n,n} H^{p_kq_k} \langle T, H_{p_1q_1} \circ_\sigma H_{p_2q_2} \circ_\sigma \cdots \circ_\sigma H_{p_{k-1}q_{k-1}} \circ_\sigma H_{p_kq_k} \rangle \\
&= \sum_{p_k,q_k=1}^{n,n} H^{p_kq_k} \sum_{i_1,\ldots,i_k=1}^{n,\ldots,n} T^{i_1\ldots i_k} (H_{p_1q_1} \circ_\sigma H_{p_2q_2} \circ_\sigma \cdots \circ_\sigma H_{p_{k-1}q_{k-1}} \circ_\sigma H_{p_kq_k})^{i_1\ldots i_k} \\
&= \sum_{p_k,q_k=1}^{n,n} H^{p_kq_k} \sum_{i_1,\ldots,i_k=1}^{n,\ldots,n} T^{i_1\ldots i_k} \delta_{i_1p_1} \delta_{p_1q_{\sigma(1)}} \cdots \delta_{i_kp_k} \delta_{p_kq_{\sigma(k)}} \\
&= \sum_{p_k,q_k=1}^{n,n} H^{p_kq_k} T^{p_1\ldots p_k} \delta_{p_1q_{\sigma(1)}} \cdots \delta_{p_kq_{\sigma(k)}}.
\end{aligned}$$

The result follows easily by considering the two cases separately. ∎

3. A partial order on P^k and a property of the σ-Hadamard product. Given two permutations σ, μ on $\mathbb{N}_k$, we say that σ [illegible] μ if for every $s \in \mathbb{N}_k$ there is an $r \in \mathbb{N}_k$ such that

$$\{\sigma^l(s) : l = 1, 2, \ldots\} \subseteq \{\mu^l(r) : l = 1, 2, \ldots\},$$

where $\sigma^l(s) = \sigma(\sigma(\cdots(\sigma(s))\cdots)$, l times. Informally, σ refines μ if the elements of every cycle of σ are contained in a cycle of μ; thus, the cycles of σ partition the cycles of μ. If σ refines μ, we denote it by

$$\mu \preceq \sigma.$$

The refinement relationship is a preorder on P^k (it is reflexive and transitive, but not antisymmetric). With respect to this preorder, the identity permutation is the biggest element (that is, bigger than any other element) and every permutation with only one cycle is a smallest element (that is, smaller than any other element).

There is a natural map between the set P^k and the [illegible] of $\mathbb{R}^k$, given as follows:

$$\mathcal{D}(\sigma) = \{x \in \mathbb{R}^k : x_s = x_{\sigma(s)} \ \forall s \in \mathbb{N}_k\}.$$

This map is onto but is not one-to-one since, for example, when $k = 3$, $\mathcal{D}((123)) = \mathcal{D}((132)) = \{x \in \mathbb{R}^3 : x_1 = x_2 = x_3\}$. The image of the identity permutation is $\mathbb{R}^k$. The following relationship helps to visualize the partial order on P^k:

$$\mu \preceq \sigma \ \Leftrightarrow \ \mathcal{D}(\mu) \subseteq \mathcal{D}(\sigma).$$

Given a permutation $\mu \in P^k$ and a tensor $T \in T^{k,n}$, we denote by $P_\mu(T)$ the tensor in $T^{k,n}$ defined by

$$(P_\mu(T))^{i_1\ldots i_k} = \begin{cases} T^{i_1\ldots i_k} & \text{if } i_s = i_{\mu(s)} \ \forall s \in \mathbb{N}_k, \\ 0 & \text{otherwise.} \end{cases}$$

Informally, the tensor $P_\mu(T)$ preserves the entries of T lying on the "subspace" $\mathcal{D}(\mu)$ of T and replaces the rest of the entries with zeros.

Next is the main result of this section. It describes exactly when one can "transfer" diagonal "subspaces" of T between different σ-Hadamard products.

THEOREM 3.1. [illegible] σ_1 σ_2 [illegible] μ [illegible] $\mathbb{N}_k$ [illegible]

$$\langle P_\mu(T), H_1 \circ_{\sigma_1} \cdots \circ_{\sigma_1} H_k \rangle = \langle P_\mu(T), H_1 \circ_{\sigma_2} \cdots \circ_{\sigma_2} H_k \rangle \tag{4}$$

[illegible] $H_1, \ldots, H_k$ [illegible] T [illegible] $T^{k,n}$ [illegible] $\mu \preceq \sigma_2^{-1} \circ \sigma_1$ [illegible]. Since both sides are linear in each of the matrices $H_1, \ldots, H_k$ separately, it is enough to prove the theorem when these matrices are basic. In other words, we show that

$$\langle P_\mu(T), H_{p_1 q_1} \circ_{\sigma_1} \cdots \circ_{\sigma_1} H_{p_k q_k} \rangle = \langle P_\mu(T), H_{p_1 q_1} \circ_{\sigma_2} \cdots \circ_{\sigma_2} H_{p_k q_k} \rangle$$

for any indexes $p_1, \ldots, p_k$, $q_1, \ldots, q_k$ and any $T \in T^{k,n}$ if and only if $\mu \preceq \sigma_2^{-1} \circ \sigma_1$. Direct calculation gives

$$\begin{aligned}
&\langle P_\mu(T), H_{p_1 q_1} \circ_{\sigma_1} \cdots \circ_{\sigma_1} H_{p_k q_k} \rangle \\
&\qquad = \sum_{i_1,\ldots,i_k=1}^{n,\ldots,n} (P_\mu(T))^{i_1 \ldots i_k} (H_{p_1 q_1} \circ_{\sigma_1} \cdots \circ_{\sigma_1} H_{p_k q_k})^{i_1 \ldots i_k} \\
&\qquad = \sum_{i_1,\ldots,i_k=1}^{n,\ldots,n} (P_\mu(T))^{i_1 \ldots i_k} H_{p_1 q_1}^{i_1 i_{\sigma_1^{-1}(1)}} \cdots H_{p_k q_k}^{i_k i_{\sigma_1^{-1}(k)}} \\
&\qquad = \sum_{i_1,\ldots,i_k=1}^{n,\ldots,n} (P_\mu(T))^{i_1 \ldots i_k} \delta_{i_1 p_1} \delta_{i_1 q_{\sigma_1(1)}} \cdots \delta_{i_k p_k} \delta_{i_k q_{\sigma_1(k)}} \\
&\qquad = (P_\mu(T))^{p_1 \ldots p_k} \delta_{p_1 q_{\sigma_1(1)}} \cdots \delta_{p_k q_{\sigma_1(k)}}.
\end{aligned}$$

The last expression is equal to $T^{p_1 \ldots p_k}$ when $p_s = p_{\mu(s)} = q_{\sigma_1(s)}$ for all $s \in \mathbb{N}_k$, and is equal to 0 otherwise.

Analogously, the right-hand side of (4) is

$$\langle P_\mu(T), H_{p_1 q_1} \circ_{\sigma_2} \cdots \circ_{\sigma_2} H_{p_k q_k} \rangle = (P_\mu(T))^{p_1 \ldots p_k} \delta_{p_1 q_{\sigma_2(1)}} \cdots \delta_{p_k q_{\sigma_2(k)}},$$

which is equal to $T^{p_1 \ldots p_k}$ when $p_s = p_{\mu(s)} = q_{\sigma_2(s)}$ for all $s \in \mathbb{N}_k$, and is equal to 0 otherwise.

Suppose that $\mu \preceq \sigma_2^{-1} \circ \sigma_1$. We consider three cases.

If there is an s_0 such that $p_{s_0} \neq p_{\mu(s_0)}$, then both sides of (4) are zero and the equality is trivial.

If $p_s = p_{\mu(s)}$ for all $s \in \mathbb{N}_k$ but for some s_0 we have that $p_{s_0} \neq q_{\sigma_1(s_0)}$, then it is not possible to have $p_s = q_{\sigma_2(s)}$ for all $s \in \mathbb{N}_k$. Indeed, suppose on the contrary that $p_s = q_{\sigma_2(s)}$ for all $s \in \mathbb{N}_k$. Letting $r = \sigma_2(s)$ we get $p_{\sigma_2^{-1}(r)} = q_r$ for every $r \in \mathbb{N}_k$. Therefore $p_{\sigma_2^{-1}(\sigma_1(s))} = q_{\sigma_1(s)}$ for every $s \in \mathbb{N}_k$ and in particular $p_{\sigma_2^{-1}(\sigma_1(s_0))} = q_{\sigma_1(s_0)} \neq p_{s_0}$. But $\mu \preceq \sigma_2^{-1} \circ \sigma_1$ implies that $\sigma_2^{-1}(\sigma_1(s_0))$ and s_0 belong to the same cycle of μ, that is, $\mu^l(s_0) = \sigma_2^{-1}(\sigma_1(s_0))$ for some $l \in \mathbb{N}$. By the assumption in this case we have that $p_{s_0} = p_{\mu^l(s_0)}$ for every l, which is a contradiction. Thus, for some $s_1 \in \mathbb{N}_k$ we have $p_{s_1} \neq q_{\sigma_2(s_1)}$, and again both sides of (4) are equal to zero.

Suppose finally that $p_s = p_{\mu(s)} = q_{\sigma_1(s)}$ for all $s \in \mathbb{N}_k$. Then the left-hand side of (4) is equal to $T^{p_1 \ldots p_k}$. We are done if we show that $p_s = q_{\sigma_2(s)}$ for every $s \in \mathbb{N}_k$.

Suppose this is not true, that is, for some s_0, $p_{s_0} \neq q_{\sigma_2(s_0)}$. Then for $r_0 = \sigma_2(s_0)$ we have $p_{\sigma_2^{-1}(r_0)} \neq q_{r_0}$, and for $s_1 = \sigma_1^{-1}(r_0)$ we have $p_{\sigma_2^{-1}(\sigma_1(s_1))} \neq q_{\sigma_1(s_1)}$. The condition $\mu \preceq \sigma_2^{-1} \circ \sigma_1$ implies that $\sigma_2^{-1}(\sigma_1(s_1))$ and s_1 belong to the same cycle of μ, and we reach a contradiction as in the previous case.

To prove the opposite direction of the theorem, suppose that

$$(P_\mu(T))^{p_1 \cdots p_k} \delta_{p_1 q_{\sigma_1(1)}} \cdots \delta_{p_k q_{\sigma_1(k)}} = (P_\mu(T))^{p_1 \cdots p_k} \delta_{p_1 q_{\sigma_2(1)}} \cdots \delta_{p_k q_{\sigma_2(k)}} \tag{5}$$

for every choice of the indexes $p_1, \ldots, p_k$ and $q_1, \ldots, q_k$ and every T. Take T to be such that $T^{i_1 \cdots i_k} \neq 0$ for every choice of the indexes $i_1, \ldots, i_k$ satisfying $i_s = i_{\mu(s)}$ for every $s \in \mathbb{N}_k$. Suppose that $\mu \not\preceq \sigma_2^{-1} \circ \sigma_1$. This means that there is a number $s_0 \in \mathbb{N}_k$ such that $\sigma_2^{-1}(\sigma_1(s_0))$ and s_0 are not in the same cycle of μ. Choose the indexes $p_1, \ldots, p_k$ and $q_1, \ldots, q_k$ so that $p_s = p_{\mu(s)}$ and $p_s = q_{\sigma_1(s)}$ for every $s \in \mathbb{N}_k$. Moreover, choose the indexes $p_1, \ldots, p_k$ so that if $s, r \in \mathbb{N}_k$ are not in the same cycle of μ, then $p_s \neq p_r$. This in particular means that

$$p_{\sigma_2^{-1}(\sigma_1(s_0))} \neq p_{s_0}. \tag{6}$$

With those choices, the left-hand side of (5) is equal to $T^{p_1 \cdots p_k} \neq 0$. We reach a contradiction by showing that for some r_0, $p_{r_0} \neq q_{\sigma_2(r_0)}$, implying that the right-hand side of (5) is zero. Suppose on the contrary that $p_r = q_{\sigma_2(r)}$ for every $r \in \mathbb{N}_k$. Then $p_{\sigma_2^{-1}(\sigma_1(s))} = q_{\sigma_1(s)} = p_s$ for every $s \in \mathbb{N}_k$, contradicting (6). We are done. □

Notice that if $\mu \preceq \nu$, then for arbitrary permutation σ in P^k we have

$$\mu \preceq \nu^{-1} = (\sigma \circ \nu)^{-1} \circ \sigma.$$

This observation leads to the next corollary.

COROLLARY 3.2. [illegible] μ [illegible] ν [illegible] P^k [illegible] $\mu \preceq \nu$ [illegible] $\sigma \in P^k$ [illegible] $H_1, \ldots, H_k$ [illegible] T [illegible] $T^{k,n}$ [illegible]

$$\langle P_\mu(T), H_1 \circ_\sigma \cdots \circ_\sigma H_k \rangle = \langle P_\mu(T), H_1 \circ_{\sigma\circ\nu} \cdots \circ_{\sigma\circ\nu} H_k \rangle.$$

[illegible] $\nu = \mu$, $\nu = \mu^{-1}$

It is useful to see explicitly the conclusions of the above theorem when $k \leq 3$. We summarize them in the next corollary.

COROLLARY 3.3. [illegible] $T \in T^{2,n}$ [illegible] H_1 [illegible] H_2 [illegible]

$$\langle P_{(12)}(T), H_1 \circ_{(1)(2)} H_2 \rangle = \langle P_{(12)}(T), H_1 \circ_{(12)} H_2 \rangle.$$

[illegible] $T \in T^{3,n}$ [illegible] H_1 H_2 [illegible] H_3 [illegible]

$$\begin{aligned} \langle P_{(13)}(T), H_1 \circ_{(132)} H_2 \circ_{(132)} H_3 \rangle &= \langle P_{(13)}(T), H_1 \circ_{(12)(3)} H_2 \circ_{(12)(3)} H_3 \rangle, \\ \langle P_{(23)}(T), H_1 \circ_{(123)} H_2 \circ_{(123)} H_3 \rangle &= \langle P_{(23)}(T), H_1 \circ_{(12)(3)} H_2 \circ_{(12)(3)} H_3 \rangle \end{aligned}$$

[illegible]

$$\begin{aligned} \langle P_{(13)}(T), H_1 \circ_{(13)(2)} H_2 \circ_{(13)(2)} H_3 \rangle &= \langle P_{(13)}(T), H_1 \circ_{(1)(2)(3)} H_2 \circ_{(1)(2)(3)} H_3 \rangle, \\ \langle P_{(23)}(T), H_1 \circ_{(1)(23)} H_2 \circ_{(1)(23)} H_3 \rangle &= \langle P_{(23)}(T), H_1 \circ_{(1)(2)(3)} H_2 \circ_{(1)(2)(3)} H_3 \rangle. \end{aligned}$$

[illegible] σ_1, σ_2 [illegible] $\mathbb{N}_3$ [illegible]

$$\langle P_{(123)}(T), H_1 \circ_{\sigma_1} H_2 \circ_{\sigma_1} H_3 \rangle = \langle P_{(123)}(T), H_1 \circ_{\sigma_2} H_2 \circ_{\sigma_2} H_3 \rangle.$$

3.4. Let us have another look at formula (1) for the first derivative of a spectral function at X. Let $X = V(\operatorname{Diag}\lambda(X))V^T$ and $\tilde{E} = V^T E V$, where E is a symmetric matrix. Using the definitions and notation in the previous subsection, we have

$$\begin{aligned}\nabla(f\circ\lambda)(X)[E] &= \langle V\big(\operatorname{Diag}\nabla f(\mu)\big)V^T, E\rangle\\ &= \langle \nabla f(\mu), \operatorname{diag}\tilde{E}\rangle\\ &= \langle \nabla f(\mu), \circ_{(1)}\tilde{E}\rangle.\end{aligned}$$

3.5. Let X be a symmetric matrix with ordered spectral decomposition $X = V(\operatorname{Diag}\lambda(X))V^T$. Take two symmetric matrices E_1 and E_2 and let $\tilde{E}_i = V^T E_i V$ for $i = 1, 2$. As we saw in the examples in section 2, we have

$$E_1 \circ_{(1)(2)} E_2 = (\operatorname{diag}E_1)(\operatorname{diag}E_2)^T \quad\text{and}\quad E_1 \circ_{(12)} E_2 = E_1 \circ E_2.$$

Then formula (2) for the Hessian of the spectral function $f\circ\lambda$ becomes

$$\begin{aligned}\nabla^2(f\circ\lambda)(X)[E_1,E_2] &= \nabla^2 f(\lambda(X))[\operatorname{diag}\tilde{E}_1, \operatorname{diag}\tilde{E}_2] + \langle \mathcal{A}(\lambda(X)), \tilde{E}_1\circ\tilde{E}_2\rangle\\ &= \langle \nabla^2 f(\lambda(X)), \tilde{E}_1 \circ_{(1)(2)} \tilde{E}_2\rangle + \langle \mathcal{A}(\lambda(X)), \tilde{E}_1 \circ_{(12)} \tilde{E}_2\rangle.\end{aligned}$$

These examples support the following conjecture, describing the structure of the higher order derivatives of spectral functions. (More instances of when the conjecture holds are given after its equivalent reformulation in Conjecture 4.1.)

3.1. The spectral function $f\circ\lambda$ is k-times (continuously) differentiable at X if and only if $f(x)$ is k-times (continuously) differentiable at the vector $\lambda(X)$. Moreover, there are k-tensor valued maps $\mathcal{A}_\sigma : \mathbb{R}^n \to T^{k,n}$, $\sigma \in P^k$, depending only on the symmetric function f, such that for any symmetric matrices $E_1, \ldots, E_k$, we have

$$\nabla^k(f\circ\lambda)(X)[E_1,\ldots,E_k] = \sum_{\sigma\in P^k} \langle \mathcal{A}_\sigma(\lambda(X)), \tilde{E}_1 \circ_\sigma \cdots \circ_\sigma \tilde{E}_k\rangle, \tag{7}$$

where $X = V(\operatorname{Diag}\lambda(X))V^T$ and $\tilde{E}_i = V^T E_i V$ for $i = 1, \ldots, k$.

The left-hand side of formula (7) is the kth derivative of the spectral function evaluated at the matrices $E_1, \ldots, E_k$, while on the right side these matrices are conjugated by V and "jumbled" into the σ-Hadamard products $\tilde{E}_1 \circ_\sigma \cdots \circ_\sigma \tilde{E}_k$. Our goal in the next section is to identify more clearly the multilinear operator on the right-hand side of (7) acting on the matrices $E_1, \ldots, E_k$.

4. The $\operatorname{Diag}^\sigma$ operator. Recall that the adjoint of the linear operator $\operatorname{Diag} : \mathbb{R}^n \to M^n$ is the operator $\operatorname{diag} : M^n \to \mathbb{R}^n$. That is, we have the identity

$$\langle \operatorname{Diag} x, H\rangle = \langle x, \operatorname{diag} H\rangle \tag{8}$$

for any vector x and any matrix H. It is also easy to verify that for any vector x, matrix H, and orthogonal matrix U, we have

$$\langle U(\operatorname{Diag} x)U^T, H\rangle = \langle x, \operatorname{diag}(U^T H U)\rangle = \langle x, \circ_{(1)}(U^T H U)\rangle, \tag{9}$$

where the last equality is trivial from Example 2.5.

In this section we generalize (8) and (9) for an arbitrary k-tensor in place of x and an arbitrary σ-Hadamard product in place of $\circ_{(1)}$.

Let T be an arbitrary k-tensor on $\mathbb{R}^n$ and let σ be a permutation on $\mathbb{N}_k$. We define $\mathrm{Diag}^{\,\sigma}T$ to be a $2k$-tensor on $\mathbb{R}^n$ in the following way:

$$(\mathrm{Diag}^{\,\sigma}T)^{i_1\dots i_k}_{j_1\dots j_k} = \begin{cases} T^{i_1\dots i_k} & \text{if } i_s = j_{\sigma(s)}\ \forall s \in \mathbb{N}_k, \\ 0 & \text{otherwise.} \end{cases}$$

Informally speaking, viewing tensors as cubes placed at the origin of the positive orthant of a Euclidean space and its indices as coordinates, the operator $\mathrm{Diag}^{\,\sigma}T$ then lifts T onto the k-dimensional diagonal plane defined by

$$\{(x,y) \in \mathbb{R}^k \times \mathbb{R}^k \,|\, x_s = y_{\sigma(s)}\ \forall s \in \mathbb{N}_k\}.$$

Notice that this map between the permutations on $\mathbb{N}_k$ and k-dimensional diagonal subspaces of $\mathbb{R}^k \times \mathbb{R}^k$ is one-to-one.

When $k = 1$ and $\sigma = (1)$, the definition of $\mathrm{Diag}^{\,\sigma}T$ coincides with the definition of the Diag operator in (8). An equivalent way to define $\mathrm{Diag}^{\,\sigma}T$ useful for calculations is

$$(\mathrm{Diag}^{\,\sigma}T)^{i_1\dots i_k}_{j_1\dots j_k} = T^{i_1\dots i_k}\delta_{i_1 j_{\sigma(1)}}\cdots\delta_{i_k j_{\sigma(k)}}.$$

We now consider an action—call it [illegible]—of the group O^n of all $n \times n$ orthogonal matrices on the space of all k-tensors on $\mathbb{R}^n$. For any k-tensor T and $U \in O^n$, this action will be denoted by UTU^T and defined by

$$(UTU^T)^{i_1\dots i_k} = \sum_{p_1=1}^{n}\cdots\sum_{p_k=1}^{n}\Big(T^{p_1\dots p_k}U^{i_1 p_1}\dots U^{i_k p_k}\Big). \tag{10}$$

When $k = 1$, this is exactly the action of O^n on $\mathbb{R}^n$, and when $k = 2$ the definition coincides with the conjugate action of O^n on the space of $n \times n$ square matrices. In general, it is not difficult to see that it is possible to order the entries of $T \in T^{k,n}$ into a vector $\mathrm{vec}\,(T) \in \mathbb{R}^{n^k}$ such that

$$UTU^T = \mathrm{vec}^{-1}\big((\otimes^k U)\mathrm{vec}\,(T)\big), \tag{11}$$

where $\otimes^k U$ is the kth tensor power of U and vec^{-1} is the inverse of the linear operation vec. The fact that $\otimes^k U$ is an orthogonal matrix whenever U is, the well-known identity $(\otimes^k V)(\otimes^k U) = \otimes^k(VU)$, and (11) show the following lemma.

LEMMA 4.1. [illegible] k [illegible] T [illegible] $\mathbb{R}^n$ [illegible] U V [illegible] O^n

$$V(UTU^T)V^T = (VU)T(VU)^T$$

[illegible]

$$\|UTU^T\| = \|T\|.$$

Any $2k$-tensor T on R^n can naturally be viewed as a k-tensor on the space M^n in the following way. Let $H_1, \dots, H_k$ be any $n \times n$ matrices; then

$$T[H_1,\dots,H_k] := \sum_{p_1,q_1=1}^{n,n}\cdots\sum_{p_k,q_k=1}^{n,n} T^{p_1\dots p_k}_{q_1\dots q_k}H_1^{p_1 q_1}\cdots H_k^{p_k q_k}.$$

Let P be an $n \times n$ permutation matrix and σ its corresponding permutation on $\mathbb{N}_n$, that is, $P^T e^i = e^{\sigma(i)}$ for all $i = 1, \ldots, n$, where $\{e^i \mid i = 1, \ldots, n\}$ is the standard basis in $\mathbb{R}^n$. The action of P on the tensors is what one expects it to be:

$$(PTP^T)^{i_1 \ldots i_k} = \sum_{p_1=1}^{n} \cdots \sum_{p_k=1}^{n} \Big(T^{p_1 \ldots p_k} \prod_{\nu=1}^{k} P^{i_\nu p_\nu} \Big) = T^{\sigma(i_1) \ldots \sigma(i_k)}.$$

The conjugation by an orthogonal matrix is defined on tensors on $\mathbb{R}^n$ of any dimension. The next lemma shows that the conjugation by a permutation matrix commutes with the lifting operation Diag^μ for any permutation μ.

LEMMA 4.2. [illegible] μ [illegible] $\mathbb{N}_k$ [illegible] P, P^n [illegible] k [illegible] T [illegible] $\mathbb{R}^n$ [illegible]

$$P(\operatorname{Diag}^\mu T)P^T = \operatorname{Diag}^\mu(PTP^T).$$

[illegible] Let σ be the permutation on $\mathbb{N}_n$ corresponding to P. Fix any multi-index $\binom{i_1 \ldots i_k}{j_1 \ldots j_k}$. We begin calculating the right-hand side entry corresponding to that index. In the third equality below, we use the fact that σ is a one-to-one map.

$$\begin{aligned}
\big(P(\operatorname{Diag}^\mu T)P^T\big)^{i_1 \ldots i_k}_{j_1 \ldots j_k} &= \big(\operatorname{Diag}^\mu T\big)^{\sigma(i_1) \ldots \sigma(i_k)}_{\sigma(j_1) \ldots \sigma(j_k)} \\
&= T^{\sigma(i_1) \ldots \sigma(i_k)} \delta_{\sigma(i_1)\sigma(j_{\mu(1)})} \cdots \delta_{\sigma(i_k)\sigma(j_{\mu(k)})} \\
&= T^{\sigma(i_1) \ldots \sigma(i_k)} \delta_{i_1 j_{\mu(1)}} \cdots \delta_{i_k j_{\mu(k)}} \\
&= (PTP^T)^{i_1 \ldots i_k} \delta_{i_1 j_{\mu(1)}} \cdots \delta_{i_k j_{\mu(k)}} \\
&= \big(\operatorname{Diag}^\mu(PTP^T)\big)^{i_1 \ldots i_k}_{j_1 \ldots j_k}. \qquad \square
\end{aligned}$$

These preparations lead to the following generalization to (9). (When $k = 1$ and $\sigma = (1)$ we obtain (9) exactly.)

THEOREM 4.3. [illegible] k [illegible] T [illegible] $\mathbb{R}^n$ [illegible] $H_1, \ldots, H_k$, M^n [illegible] [illegible] U, O^n [illegible] σ [illegible] $\mathbb{N}_k$ [illegible]

$$\langle T, \tilde{H}_1 \circ_\sigma \cdots \circ_\sigma \tilde{H}_k \rangle = \big(U(\operatorname{Diag}^\sigma T)U^T\big)[H_1, \ldots, H_k], \tag{12}$$

[illegible] $\tilde{H}_i = U^T H_i U$ [illegible] $i = 1, 2, \ldots, k$ [illegible]. Since both sides are linear in each argument separately, it is enough to show that the equality holds for k-tuples $(H_{i_1 j_1}, \ldots, H_{i_k j_k})$ of basic matrices.

Using Lemma 2.2 and the fact that $\tilde{H}^{pq}_{ij} = U^{ip} U^{jq}$, we develop the left-hand side of (12):

$$\begin{aligned}
\langle T, \tilde{H}_{i_1 j_1} \circ_\sigma \cdots \circ_\sigma \tilde{H}_{i_k j_k} \rangle &= \sum_{p_1, \ldots, p_k = 1}^{n, \ldots, n} T^{p_1 \ldots p_k} \tilde{H}^{p_1 p_{\sigma^{-1}(1)}}_{i_1 j_1} \cdots \tilde{H}^{p_k p_{\sigma^{-1}(k)}}_{i_k j_k} \\
&= \sum_{p_1, \ldots, p_k = 1}^{n, \ldots, n} T^{p_1 \ldots p_k} U^{i_1 p_1} U^{j_1 p_{\sigma^{-1}(1)}} \cdots U^{i_k p_k} U^{j_k p_{\sigma^{-1}(k)}}.
\end{aligned}$$

On the other hand, using the definitions we calculate that the right-hand side is

$$\begin{aligned}
&(U(\operatorname{Diag}^{\sigma}T)U^T)[H_{i_1j_1},\dots,H_{i_kj_k}]\\
&\quad=\sum_{p_1,q_1=1}^{n,n}\cdots\sum_{p_k,q_k=1}^{n,n}\big((U(\operatorname{Diag}^{\sigma}T)U^T)^{p_1\dots p_k}_{q_1\dots q_k}H^{p_1q_1}_{i_1j_1}\cdots H^{p_kq_k}_{i_kj_k}\big)\\
&\quad=(U(\operatorname{Diag}^{\sigma}T)U^T)^{i_1\dots i_k}_{j_1\dots j_k}\\
&\quad=\sum_{p_1,q_1=1}^{n,n}\cdots\sum_{p_k,q_k=1}^{n,n}\Big((\operatorname{Diag}^{\sigma}T)^{p_1\dots p_k}_{q_1\dots q_k}\prod_{\nu=1}^{k}U^{i_\nu p_\nu}U^{j_\nu q_\nu}\Big)\\
&\quad=\sum_{p_1=1}^{n}\cdots\sum_{p_k=1}^{n}\Big(T^{p_1\dots p_k}\prod_{\nu=1}^{k}U^{i_\nu p_\nu}U^{j_\nu p_{\sigma^{-1}(\nu)}}\Big).
\end{aligned}$$

This shows that both sides are equal. □

COROLLARY 4.4. [illegible] k [illegible] T [illegible] $H_1,\dots,H_k$ [illegible] σ [illegible] $\mathbb{N}_k$ [illegible]

$$\langle T, H_1\circ_\sigma\dots\circ_\sigma H_k\rangle=(\operatorname{Diag}^{\sigma}T)[H_1,\dots,H_k]. \tag{13}$$

If in Corollary 4.4 we substitute the matrices $H_1,\dots,H_k$ with $\tilde H_1,\dots,\tilde H_k$ and use Theorem 4.3, we obtain the next result.

COROLLARY 4.5. [illegible] k [illegible] T [illegible] $U\in O^n$ [illegible] σ [illegible] $\mathbb{N}_k$ [illegible] $H_1,\dots,H_k$ [illegible]

$$(\operatorname{Diag}^{\sigma}T)[\tilde H_1,\dots,\tilde H_k]=\big(U(\operatorname{Diag}^{\sigma}T)U^T\big)[H_1,\dots,H_k]. \tag{14}$$

If in Corollary 4.4 we take σ to be the identity permutation, then we get the next corollary, which generalizes (8).

COROLLARY 4.6. [illegible] k [illegible] T [illegible] $H_1,\dots,H_k$ [illegible]

$$T[\operatorname{diag}H_1,\dots,\operatorname{diag}H_k]=(\operatorname{Diag}^{(\mathrm{id})}T)[H_1,\dots,H_k]. \tag{15}$$

We conclude this section with a second look at the first two derivatives of spectral functions.

[illegible] 4.7. As we saw in Example 3.4, the first derivative of the spectral function $f\circ\lambda$ at the point $X=V(\operatorname{Diag}\lambda(X))V^T$, applied to the symmetric matrix E, is given by the formula

$$\nabla(f\circ\lambda)(X)[E]=\langle V\big(\operatorname{Diag}\nabla f(\lambda(X))\big)V^T,E\rangle=V\big(\operatorname{Diag}^{(1)}\nabla f(\lambda(X))\big)V^T[E].$$

The usefulness of the notation becomes more evident below.

[illegible] 4.8. Let X be a symmetric matrix with ordered spectral decomposition $X=V(\operatorname{Diag}\lambda(X))V^T$. Take two symmetric matrices E_1 and E_2 and let $\tilde E_i=V^TE_iV$ for $i=1,2$. As we saw in Example 3.4, the Hessian of the spectral function $f\circ\lambda$ at the point $X=V(\operatorname{Diag}\lambda(X))V^T$, applied to the symmetric matrices E_1 and E_2, is given by the formula

$$\nabla^2(f\circ\lambda)(X)[E_1,E_2]=\langle\nabla^2f(\lambda(X)),\tilde E_1\circ_{(1)(2)}\tilde E_2\rangle+\langle\mathcal{A}(\lambda(X)),\tilde E_1\circ_{(12)}\tilde E_2\rangle.$$

With the notation introduced in this section we can rewrite it as

$$\nabla^2(f\circ\lambda)(X)[E_1,E_2] = \big(V\big(\mathrm{Diag}^{(1)(2)}\nabla^2 f(\lambda(X))\big)V^T\big)[E_1,E_2] + \big(V\big(\mathrm{Diag}^{(12)}\mathcal{A}(\lambda(X))\big)V^T\big)[E_1,E_2],$$

or, in other words,

$$\nabla^2(f\circ\lambda)(X) = V\big(\mathrm{Diag}^{(1)(2)}\nabla^2 f(\lambda(X)) + \mathrm{Diag}^{(12)}\mathcal{A}(\lambda(X))\big)V^T.$$

Finally, we express Conjecture 3.1 in the new language.

[illegible] 4.1. The spectral function $f\circ\lambda$ is k-times (continuously) differentiable at X if and only if $f(x)$ is k-times (continuously) differentiable at the vector $\lambda(X)$. Moreover, there are k-tensor valued maps $\mathcal{A}_\sigma : \mathbb{R}^n \to T^{k,n}$, $\sigma \in P^k$, depending only on the symmetric function f, such that

$$\nabla^k(f\circ\lambda)(X) = V\Big(\sum_{\sigma\in P^k}\mathrm{Diag}^{\sigma}\mathcal{A}_\sigma(\lambda(X))\Big)V^T, \tag{16}$$

where $X = V(\mathrm{Diag}\,\lambda(X))V^T$.

Formula (16) says that the orthogonal matrix V in the ordered spectral decomposition of X also "diagonalizes" the kth derivative of $f\circ\lambda$ at X. Moreover, the effect of the eigenvalues in the right-hand side of (16) can very clearly be seen: only V and $\lambda(X)$ depend on the eigenvalues. In addition, we can easily evaluate the derivative, as a multilinear function, at any k symmetric matrices, using Theorem 4.3 and the σ-Hadamard product. Finally, there are precisely $k!$ summands in the right-hand side of (16); this should be compared with the classical Faà de Bruno formula [5, Lemma 1.3.1] for the kth derivative of the composition of two (smooth) functions, in which the number of summands in highly nontrivial.

In [16] we show that this conjecture holds for the derivatives of any function (not necessarily symmetric) of the eigenvalues of symmetric matrices, at a symmetric matrix X with distinct eigenvalues, as well as for the derivatives of separable spectral functions at an arbitrary symmetric matrix X. (Separable spectral functions are those arising from symmetric functions $f(x) = g(x_1) + \cdots + g(x_n)$ for some function g on a scalar argument.) There we also describe how, for every σ in P^k, to compute the operators $\mathcal{A}_\sigma(x)$, depending only on the symmetric function $f(x)$.

5. Sufficient condition for Conjecture 4.1. Recall that Examples 4.7, and 4.8 show that Conjecture 4.1 holds for $k = 1$ and $k = 2$. The next theorem summarizes this section.

THEOREM 5.1. [illegible] 4.1 [illegible]

- [illegible] 4.1 [illegible] $X = \mathrm{Diag}\,x$ [illegible] $x \in \mathbb{R}^n$ [illegible] $x_1 \geq \cdots \geq x_n$
- [illegible] $\mathcal{A}_\sigma$ [illegible] $\lambda(X)$ [illegible] $\sigma \in P^k$ [illegible] $\nabla^k(f\circ\lambda)$ [illegible] X

We begin with a simple lemma. For brevity, given a k-tensor T on M^n by $T[H]$, we denote the $(k-1)$-tensor $T[\cdot,\ldots,H]$.

LEMMA 5.2. [illegible] T [illegible] $2k$ [illegible] R^n $U \in O^n$ [illegible] H [illegible]

$$U(T[\tilde{H}])U^T = (UTU^T)[H],$$

[illegible] $\tilde{H} = U^T H U$

Since both sides are linear with respect to H, it is enough to prove the identity only for basic matrices $H_{i_kj_k}$. By the definition of conjugation, and using the fact that $\tilde{H}^{pq}_{i_kj_k} = U^{i_kp}U^{j_kq}$, we obtain

$$\begin{aligned}
&\big(U(T[\tilde{H}_{i_kj_k}])U^T\big)^{i_1\dots i_{k-1}}_{j_1\dots j_{k-1}} \\
&\qquad = \sum_{\substack{p_s,q_s=1\\ s=1,\dots,k-1}}^{n,\dots,n} (T[\tilde{H}_{i_kj_k}])^{p_1\dots p_{k-1}}_{q_1\dots q_{k-1}} U^{i_1p_1}U^{j_1q_1}\dots U^{i_{k-1}p_{k-1}}U^{j_{k-1}q_{k-1}} \\
&\qquad = \sum_{\substack{p_s,q_s=1\\ s=1,\dots,k}}^{n,\dots,n} T^{p_1\dots p_k}_{q_1\dots q_k} U^{i_1p_1}U^{j_1q_1}\dots U^{i_kp_k}U^{j_kq_k} \\
&\qquad = (UTU^T)^{i_1\dots i_k}_{j_1\dots j_k} \\
&\qquad = \big((UTU^T)[H_{i_kj_k}]\big)^{i_1\dots i_{k-1}}_{j_1\dots j_{k-1}}. \qquad \square
\end{aligned}$$

We now establish the first part of Theorem 5.1. Suppose that Conjecture 4.1 holds for all derivatives of order less than k and for the kth derivative it holds only for ordered diagonal matrices. We show that the conjecture holds for the kth derivative at an arbitrary matrix. Indeed, let $X = V(\operatorname{Diag}\lambda(X))V^T$, let E be arbitrary symmetric matrix, and denote $\tilde{E} = V^TEV$. Then

$$\begin{aligned}
&\nabla^{k-1}F(X+E) = \nabla^{k-1}F\big(V(\operatorname{Diag}\lambda(X)+\tilde{E})V^T\big) \\
&= V\big(\nabla^{k-1}F(\operatorname{Diag}\lambda(X)+\tilde{E})\big)V^T \\
&= V\big(\nabla^{k-1}F(\operatorname{Diag}\lambda(X))\big)V^T + V\big(\nabla^kF(\operatorname{Diag}\lambda(X))[\tilde{E}]\big)V^T + o(\|E\|) \\
&= \nabla^{k-1}F(X) + \big(V(\nabla^kF(\operatorname{Diag}\lambda(X)))V^T\big)[E] + o(\|E\|),
\end{aligned}$$

where in the last equality we used Lemma 5.2. This shows that $\nabla^{k-1}F$ is differentiable at X and that $V(\nabla^kF(\operatorname{Diag}\lambda(X)))V^T$ is the kth derivative of F at X.

The second part of Theorem 5.1 is the next proposition.

PROPOSITION 5.3. [illegible] k [illegible] $F = f\circ\lambda$ [illegible] (16) [illegible] X [illegible] $\sigma\in P^k$ [illegible] $x\in\mathbb{R}^n \to \mathcal{A}_\sigma(x)\in T^{k,n}$ [illegible] $\nabla^kF(X)$ [illegible] X [illegible] $F\in C^k$ [illegible]. Suppose that there is a sequence of symmetric matrices X_m approaching X and an $\epsilon>0$ such that

$$\|\nabla^kF(X_m) - \nabla^kF(X)\| > \epsilon\ \forall m.$$

Let $X_m = V_m(\operatorname{Diag}\lambda(X_m))V_m^T$ and suppose without loss of generality that the orthogonal V_m approaches V (otherwise, take a subsequence.) By continuity of the eigenvalues, we have that $X = V(\operatorname{Diag}\lambda(X))V^T$ and that $\lambda(X_m)$ approaches $\lambda(X)$. Using the formula for the kth derivative and the continuity of the maps $\mathcal{A}_\sigma(x)$, the contradiction follows. $\square$

Acknowledgment. We would like to thank the first referee for a thorough reading of the manuscript, useful comments, and pointing out several typos.

REFERENCES

[1] J. Dadok, *On the C^∞ Chevalley's theorem*, Adv. in Math., 44 (1982), pp. 121–131.

[2] C. Davis, *All convex invariant functions of hermitian matrices*, Arch. Math., 8 (1957), pp. 276–278.

[3] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Grundlehren Math. Wiss. 132, Springer-Verlag, New York, Berlin, 1976.

[4] E. C. Kemble, *The Fundamental Principles of Quantum Mechanics*, Dover, New York, 1958.

[5] S. G. Krantz and H. R. Parks, *A Primer of Real Analytic Functions*, Basler Lehrbücker 4, Birkhäuser Verlag, Basel, Switzerland, 1992.

[6] A. S. Lewis, *Convex analysis on the Hermitian matrices*, SIAM J. Optim., 6 (1996), pp. 164–177.

[7] A. S. Lewis, *Derivatives of spectral functions*, Math. Oper. Res., 21 (1996), pp. 576–588.

[8] A. S. Lewis, *Nonsmooth analysis of eigenvalues*, Math. Program., 84 (1999), pp. 1–24.

[9] A. S. Lewis, *The mathematics of eigenvalue optimization*, Math. Program., 97 (2003), pp. 155–176.

[10] A. S. Lewis and M. L. Overton, *Eigenvalue optimization*, Acta Numer., 5 (1996), pp. 149–190.

[11] A. S. Lewis and H. S. Sendov, *Twice differentiable spectral functions*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 368–386.

[12] A. S. Lewis and H. S. Sendov, *Quadratic expansions of spectral functions*, Linear Algebra Appl., 340 (2002), pp. 97–121.

[13] H. Qi and X. Yang, *Semismoothness of spectral functions*, SIAM J. Matrix Anal. Appl., 25 (2004), pp. 766–783.

[14] A. Seeger, *Convex analysis of spectrally defined matrix functions*, SIAM J. Optim., 7 (1997), pp. 679–696.

[15] H. S. Sendov, *Interactions between the Generalized Hadamard Product and the Eigenvalues of Symmetric Matrices*, Research report, Mathematical Series 2004-309, Department of Mathematics and Statistics, University of Guelph, Ontario, 2004.

[16] H. S. Sendov, *On the Higher-Order Derivatives of Spectral Functions: Two Special Cases*, Research report, Mathematical Series 2004-310, Department of Mathematics and Statistics, University of Guelph, Ontario, Canada, 2004.

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

[18] M. Torki, *Second-Order Epi-differentiability and Subdifferentiability of Spectral Convex Functions*, personal communication, 1999.

[19] N.-K. Tsing, M. K. H. Fan, and E. I. Verriest, *On analyticity of functions involving eigenvalues*, Linear Algebra Appl., 207 (1994), pp. 159–180.

[20] G. Warner, *Harmonic Analysis on Semi-Simple Lie Groups*, Springer-Verlag, Berlin, Heidelberg, New York, 1972.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 3, pp. 682–702

A MATRIX PENCIL APPROACH TO THE ROW BY ROW DECOUPLING PROBLEM FOR DESCRIPTOR SYSTEMS*

DELIN CHU† AND Y. S. HUNG‡

Abstract. The row by row decoupling problem (RRDP) for descriptor systems is considered using proportional state feedback and input transformation. Necessary and sufficient conditions for the solvability of the RRDP are provided. These solvability conditions can be readily verified. A constructive solution to the RRDP is given so that the desired feedback and input transformation matrices can be obtained by a numerically reliable procedure.

Key words. row by row decoupling, descriptor systems, orthogonal transformation

AMS subject classifications. 93B05, 93B40, 93B52, 65F35

DOI. 10.1137/S089547980343905X

1. Introduction.

The row by row decoupling problem (RRDP) has played a central role in classical as well as modern control theory, since it provides a powerful methodology to reduce a multi-input/multioutput complex system to a set of single input/single output systems, thus facilitating a decoupled control strategy of such systems. The row by row decoupling is usually required for ease of system operations, for example, in the process and chemical industries [1, 2].

Consider descriptor systems of the form

$$\begin{cases} E\dot{x} = Ax + Bu, \\ y = Cx, \end{cases} \tag{1}$$

where $E, A \in \mathbf{R}^{n\times n}$, $B \in \mathbf{R}^{n\times m}$, $C \in \mathbf{R}^{m\times n}$, E is [illegible], $x \in \mathbf{R}^n$ is the state, $u \in \mathbf{R}^m$ is the control input, and $y \in \mathbf{R}^m$ is the output. It is well known that the existence and uniqueness of (classical) solutions to (1) are guaranteed if (E, A) is regular, i.e., $\det(\alpha E - \beta A) \neq 0$ for some $(\alpha, \beta) \in \mathbf{C}^2$. The system (1) is said to have [illegible] if the dimension of the largest nilpotent block in the Kronecker canonical form of (E, A) is at most one [17].

Descriptor systems that are regular and of index at most one can be separated into purely dynamical and purely algebraic parts (fast and slow modes). If the index is larger than 1, then impulses can arise in the response of the system if the control is not sufficiently smooth [7, 17]. Therefore, in the design of feedback control, one should ensure that the closed-loop system is regular and of index at most one.

If we apply state feedback of the form

$$u = Fx + Hv \tag{2}$$

to the descriptor system (1), then the closed-loop system becomes

$$\begin{cases} E\dot{x} = (A + BF)x + BHv, \\ y = Cx. \end{cases} \tag{3}$$

*Received by the editors December 23, 2003; accepted for publication (in revised form) by B. T. Kågström March 7, 2006; published electronically September 19, 2006.
http://www.siam.org/journals/simax/28-3/43905.html

†Department of Mathematics, National University of Singapore, 2 Science Drive 2, Singapore 117543 (matchudl@math.nus.edu.sg).

‡Department of Electrical and Electronic Engineering, University of Hong Kong, Pokfulam Road, Hong Kong (yshung@hkueee.hku.hk). The work of this author was supported by a CRCG research grant from the University of Hong Kong.

The problem to be considered can be stated as follows.

. Given a descriptor system of the form (1), determine a state feedback matrix $F \in \mathbf{R}^{m\times n}$ and a nonsingular input transformation matrix $H \in \mathbf{R}^{m\times m}$ such that

(a) the pencil $(E, A + BF)$ is regular and of index at most one;

(b) the closed-loop transfer function matrix

$$C(sE - A - BF)^{-1}BH \quad \text{is nonsingular and diagonal.} \tag{4}$$

Here $C(sE - A - BF)^{-1}BH$ being nonsingular means that

$$\operatorname{rank}(C(sE - A - BF)^{-1}BH) = m \quad \text{for some } s \in \mathbf{C}.$$

Before studying the RRDP for the descriptor system (1), we summarize the main results on the RRDP available in the literature.

The RRDP for linear time-invariant systems of the form

$$\begin{cases} \mathcal{E}\dot{x} = \mathcal{A}x + \mathcal{B}u, \\ y = \mathcal{C}x \end{cases} \tag{5}$$

with $\mathcal{E}, \mathcal{A} \in \mathbf{R}^{k\times k}$, $\mathcal{B} \in \mathbf{R}^{k\times p}$, $\mathcal{C} \in \mathbf{R}^{p\times k}$, and $\mathcal{E}$ has been investigated extensively over the last three decades and is still attracting continuing interests [2, 3, 4, 5, 6, 9, 10, 21]. In particular, because system (5) is equivalent to

$$\begin{cases} \dot{x} = \mathcal{E}^{-1}\mathcal{A}x + \mathcal{E}^{-1}\mathcal{B}u, \\ y = \mathcal{C}x, \end{cases}$$

we have the following theorem.

THEOREM 1 (see [3]). (5) $\mathcal{E}$ c_i i $\mathcal{C}$

$$c_i(\mathcal{E}^{-1}\mathcal{A})^j(\mathcal{E}^{-1}\mathcal{B}) \neq 0$$

j

$$l_i = \min\{j \geq 0 : j \text{ is integer satisfying } c_i(\mathcal{E}^{-1}\mathcal{A})^j(\mathcal{E}^{-1}\mathcal{B}) \neq 0\};$$

$l_i = k - 1$

$$\mathcal{L} = \begin{bmatrix} c_1(\mathcal{E}^{-1}\mathcal{A})^{l_1} \\ c_2(\mathcal{E}^{-1}\mathcal{A})^{l_2} \\ \\ c_p(\mathcal{E}^{-1}\mathcal{A})^{l_p} \end{bmatrix} (\mathcal{E}^{-1}\mathcal{B}), \quad \mathcal{K} = \begin{bmatrix} c_1(\mathcal{E}^{-1}\mathcal{A})^{l_1+1} \\ c_2(\mathcal{E}^{-1}\mathcal{A})^{l_2+1} \\ \\ c_p(\mathcal{E}^{-1}\mathcal{A})^{l_p+1} \end{bmatrix} (\mathcal{E}^{-1}\mathcal{B}).$$

(5) $\mathcal{L}$ $(\mathcal{F}, \mathcal{H})$

$$\mathcal{F} = -\mathcal{L}^{-1}\mathcal{K}, \quad \mathcal{H} = \mathcal{L}^{-1}.$$

Although Theorem 1 provides an explicit solution for the RRDP of the linear time-invariant system (5) with $\mathcal{E}$ nonsingular, some natural questions remain.

(a) If the RRDP is solvable,

(i) can we solve the RRDP with the additional requirement of stability?

(ii) does numerically reliable solution exist for the RRDP with stability?

(b) If the RRDP is not solvable, can one resort to a relaxed problem of triangular decoupling?

Regarding (a)(i), the RRDP with stability for the system (5) has been investigated in [5, 6] using geometric and structural approaches, giving coordinate-free solvability conditions. But, the results in [5, 6] cannot lead to numerically reliable methods for computing a solution to the problem. To address (a)(ii), a numerically reliable method has been developed in [14] based on orthogonal transformations. If the condition of Theorem 1 is not satisfied so that the RRDP is not solvable, it is shown in [15] that a triangular decoupling problem may be solvable under less restrictive conditions. In [15], explicit solvability conditions are provided with a parameterization of all solutions to the triangular decoupling problem.

Unfortunately, the above results for the RRDP for the system (5) with $\mathcal{E}$ nonsingular cannot be readily extended to the general descriptor system (1). For example, it is not possible to apply existing results to system (1) by decomposing it into differential and algebraic parts and then deal with them separately. Instead, it is necessary to develop a separate theory to handle the RRDP for descriptor systems. The RRDP for descriptor system (1) has been studied in [7, 11, 19]. In [7], it is shown that the RRDP for system (1) is solvable using combined proportional and derivative state feedback if and only if the input-output transfer function is invertible. However, the use of derivative feedback is undesirable due to noise accentuation and an increase in the system order. To our knowledge, the solution is still not known for the RRDP of the descriptor system (1) using only proportional state feedback.

A problem related to the RRDP is the disturbance decoupling problem. Although the objectives of the RRDP are different from the disturbance decoupling problem, we will make use of the matrix pencil approach developed in [12, 13] to characterize necessary and sufficient conditions for the solvability of the RRDP for the system (1). In this paper, we provide numerically reliable methods for verifying the solvability of the RRDP for the system (1) and for computing the solution matrices F and H. These results are new to our knowledge and are valuable, as real descriptor systems with singular E do exist in practice. However, the RRDP with stability for the descriptor system (1) remains an open problem.

The paper is organized as follows. Some necessary preliminary results for matrix pencils are collected in section 2. In section 3, necessary and sufficient solvability conditions as well as a numerically reliable algorithm for the RRDP of descriptor system (1) are established. Concluding remarks are included in section 4.

2. Preliminaries. The following two lemmas are basic results for matrix pencils and will be needed in the development to be given in the next section.

LEMMA 2 (see [12, 14]). [illegible] $\mathcal{E}, \mathcal{A} \in \mathbf{R}^{n\times n}, \mathcal{B} \in \mathbf{R}^{n\times m}, \mathcal{C} \in \mathbf{R}^{p\times n}$ [illegible] $\mathcal{D} \in \mathbf{R}^{p\times m}$ [illegible] $\mathcal{E}$ [illegible]

(i)

$$\mathcal{C}(s\mathcal{E} - \mathcal{A})^{-1}\mathcal{B} + \mathcal{D} = 0$$

[illegible]

$$\mathcal{D} = 0, \quad \max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} s\mathcal{E} - \mathcal{A} & \mathcal{B} \\ \mathcal{C} & 0 \end{bmatrix} = n;$$

(ii) [illegible]

$$\operatorname{rank} \begin{bmatrix} s\mathcal{E} - \mathcal{A} & \mathcal{B} \end{bmatrix} = n \quad \forall s \in \mathbf{C}.$$

$$\max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} s\mathcal{E}-\mathcal{A} & \mathcal{B} \\ \mathcal{C} & -\mathcal{D} \end{bmatrix} = n$$

$$\mathcal{C}=0, \quad \mathcal{D}=0.$$

LEMMA 3. $\mathcal{E},\mathcal{A}\in\mathbf{R}^{n\times l},\mathcal{B}\in\mathbf{R}^{n\times m},\mathcal{C}\in\mathbf{R}^{p\times n}$ $\mathcal{D}\in\mathbf{R}^{p\times m}$
(i)

$$\max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} s\mathcal{E}-\mathcal{A} & \mathcal{B} \\ \mathcal{C} & -\mathcal{D} \end{bmatrix} \geq \operatorname{rank}(\mathcal{E}) + \operatorname{rank}(\mathcal{D});$$

(ii) $\mathcal{E}$ $\mathcal{D}$

$$\max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} s\mathcal{E}-\mathcal{A} & \mathcal{B} \\ \mathcal{C} & -\mathcal{D} \end{bmatrix} = n+p. \tag{6}$$

(i). We can assume without loss of generality that

$$\mathcal{D} = \begin{matrix} \\ \\ \end{matrix} \begin{matrix} \tau & m-\tau \end{matrix} \begin{bmatrix} \mathcal{D}_{11} & 0 \\ 0 & 0 \end{bmatrix} \begin{matrix} \}\tau \\ \}p-\tau \end{matrix}$$

with $\mathcal{D}_{11}$ nonsingular and $\operatorname{rank}(\mathcal{D}) = \operatorname{rank}(\mathcal{D}_{11}) = \tau$. Denote

$$\mathcal{B} = \begin{matrix} \end{matrix} \begin{matrix} \tau & m-\tau \end{matrix} \begin{bmatrix} \mathcal{B}_1 & \mathcal{B}_2 \end{bmatrix}, \quad \mathcal{C} = \begin{bmatrix} \mathcal{C}_1 \\ \mathcal{C}_2 \end{bmatrix} \begin{matrix} \}\tau \\ \}p-\tau \end{matrix},$$

and let the generalized upper triangular form [8, 20] of

$$\begin{bmatrix} s\mathcal{E}-\mathcal{A}+\mathcal{B}_1\mathcal{D}_{11}^{-1}\mathcal{C}_1 & \mathcal{B}_2 \\ \mathcal{C}_2 & 0 \end{bmatrix}$$

be

$$\mathcal{P}\begin{bmatrix} s\mathcal{E}-\mathcal{A}+\mathcal{B}_1\mathcal{D}_{11}^{-1}\mathcal{C}_1 & \mathcal{B}_2 \\ \mathcal{C}_2 & 0 \end{bmatrix}\mathcal{Q}$$

$$= \begin{matrix} \\ \\ \\ \\ \end{matrix} \begin{matrix} \mu_1 & \mu_2 & \mu_3 & \mu_4 \end{matrix} \begin{bmatrix} s\Theta_{11}-\Phi_{11} & s\Theta_{12}-\Phi_{12} & s\Theta_{13}-\Phi_{13} & s\Theta_{14}-\Phi_{14} \\ 0 & s\Theta_{22}-\Phi_{22} & s\Theta_{23}-\Phi_{23} & s\Theta_{24}-\Phi_{24} \\ 0 & 0 & s\Theta_{33}-\Phi_{33} & s\Theta_{34}-\Phi_{34} \\ 0 & 0 & 0 & s\Theta_{44}-\Phi_{44} \end{bmatrix} \begin{matrix} \}\nu_1 \\ \}\mu_2 \\ \}\mu_3 \\ \}\nu_4 \end{matrix},$$

where $\mathcal{P}$ and $\mathcal{Q}$ are orthogonal, Θ_{11} is of full row rank, Θ_{44} is of full column rank, Θ_{22} is nonsingular, and

$$\operatorname{rank}(s\Theta_{11}-\Phi_{11})=\nu_1, \quad \operatorname{rank}(s\Theta_{33}-\Phi_{33})=\mu_3, \quad \operatorname{rank}(s\Theta_{44}-\Phi_{44})=\mu_4 \quad \forall s\in\mathbf{C}. \tag{7}$$

Now we have that

$$\operatorname{rank}(\mathcal{E}) = \operatorname{rank}\begin{bmatrix} \Theta_{11} & \Theta_{12} & \Theta_{13} & \Theta_{14} \\ 0 & \Theta_{22} & \Theta_{23} & \Theta_{24} \\ 0 & 0 & \Theta_{33} & \Theta_{34} \\ 0 & 0 & 0 & \Theta_{44} \end{bmatrix} = \nu_1+\mu_2+\operatorname{rank}(\Theta_{33})+\mu_4 \le \nu_1+\mu_2+\mu_3+\mu_4$$

and the equality $\operatorname{rank}(\mathcal{D}) = \tau$ yield that

$$\begin{aligned}
&\max_{s\in\mathbf{C}} \operatorname{rank}\begin{bmatrix} s\mathcal{E}-\mathcal{A} & \mathcal{B} \\ \mathcal{C} & -\mathcal{D} \end{bmatrix} \\
&\quad = \max_{s\in\mathbf{C}} \operatorname{rank}\begin{bmatrix} s\mathcal{E}-\mathcal{A} & \mathcal{B}_1 & \mathcal{B}_2 \\ \mathcal{C}_1 & -\mathcal{D}_{11} & 0 \\ \mathcal{C}_2 & 0 & 0 \end{bmatrix} \\
&\quad = \max_{s\in\mathbf{C}} \operatorname{rank}\begin{bmatrix} s\mathcal{E}-\mathcal{A}+\mathcal{B}_1\mathcal{D}_{11}^{-1}\mathcal{C}_1 & \mathcal{B}_2 \\ \mathcal{C}_2 & 0 \end{bmatrix} + \tau \\
&\quad = \max_{s\in\mathbf{C}} \operatorname{rank}\begin{bmatrix} s\Theta_{11}-\Phi_{11} & s\Theta_{12}-\Phi_{12} & s\Theta_{13}-\Phi_{13} & s\Theta_{14}-\Phi_{14} \\ 0 & s\Theta_{22}-\Phi_{22} & s\Theta_{23}-\Phi_{23} & s\Theta_{24}-\Phi_{24} \\ 0 & 0 & s\Theta_{33}-\Phi_{33} & s\Theta_{34}-\Phi_{34} \\ 0 & 0 & 0 & s\Theta_{44}-\Phi_{44} \end{bmatrix} + \tau \\
&\quad = \nu_1+\mu_2+\mu_3+\mu_4+\tau \ge \operatorname{rank}(\mathcal{E})+\operatorname{rank}(\mathcal{D}).
\end{aligned}$$

(ii). Since $\mathcal{E}$ and $\mathcal{D}$ are of full row rank, by part (i) we obtain that

$$\max_{s\in\mathbf{C}} \operatorname{rank}\begin{bmatrix} s\mathcal{E}-\mathcal{A} & \mathcal{B} \\ \mathcal{C} & -\mathcal{D} \end{bmatrix} \ge \operatorname{rank}(\mathcal{E})+\operatorname{rank}(\mathcal{D}) = n+p.$$

But, it is obvious that

$$\max_{s\in\mathbf{C}} \operatorname{rank}\begin{bmatrix} s\mathcal{E}-\mathcal{A} & \mathcal{B} \\ \mathcal{C} & -\mathcal{D} \end{bmatrix} \le n+p.$$

Hence,

$$\max_{s\in\mathbf{C}} \operatorname{rank}\begin{bmatrix} s\mathcal{E}-\mathcal{A} & \mathcal{B} \\ \mathcal{C} & -\mathcal{D} \end{bmatrix} = n+p. \qquad \square$$

The next lemma provides necessary and sufficient conditions for a matrix pencil to be regular and of index at most one.

LEMMA 4 (see [7, 17]). [illegible] $E, A \in \mathbf{R}^{n\times n}$ [illegible]

(i) (E, A)[illegible]

(ii) $\operatorname{rank}\begin{bmatrix} E & AS_\infty(E) \end{bmatrix} = n$ [illegible] $S_\infty(E)$ [illegible] E

(iii) $\deg(\det(sE-A)) = \operatorname{rank}(E)$

3. Main results. The purpose of this section is to present necessary and sufficient solvability conditions as well as a numerically reliable algorithm for the RRDP of descriptor system (1). For this purpose, first we transform the RRDP for descriptor system (1) into the RRDP for a linear time-invariant system using orthogonal transformations.

THEOREM 5. [illegible] (1) [illegible] n_1 n_2 n_3 $\tilde{n}_2$ $\tilde{n}_3$ [illegible] $U, V, Q \in \mathbf{R}^{n\times n}$ [illegible] $W \in \mathbf{R}^{m\times m}$ [illegible]

$$Q = \begin{matrix} n_1 & n_2+n_3 \end{matrix} \begin{matrix} \\ \end{matrix} \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{matrix} \}n_1 \\ \}n_2+n_3 \end{matrix} \tag{8}$$

[illegible] $n_1+n_2+n_3 = n_1+\tilde{n}_2+\tilde{n}_3 = n$ Q_{11} [illegible]

$$\left[\begin{array}{ccc} Q_{11} & Q_{12} & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{array}\right] \left[\begin{array}{cc} U & 0 \\ 0 & I \end{array}\right] \left[\begin{array}{c|cc} sE-A & B & \\ \hline C & 0 & \end{array}\right] \left[\begin{array}{cc} V & 0 \\ 0 & W \end{array}\right]$$

$$= \begin{matrix} n_1 & n_2 & n_3 & \tilde{n}_2 & m-\tilde{n}_2 \end{matrix} \begin{matrix} \\ \end{matrix} \left[\begin{array}{ccc|cc} sE_{11}-A_{11} & -A_{12} & sE_{13}-A_{13} & 0 & B_{12} \\ -A_{21} & -A_{22} & sE_{23}-A_{23} & B_{21} & 0 \\ 0 & 0 & sE_{33}-A_{33} & 0 & 0 \\ \hline C_1 & C_2 & C_3 & 0 & 0 \end{array}\right] \begin{matrix} \}n_1 \\ \}\tilde{n}_2 \\ \}\tilde{n}_3 \\ \\ \}m \end{matrix}, \tag{9}$$

[illegible] E_{11} [illegible] B_{21} [illegible] $sE_{33}-A_{33}$ [illegible] $s \in \mathbf{C}$

[illegible]. The form (9) is constructed in [23]. □

In the following, we give a system interpretation of the form (9).

With respect to the coordinate transformations in the form (9), the system (1) can be expressed as

$$\begin{cases} \left(\left[\begin{array}{cc} Q_{11} & Q_{12} \\ 0 & I \end{array}\right] UEV\right) V^T\dot{x} = \left(\left[\begin{array}{cc} Q_{11} & Q_{12} \\ 0 & I \end{array}\right] UAV\right) V^T x \\ \qquad + \left(\left[\begin{array}{cc} Q_{11} & Q_{12} \\ 0 & I \end{array}\right] UBW\right) W^T u, \\ y = (CV)V^T x, \end{cases} \tag{10}$$

where $V^T x$ represents the transformed state vector and $W^T u$ the transformed input. Let

$$V^T x = \begin{bmatrix} \tilde{x} \\ x_2 \\ x_3 \end{bmatrix} \begin{matrix} \}n_1 \\ \}n_2 \\ \}n_3 \end{matrix}, \quad W^T u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \begin{matrix} \}\tilde{n}_2 \\ \}m-\tilde{n}_2 \end{matrix}. \tag{11}$$

Then system (10) is equivalent to

$$\begin{cases} E_{11}\dot{\tilde{x}} + E_{13}\dot{x}_3 = A_{11}\tilde{x} + A_{12}x_2 + A_{13}x_3 + B_{12}u_2, \\ E_{23}\dot{x}_3 = A_{21}\tilde{x} + A_{22}x_2 + A_{23}x_3 + B_{21}u_1, \\ E_{33}\dot{x}_3 = A_{33}x_3, \\ y = C_1\tilde{x} + C_2x_2 + C_3x_3. \end{cases} \tag{12}$$

Because $sE_{33}-A_{33}$ is of full column rank for any $s \in \mathbf{C}$, according to [22], we know that

$$x_3 = 0 \quad \forall t \geq 0.$$

Consequently, $E_{33}\dot{x}_3 = A_{33}x_3$ is a redundant subsystem (associated with x_3 constrained to be zero). As the redundant subsystem has a zero trajectory $x_3 = 0$, we can delete this part. Therefore, (1) is reduced to

- a regular subsystem (with nonsingular E_{11})

$$\begin{cases} E_{11}\dot{\tilde{x}} = A_{11}\tilde{x} + \begin{bmatrix} B_{12} & A_{12} \end{bmatrix} \begin{bmatrix} u_2 \\ x_2 \end{bmatrix}, \\ y = C_1\tilde{x} + \begin{bmatrix} 0 & C_2 \end{bmatrix} \begin{bmatrix} u_2 \\ x_2 \end{bmatrix}; \end{cases} \tag{13}$$

- an algebraic subsystem (associated with x_2)

$$0 = A_{21}\tilde{x} + A_{22}x_2 + B_{21}u_1. \tag{14}$$

The algebraic part of the system results in the algebraic condition (14), which must be satisfied. This can be taken as an algebraic constraint on the feasibility of the system (1). Since B_{21} is nonsingular, we can always find an input u_1 to ensure that the descriptor system (1) is consistent. If we consider

$$\begin{bmatrix} u_2 \\ x_2 \end{bmatrix} = \tilde{u}$$

as a new input and choose $u_1 = -B_{21}^{-1}(A_{21}\tilde{x} + A_{22}x_2)$, then the regular subsystem (13) becomes

$$\begin{cases} E_{11}\dot{\tilde{x}} = A_{11}\tilde{x} + \begin{bmatrix} B_{12} & A_{12} \end{bmatrix} \tilde{u}, \\ y = C_1\tilde{x} + \begin{bmatrix} 0 & C_2 \end{bmatrix} \tilde{u}, \end{cases} \tag{15}$$

and the algebraic constraint (14) is satisfied. The regular subsystem (15) preserves the finite zeros of the descriptor system (1), as shown in the next corollary.

COROLLARY 6. [illegible] (1), [illegible] (9) [illegible] (1) [illegible] (15).

[illegible]. The finite zeros of systems (1) and (15) are the finite eigenvalues of matrix pencils

$$\begin{bmatrix} A - sE & B \\ C & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} A_{11} - sE_{11} & B_{12} & A_{12} \\ C_1 & 0 & C_2 \end{bmatrix},$$

respectively. By construction, B_{21} is nonsingular and $sE_{33} - A_{33}$ has full column rank for any $s \in \mathbf{C}$. Hence, we obtain by means of the form (9) that matrix pencils

$$\begin{bmatrix} A - sE & B \\ C & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} A_{11} - sE_{11} & B_{12} & A_{12} \\ C_1 & 0 & C_2 \end{bmatrix}$$

have the same finite eigenvalues. Therefore, systems (1) and (15) have the same set of finite zeros. □

After removing the redundant subsystem $E_{33}\dot{x}_3 = A_{33}x_3$ and assuming that the algebraic consistency (14) is satisfied, it is therefore natural to focus on the regular subsystem (15) of the descriptor system (1).

The following lemma shows that the form (9) can be used to characterize the existence of a feedback matrix F such that the pencil $(E, A + BF)$ is regular and of index at most one.

LEMMA 7. [illegible] (1) [illegible] F [illegible] $(E, A + BF)$ [illegible]

$$n_3 = \tilde{n}_3, \quad E_{23} = 0, \quad E_{33} = 0. \tag{16}$$

. For any $F \in \mathbf{R}^{m\times n}$, denote

$$W^T F V =: \begin{array}{c} \\ \\ \end{array} \begin{array}{c} \begin{matrix} n_1 & n_2 & n_3 \end{matrix} \\ \begin{bmatrix} F_{11} & F_{12} & F_{13} \\ F_{21} & F_{22} & F_{23} \end{bmatrix} \end{array} \begin{array}{l} \}\tilde{n}_2 \\ \}m-\tilde{n}_2 \end{array}. \tag{17}$$

We have

$$\begin{aligned} &\begin{bmatrix} Q_{11} & Q_{12} \\ 0 & I \end{bmatrix} U(sE-A-BF)V \\ &= \begin{bmatrix} sE_{11}-A_{11}-B_{12}F_{21} & -A_{12}-B_{12}F_{22} & sE_{13}-A_{13}-B_{12}F_{23} \\ -(A_{21}+B_{21}F_{11}) & -(A_{22}+B_{21}F_{12}) & sE_{23}-(A_{23}+B_{21}F_{13}) \\ 0 & 0 & sE_{33}-A_{33} \end{bmatrix}. \end{aligned} \tag{18}$$

. Let $F \in \mathbf{R}^{m\times n}$ be such that the pencil $(E, A+BF)$ is regular and of index at most one. Then by the regularity of $(E, A+BF)$ we have

$$\max_{s\in\mathbf{C}} \operatorname{rank}(sE-A-BF) = n,$$

which together with (18) yields that

$$\max_{s\in\mathbf{C}} \operatorname{rank}(sE_{33}-A_{33}) = \tilde{n}_3.$$

Note that $sE_{33}-A_{33}$ is of full column rank for any $s \in \mathbf{C}$. Thus,

$$\tilde{n}_3 = n_3. \tag{19}$$

Since the pencil $(E, A+BF)$ is regular and of index at most one, $sE_{33}-A_{33}$ is of full column rank for any $s \in \mathbf{C}$, and we have using Lemma 4(iii) and (19) that

$$\begin{aligned} \operatorname{rank}(E) &= \deg(\det(sE-A-BF)) \\ &= \deg\left(\det\left(\begin{bmatrix} sE_{11}-A_{11}-B_{12}F_{21} & -A_{12}-B_{12}F_{22} \\ -(A_{21}+B_{21}F_{11}) & -(A_{22}+B_{21}F_{12}) \end{bmatrix}\right)\right) \\ &\quad + \deg(\det(sE_{33}-A_{33})) \\ &= \deg\left(\det\left(\begin{bmatrix} sE_{11}-A_{11}-B_{12}F_{21} & -A_{12}-B_{12}F_{22} \\ -(A_{21}+B_{21}F_{11}) & -(A_{22}+B_{21}F_{12}) \end{bmatrix}\right)\right) \\ &\le \operatorname{rank}(E_{11}) \\ &= n_1. \end{aligned} \tag{20}$$

Because $E_{11} \in \mathbf{R}^{n_1\times n_1}$ is nonsingular, we also have

$$\operatorname{rank}(E) = \operatorname{rank}(E_{11}) + \operatorname{rank}\begin{bmatrix} E_{23} \\ E_{33} \end{bmatrix} = n_1 + \operatorname{rank}\begin{bmatrix} E_{23} \\ E_{33} \end{bmatrix}. \tag{21}$$

Hence, we obtain

$$E_{23} = 0, \quad E_{33} = 0, \tag{22}$$

which together with (19) give the condition (16).

. Assume condition (16) holds. It follows that $\tilde{n}_2 = n_2$ and $\operatorname{rank}(E) = n_1$. Let

$$F_{12} = -B_{21}^{-1}(A_{22}+I), \quad F_{11}, F_{13}, F_{21}, F_{22}, \text{ and } F_{23} \text{ are arbitrary.} \tag{23}$$

By Lemma 4(ii), we know that $(E, A+BF)$ is regular and of index at most one. ∎

In Corollary 6, it is shown that systems (1) and (15) have the same set of finite zeros. In the next result it will be shown that the RRDP for descriptor system (1) can be reduced to the RRDP for the linear time-invariant system (15).

THEOREM 8. [illegible] (1) [illegible] (1) [illegible] (16) [illegible]

(24) the RRDP for system (15) is solvable;

[illegible] $\mathcal{F}$ [illegible] $\mathcal{H}$ [illegible] $\mathcal{H}$ [illegible]

$$\mathcal{T}_{\mathcal{F},\mathcal{H}}(s) = ((C_1 + \begin{bmatrix} 0 & C_2 \end{bmatrix} \mathcal{F})(sE_{11} - A_{11} - \begin{bmatrix} B_{12} & A_{12} \end{bmatrix} \mathcal{F})^{-1} \begin{bmatrix} B_{12} & A_{12} \end{bmatrix} + \begin{bmatrix} 0 & C_2 \end{bmatrix})\mathcal{H} \tag{25}$$

[illegible]. For any $F \in \mathbf{R}^{m\times n}$ and $H \in \mathbf{R}^{m\times m}$, denote $W^T F V$ as in (17), and let

$$\tilde{H} = W^T H. \tag{26}$$

Clearly, if H is nonsingular, then $\tilde{H}$ is nonsingular. We will first prove "necessity" and then "sufficiency."

[illegible]. Let $F \in \mathbf{R}^{m\times n}$ and $H \in \mathbf{R}^{m\times m}$ with H nonsingular be such that the pencil $(E, A + BF)$ is regular, of index at most one, and (4) is true. Then condition (16) of the theorem follows directly from Lemma 7. Note that the condition (16) implies that

$$\tilde{n}_2 = n_2, \quad \operatorname{rank}(E) = n_1. \tag{27}$$

We have shown in the proof of Lemma 7 that (20) holds, from which it follows that the pencil

$$\left(\begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} + B_{12}F_{21} & A_{12} + B_{12}F_{22} \\ A_{21} + B_{21}F_{11} & A_{22} + B_{21}F_{12} \end{bmatrix} \right)$$

is regular and of index at most one. Hence, by Lemma 4(ii), we have that $A_{22}+B_{21}F_{12}$ is nonsingular. Now a simple calculation yields that

$$\begin{aligned} & C(sE - A - BF)^{-1}BH \\ &= \begin{bmatrix} C_1 & C_2 & C_3 \end{bmatrix} \begin{bmatrix} sE_{11} - A_{11} - B_{12}F_{21} & -A_{12} - B_{12}F_{22} & sE_{13} - A_{13} - B_{12}F_{23} \\ -(A_{21} + B_{21}F_{11}) & -(A_{22} + B_{21}F_{12}) & -(A_{23} + B_{21}F_{13}) \\ 0 & 0 & -A_{33} \end{bmatrix}^{-1} \\ & \quad \times \begin{bmatrix} 0 & B_{12} \\ B_{21} & 0 \\ 0 & 0 \end{bmatrix} \tilde{H} \\ &= \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} sE_{11} - A_{11} - B_{12}F_{21} & -A_{12} - B_{12}F_{22} \\ -(A_{21} + B_{21}F_{11}) & -(A_{22} + B_{21}F_{12}) \end{bmatrix}^{-1} \begin{bmatrix} 0 & B_{12} \\ B_{21} & 0 \end{bmatrix} \tilde{H} \\ &= \mathcal{T}_{\mathcal{F},\mathcal{H}}(s), \end{aligned} \tag{28}$$

where

$$\mathcal{F} = \begin{bmatrix} F_{21} - F_{22}(A_{22} + B_{21}F_{12})^{-1}(A_{21} + B_{21}F_{11}) \\ -(A_{22} + B_{21}F_{12})^{-1}(A_{21} + B_{21}F_{11}) \end{bmatrix}, \tag{29}$$

$$\mathcal{H} = \begin{bmatrix} -F_{22}(A_{22} + B_{21}F_{12})^{-1} & I \\ -(A_{22} + B_{21}F_{12})^{-1} & 0 \end{bmatrix} \begin{bmatrix} B_{21} & 0 \\ 0 & I \end{bmatrix} \tilde{H}. \tag{30}$$

Since $\tilde{H}$, B_{21}, and $A_{22}+B_{21}F_{12}$ are nonsingular, so is $\mathcal{H}$. Hence, the nonsingularity and diagonality of $C(sE-A-BF)^{-1}BH$ imply that $\mathcal{T}_{\mathcal{F},\mathcal{H}}(s)$ is nonsingular and diagonal. This is equivalent to the solvability of the RRDP for system (15).

. We will prove the sufficiency constructively. Assume that conditions (16) and (24) hold. Condition (16) implies that $n_2=\tilde{n}_2$, and so the system (15) is square. From the condition (24) there are matrices $\mathcal{F}$ and $\mathcal{H}$ with $\mathcal{H}$ nonsingular and

$$\mathcal{F}=\begin{bmatrix}\mathcal{F}_1\\ \mathcal{F}_2\end{bmatrix}\begin{matrix}\}m-n_2\\ \}n_2\end{matrix}$$

such that $\mathcal{T}_{\mathcal{F},\mathcal{H}}(s)$ is nonsingular and diagonal. Let (F,H) be determined by

$$\begin{cases}\begin{bmatrix}B_{21}&0\\0&I\end{bmatrix}W^TF=\begin{bmatrix}\mathcal{F}_2-A_{21}&-I-A_{22}&0\\ \mathcal{F}_1&0&0\end{bmatrix}V^T,\\ \begin{bmatrix}B_{21}&0\\0&I\end{bmatrix}W^TH=\begin{bmatrix}0&I_{n_2}\\ I_{m-n_2}&0\end{bmatrix}\mathcal{H},\end{cases}\tag{31}$$

partition F as in (17), and define $\tilde{H}$ by (26). We have that $A_{22}+B_{21}F_{12}=-I$ and (29) and (30) hold. By condition (16) and the proof of the sufficiency of Lemma 7, $(E,A+BF)$ is regular and of index at most one. Moreover, (28) yields that $C(sE-A-BF)^{-1}BH$ is nonsingular and diagonal. □

In general, $\begin{bmatrix}0&C_2\end{bmatrix}\neq 0$, and, consequently, Theorem 1 cannot be extended to system (15). Hence, we reduce the RRDP for system (15) to the one for a system of the form (5) via the following factorization.

THEOREM 9. (1) (9) $\mathcal{U},\mathcal{V}$ $\mathcal{W}$ $\mathcal{P}$

$$\begin{bmatrix}\mathcal{U}&0\\0&\mathcal{P}\end{bmatrix}\left[\begin{array}{c|cc}sE_{11}-A_{11}&B_{12}&A_{12}\\ \hline C_1&0&C_2\end{array}\right]\begin{bmatrix}\mathcal{V}&0\\0&\mathcal{W}\end{bmatrix}$$

$$=\begin{array}{c}\begin{matrix}\mu_1&\mu_2&\nu&m+n_2-\tilde{n}_2-\nu\end{matrix}\\ \left[\begin{array}{cc|cc}s\mathcal{E}_{11}-\mathcal{A}_{11}&s\mathcal{E}_{12}-\mathcal{A}_{12}&\mathcal{B}_{11}&\mathcal{B}_{12}\\ -\mathcal{A}_{21}&s\mathcal{E}_{22}-\mathcal{A}_{22}&\mathcal{B}_{21}&\mathcal{B}_{22}\\ 0&s\mathcal{E}_{32}-\mathcal{A}_{32}&0&0\\ \hline \mathcal{C}_{11}&\mathcal{C}_{12}&\mathcal{D}_{11}&0\\ 0&\mathcal{C}_{22}&\mathcal{D}_{21}&0\end{array}\right]\begin{matrix}\}\mu_1\\ \}\tau_2\\ \}\tau_3\\ \\ \}\nu\\ \}m-\nu\end{matrix}\end{array},\tag{32}$$

$\mu_1+\mu_2=\mu_1+\tau_2+\tau_3=n_1$ $\mathcal{E}_{11}$ $\mathcal{D}_{11}$ $\begin{bmatrix}\mathcal{B}_{21}&\mathcal{B}_{22}\end{bmatrix}$

$$\operatorname{rank}\begin{bmatrix}s\mathcal{E}_{11}-\mathcal{A}_{11}&\mathcal{B}_{11}&\mathcal{B}_{12}\\ -\mathcal{A}_{21}&\mathcal{B}_{21}&\mathcal{B}_{22}\end{bmatrix}=\mu_1+\tau_2\quad\forall s\in\mathbf{C},\tag{33}$$

$$\max_{s\in\mathbf{C}}\operatorname{rank}\begin{bmatrix}s\mathcal{E}_{32}-\mathcal{A}_{32}\\ \mathcal{C}_{22}\end{bmatrix}=\mu_2.\tag{34}$$

. The forms (32) are constructed in [23]. □

Let

$$\mathcal{V}^T\tilde{x} = \begin{bmatrix} x_{11} \\ x_{21} \end{bmatrix} \begin{matrix} \}\mu_1 \\ \}\mu_2 \end{matrix}, \quad \mathcal{W}^T\tilde{u} = \begin{bmatrix} u_{11} \\ u_{21} \end{bmatrix} \begin{matrix} \}\nu \\ \}m+n_2-\tilde{n}_2-\nu \end{matrix}, \quad \mathcal{P}y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \begin{matrix} \}\nu \\ \}m-\nu \end{matrix}. \tag{35}$$

Then system (15) is equivalent to

$$\begin{cases} \mathcal{E}_{11}\dot{x}_{11} + \mathcal{E}_{12}\dot{x}_{21} = \mathcal{A}_{11}x_{11} + \mathcal{A}_{12}x_{21} + \mathcal{B}_{11}u_{11} + \mathcal{B}_{12}u_{21}, \\ \begin{bmatrix} \mathcal{E}_{22} \\ \mathcal{E}_{32} \end{bmatrix}\dot{x}_{21} = \begin{bmatrix} \mathcal{A}_{22} \\ \mathcal{A}_{32} \end{bmatrix}x_{21} + \begin{bmatrix} \mathcal{A}_{21} & \mathcal{B}_{21} \\ 0 & 0 \end{bmatrix}\begin{bmatrix} x_{11} \\ u_{11} \end{bmatrix} + \begin{bmatrix} \mathcal{B}_{22} \\ 0 \end{bmatrix}u_{21}, \\ y_1 = \mathcal{C}_{11}x_{11} + \mathcal{C}_{12}x_{21} + \mathcal{D}_{11}u_{11}, \\ y_2 = \mathcal{C}_{22}x_{21} + \mathcal{D}_{21}u_{11}. \end{cases} \tag{36}$$

Now, the nonsingularity of E_{11} implies that

$$\begin{bmatrix} \mathcal{E}_{22} \\ \mathcal{E}_{32} \end{bmatrix}$$

is nonsingular. So, if $\mathcal{B}_{22}$ is nonsingular, we take $u_{11} = 0$, and we denote $v_{21} = u_{21} + \mathcal{B}_{22}^{-1}\mathcal{A}_{21}x_{11}$, then the system (36) becomes

$$\begin{cases} \begin{bmatrix} \mathcal{E}_{22} \\ \mathcal{E}_{32} \end{bmatrix}\dot{x}_{21} = \begin{bmatrix} \mathcal{A}_{22} \\ \mathcal{A}_{32} \end{bmatrix}x_{21} + \begin{bmatrix} \mathcal{B}_{22} \\ 0 \end{bmatrix}v_{21}, \\ y_2 = \mathcal{C}_{22}x_{21} \end{cases} \tag{37}$$

and

$$\begin{cases} \mathcal{E}_{11}\dot{x}_{11} = (\mathcal{A}_{11} - \mathcal{B}_{12}\mathcal{B}_{22}^{-1}\mathcal{A}_{21})x_{11} + \left(\mathcal{A}_{12} - \mathcal{E}_{12}\begin{bmatrix} \mathcal{E}_{22} \\ \mathcal{E}_{32} \end{bmatrix}^{-1}\begin{bmatrix} \mathcal{A}_{22} \\ \mathcal{A}_{32} \end{bmatrix}\right)x_{21} \\ \qquad\qquad + \left(\mathcal{B}_{12} - \mathcal{E}_{12}\begin{bmatrix} \mathcal{E}_{22} \\ \mathcal{E}_{32} \end{bmatrix}^{-1}\begin{bmatrix} \mathcal{B}_{22} \\ 0 \end{bmatrix}\right)v_{21}, \\ y_1 = \mathcal{C}_{11}x_{11} + \mathcal{C}_{12}x_{21}. \end{cases}$$

The main feature of system (37) is that it is a linear time-invariant system of the form (5) because

$$\begin{bmatrix} \mathcal{E}_{22} \\ \mathcal{E}_{32} \end{bmatrix}$$

is nonsingular.

Remark 1. In the descriptions above, we set $u_{11} = 0$ to motivate how the system (37) may be deduced from the system (15) by a particular choice of the input. Hence, if the RRDP for the system (15) is solvable, so must the RRDP for the system (37) (i.e., the latter is a necessary condition for the former). However, whether the RRDP for the system (15) is solvable or not should not depend on the choice of the input. Indeed, in Theorem 10, we show that a necessary and sufficient condition for the solvability of the RRDP for the system (15) can be expressed in terms of the solvability of the RRDP for the system (37) independent of the choice of the input, and an examination of the proof of Theorem 10 will reveal that we never make use of $u_{11} = 0$ in the proof. To summarize, the descriptions above are just a simple argument intended to introduce the form of the system (37) for easy reference in Theorem 10;

otherwise, the choice $u_{11} = 0$ is not used in the rigorous proof of the equivalence between the RRDPs for the systems (15) and (37).

The following theorem shows that we can reduce the RRDP for system (15) further to the RRDP for system (37).

THEOREM 10. (1) (9) (32) (16) (15)

(38) $$\mathcal{D}_{21} = 0, \quad \mathcal{B}_{22} \text{ is nonsingular,}$$

(39) the RRDP for system (37) is solvable.

. Since $n_3 = \tilde{n}_3$ in the form (9), $n_2 = \tilde{n}_2$ and

(40) $$m + n_2 - \tilde{n}_2 - \nu = m - \nu, \quad \mathcal{B}_{12} \in \mathbf{R}^{\mu_1 \times (m-\nu)}, \quad \mathcal{B}_{22} \in \mathbf{R}^{\tau_2 \times (m-\nu)}.$$

. Assume that $\mathcal{F}$ and $\mathcal{H}$ with $\mathcal{H}$ nonsingular solve the RRDP of system (15), $\mathcal{T}_{\mathcal{F},\mathcal{H}}(s)$ defined by (25) being diagonal and nonsingular. Let

(41) $$\mathcal{W}^T\mathcal{F}\mathcal{V} = \begin{matrix} \\ \nu \\ m-\nu \end{matrix} \overset{\begin{matrix} \mu_1 & \mu_2 \end{matrix}}{\begin{bmatrix} \mathcal{F}_{11} & \mathcal{F}_{12} \\ \mathcal{F}_{21} & \mathcal{F}_{22} \end{bmatrix}}, \quad \mathcal{W}^T\mathcal{H}\mathcal{P}^T = \begin{matrix} \\ \nu \\ m-\nu \end{matrix} \overset{\begin{matrix} \nu & m-\nu \end{matrix}}{\begin{bmatrix} \mathcal{H}_{11} & \mathcal{H}_{12} \\ \mathcal{H}_{21} & \mathcal{H}_{22} \end{bmatrix}}.$$

Since $\mathcal{P}$ is a permutation matrix, $\mathcal{P}\mathcal{T}_{\mathcal{F},\mathcal{H}}(s)\mathcal{P}^T$ is also diagonal and nonsingular, or, equivalently,

$$\begin{aligned} &\begin{bmatrix} \mathcal{T}_{11}(s) & \mathcal{T}_{12}(s) \\ \mathcal{T}_{21}(s) & \mathcal{T}_{22}(s) \end{bmatrix} \\ &= \left(\begin{bmatrix} \mathcal{D}_{11} & 0 \\ \mathcal{D}_{21} & 0 \end{bmatrix} + \begin{bmatrix} \mathcal{C}_{11} + \mathcal{D}_{11}\mathcal{F}_{11} & \mathcal{C}_{12} + \mathcal{D}_{11}\mathcal{F}_{12} \\ \mathcal{D}_{21}\mathcal{F}_{11} & \mathcal{C}_{22} + \mathcal{D}_{21}\mathcal{F}_{12} \end{bmatrix} \right. \\ &\qquad \left. \times (\mathcal{U}(sE_{11} - A_{11} - \begin{bmatrix} B_{12} & A_{12} \end{bmatrix}\mathcal{F})\mathcal{V})^{-1} \begin{bmatrix} \mathcal{B}_{11} & \mathcal{B}_{12} \\ \mathcal{B}_{21} & \mathcal{B}_{22} \\ 0 & 0 \end{bmatrix} \right) \begin{bmatrix} \mathcal{H}_{11} & \mathcal{H}_{12} \\ \mathcal{H}_{21} & \mathcal{H}_{22} \end{bmatrix} \\ &= \mathcal{P}\mathcal{T}_{\mathcal{F},\mathcal{H}}(s)\mathcal{P}^T \end{aligned}$$

is diagonal and nonsingular; here

$$\mathcal{T}_{11}(s) \in \mathbf{R}^{\nu \times \nu}, \ \mathcal{T}_{12}(s) \in \mathbf{R}^{\nu \times (m-\nu)}, \ \mathcal{T}_{21}(s) \in \mathbf{R}^{(m-\nu) \times \nu}, \ \mathcal{T}_{22}(s) \in \mathbf{R}^{(m-\nu) \times (m-\nu)},$$

so we get

$$\mathcal{T}_{11}(s) \text{ and } \mathcal{T}_{22}(s) \text{ are diagonal and nonsingular, } \mathcal{T}_{12}(s) = 0, \ \mathcal{T}_{21}(s) = 0.$$

Hence, we have

$$\begin{aligned} &\mathcal{D}_{11}\mathcal{H}_{12} + \begin{bmatrix} \mathcal{C}_{11} + \mathcal{D}_{11}\mathcal{F}_{11} & \mathcal{C}_{12} + \mathcal{D}_{11}\mathcal{F}_{12} \end{bmatrix} (\mathcal{U}(sE_{11} - A_{11} - \begin{bmatrix} B_{12} & A_{12} \end{bmatrix}\mathcal{F})\mathcal{V})^{-1} \\ &\quad \times \begin{bmatrix} \mathcal{B}_{11}\mathcal{H}_{12} + \mathcal{B}_{12}\mathcal{H}_{22} \\ \mathcal{B}_{21}\mathcal{H}_{12} + \mathcal{B}_{22}\mathcal{H}_{22} \\ 0 \end{bmatrix} \\ &\qquad = \mathcal{T}_{12}(s) = 0, \end{aligned} \tag{42}$$

$$\mathcal{D}_{21}\mathcal{H}_{11} + \begin{bmatrix} \mathcal{D}_{21}\mathcal{F}_{11} & \mathcal{C}_{22} + \mathcal{D}_{21}\mathcal{F}_{12} \end{bmatrix} (\mathcal{U}(sE_{11} - A_{11} - \begin{bmatrix} B_{12} & A_{12} \end{bmatrix} \mathcal{F})\mathcal{V})^{-1} \times \begin{bmatrix} \mathcal{B}_{11}\mathcal{H}_{11} + \mathcal{B}_{12}\mathcal{H}_{21} \\ \mathcal{B}_{21}\mathcal{H}_{11} + \mathcal{B}_{22}\mathcal{H}_{21} \\ 0 \end{bmatrix}$$

$$= \mathcal{T}_{21}(s) = 0. \tag{43}$$

Thus, Lemma 2(i) gives that

$$\mathcal{D}_{11}\mathcal{H}_{12} = 0, \quad \mathcal{D}_{21}\mathcal{H}_{11} = 0. \tag{44}$$

Since $\mathcal{D}_{11}$ and

$$\begin{bmatrix} \mathcal{H}_{11} & \mathcal{H}_{12} \\ \mathcal{H}_{21} & \mathcal{H}_{22} \end{bmatrix}$$

are nonsingular, we have

$$\mathcal{H}_{12} = 0, \quad \mathcal{D}_{21} = 0, \quad \mathcal{H}_{11} \text{ and } \mathcal{H}_{22} \text{ are nonsingular.} \tag{45}$$

Using (45), (43) becomes

$$\begin{bmatrix} 0 & \mathcal{C}_{22} \end{bmatrix} (\mathcal{U}(sE_{11} - A_{11} - \begin{bmatrix} B_{12} & A_{12} \end{bmatrix} \mathcal{F})\mathcal{V})^{-1} \begin{bmatrix} \mathcal{B}_{11} + \mathcal{B}_{12}\mathcal{H}_{21}\mathcal{H}_{11}^{-1} \\ \mathcal{B}_{21} + \mathcal{B}_{22}\mathcal{H}_{21}\mathcal{H}_{11}^{-1} \\ 0 \end{bmatrix} = 0. \tag{46}$$

By Lemma 2(i) and properties (34) and (46), we get

$$n_1 = \mu_1 + \mu_2 = \max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} \mathcal{U}(sE_{11} - A_{11} - \begin{bmatrix} B_{12} & A_{12} \end{bmatrix} \mathcal{F})\mathcal{V} \begin{bmatrix} \mathcal{B}_{11} + \mathcal{B}_{12}\mathcal{H}_{21}\mathcal{H}_{11}^{-1} \\ \mathcal{B}_{21} + \mathcal{B}_{22}\mathcal{H}_{21}\mathcal{H}_{11}^{-1} \\ 0 \end{bmatrix} \\ \begin{bmatrix} 0 & \mathcal{C}_{22} \end{bmatrix} \qquad\qquad 0 \end{bmatrix}$$

$$= \max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} s\mathcal{E}_{11} - \mathcal{A}_{11} - \mathcal{B}_{11}\mathcal{F}_{11} - \mathcal{B}_{12}\mathcal{F}_{21} & s\mathcal{E}_{12} - \mathcal{A}_{12} - \mathcal{B}_{11}\mathcal{F}_{12} - \mathcal{B}_{12}\mathcal{F}_{22} & \mathcal{B}_{11} + \mathcal{B}_{12}\mathcal{H}_{21}\mathcal{H}_{11}^{-1} \\ -\mathcal{A}_{21} - \mathcal{B}_{21}\mathcal{F}_{11} - \mathcal{B}_{22}\mathcal{F}_{21} & s\mathcal{E}_{22} - \mathcal{A}_{22} - \mathcal{B}_{21}\mathcal{F}_{12} - \mathcal{B}_{22}\mathcal{F}_{22} & \mathcal{B}_{21} + \mathcal{B}_{22}\mathcal{H}_{21}\mathcal{H}_{11}^{-1} \\ 0 & s\mathcal{E}_{32} - \mathcal{A}_{32} & 0 \\ 0 & \mathcal{C}_{22} & 0 \end{bmatrix}$$

$$= \max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} s\mathcal{E}_{11} - \mathcal{A}_{11} - \mathcal{B}_{11}\mathcal{F}_{11} - \mathcal{B}_{12}\mathcal{F}_{21} & \mathcal{B}_{11} + \mathcal{B}_{12}\mathcal{H}_{21}\mathcal{H}_{11}^{-1} \\ -\mathcal{A}_{21} - \mathcal{B}_{21}\mathcal{F}_{11} - \mathcal{B}_{22}\mathcal{F}_{21} & \mathcal{B}_{21} + \mathcal{B}_{22}\mathcal{H}_{21}\mathcal{H}_{11}^{-1} \end{bmatrix} + \mu_2$$

$$\geq \operatorname{rank}(\mathcal{E}_{11}) + \operatorname{rank}(\mathcal{B}_{21} + \mathcal{B}_{22}\mathcal{H}_{21}\mathcal{H}_{11}^{-1}) + \mu_2 \qquad \text{(by Lemma 3(i))}$$

$$= \mu_1 + \mu_2 + \operatorname{rank}(\mathcal{B}_{21} + \mathcal{B}_{22}\mathcal{H}_{21}\mathcal{H}_{11}^{-1}).$$

(47)

Thus,

$$\mathcal{B}_{21} + \mathcal{B}_{22}\mathcal{H}_{21}\mathcal{H}_{11}^{-1} = 0, \tag{48}$$

which implies that

$$\operatorname{rank}(\mathcal{B}_{22}) = \operatorname{rank} \begin{bmatrix} \mathcal{B}_{21} & \mathcal{B}_{22} \end{bmatrix}. \tag{49}$$

By construction, $\begin{bmatrix} \mathcal{B}_{21} & \mathcal{B}_{22} \end{bmatrix}$ is of full row rank ($= \tau_2$). Hence,

$$\operatorname{rank}(\mathcal{B}_{22}) = \tau_2. \tag{50}$$

Note that $\mathcal{E}_{11}$ is nonsingular and $\mathcal{B}_{22}$ is of full row rank (see (50)); Lemma 3(ii) yields that

$$\max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} s\mathcal{E}_{11} - \mathcal{A}_{11} - \mathcal{B}_{11}\mathcal{F}_{11} - \mathcal{B}_{12}\mathcal{F}_{21} & \mathcal{B}_{12} \\ -\mathcal{A}_{21} - \mathcal{B}_{21}\mathcal{F}_{11} - \mathcal{B}_{22}\mathcal{F}_{21} & \mathcal{B}_{22} \end{bmatrix} = \mu_1 + \tau_2. \tag{51}$$

Since

$$\mathcal{T}_{22}(s) = \begin{bmatrix} 0 & \mathcal{C}_{22} \end{bmatrix} (\mathcal{U}(sE_{11} - A_{11} - \begin{bmatrix} B_{12} & A_{12} \end{bmatrix} \mathcal{F})\mathcal{V})^{-1} \begin{bmatrix} \mathcal{B}_{12} \\ \mathcal{B}_{22} \\ 0 \end{bmatrix} \mathcal{H}_{22} \in \mathbf{R}^{(m-\nu)\times(m-\nu)}$$

is diagonal and nonsingular, and $E_{11} \in \mathbf{R}^{n_1\times n_1}$, we have using (34), (50), (51), and the nonsingularity of $\mathcal{H}_{22}$ and $\mathcal{E}_{11}$ that

$$\begin{aligned}
&n_1 + (m-\nu) = \mu_1 + \mu_2 + (m-\nu) \\
&= \max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} \mathcal{U}(sE_{11} - A_{11} - \begin{bmatrix} B_{12} & A_{12} \end{bmatrix} \mathcal{F})\mathcal{V} & \begin{bmatrix} \mathcal{B}_{12} \\ \mathcal{B}_{22} \\ 0 \end{bmatrix} \mathcal{H}_{22} \\ \begin{bmatrix} 0 & \mathcal{C}_{22} \end{bmatrix} & 0 \end{bmatrix} \\
&= \max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} s\mathcal{E}_{11} - \mathcal{A}_{11} - \mathcal{B}_{11}\mathcal{F}_{11} - \mathcal{B}_{12}\mathcal{F}_{21} & s\mathcal{E}_{12} - \mathcal{A}_{12} - \mathcal{B}_{11}\mathcal{F}_{12} - \mathcal{B}_{12}\mathcal{F}_{22} & \mathcal{B}_{12}\mathcal{H}_{22} \\ -\mathcal{A}_{21} - \mathcal{B}_{21}\mathcal{F}_{11} - \mathcal{B}_{22}\mathcal{F}_{21} & s\mathcal{E}_{22} - \mathcal{A}_{22} - \mathcal{B}_{21}\mathcal{F}_{12} - \mathcal{B}_{22}\mathcal{F}_{22} & \mathcal{B}_{22}\mathcal{H}_{22} \\ 0 & s\mathcal{E}_{32} - \mathcal{A}_{32} & 0 \\ 0 & \mathcal{C}_{22} & 0 \end{bmatrix} \\
&= \max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} s\mathcal{E}_{11} - \mathcal{A}_{11} - \mathcal{B}_{11}\mathcal{F}_{11} - \mathcal{B}_{12}\mathcal{F}_{21} & \mathcal{B}_{12} \\ -\mathcal{A}_{21} - \mathcal{B}_{21}\mathcal{F}_{11} - \mathcal{B}_{22}\mathcal{F}_{21} & \mathcal{B}_{22} \end{bmatrix} + \mu_2 \\
&= (\mu_1 + \tau_2) + \mu_2 \qquad \text{(by Lemma 3(ii))}.
\end{aligned}$$

Hence,

$$\tau_2 = m - \nu. \tag{52}$$

Since $\mathcal{B}_{22} \in \mathbf{R}^{\tau_2\times(m-\nu)}$ (see (40)), it follows from (50) and (52) that $\mathcal{B}_{22}$ is nonsingular, which together with (45) proves that condition (38) of the theorem holds.

Using the nonsingularity of $\mathcal{B}_{22}$, we have from (48) that

$$\mathcal{H}_{21}\mathcal{H}_{11}^{-1} = -\mathcal{B}_{22}^{-1}\mathcal{B}_{21}, \tag{53}$$

which gives that

$$\mathcal{B}_{11} + \mathcal{B}_{12}\mathcal{H}_{21}\mathcal{H}_{11}^{-1} = \mathcal{B}_{11} - \mathcal{B}_{12}\mathcal{B}_{22}^{-1}\mathcal{B}_{21}. \tag{54}$$

From (47), (48), and (54), we obtain

$$\begin{aligned}
\mu_1 &= \max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} s\mathcal{E}_{11} - \mathcal{A}_{11} - \mathcal{B}_{11}\mathcal{F}_{11} - \mathcal{B}_{12}\mathcal{F}_{21} & \mathcal{B}_{11} - \mathcal{B}_{12}\mathcal{B}_{22}^{-1}\mathcal{B}_{21} \\ -\mathcal{A}_{21} - \mathcal{B}_{21}\mathcal{F}_{11} - \mathcal{B}_{22}\mathcal{F}_{21} & 0 \end{bmatrix} \\
&= \max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} s\mathcal{E}_{11} - \mathcal{A}_{11} - \mathcal{B}_{11}\mathcal{F}_{11} - \mathcal{B}_{12}\mathcal{F}_{21} - \mathcal{B}_{12}\mathcal{B}_{22}^{-1}(-\mathcal{A}_{21} - \mathcal{B}_{21}\mathcal{F}_{11} - \mathcal{B}_{22}\mathcal{F}_{21}) & \mathcal{B}_{11} - \mathcal{B}_{12}\mathcal{B}_{22}^{-1}\mathcal{B}_{21} \\ -\mathcal{A}_{21} - \mathcal{B}_{21}\mathcal{F}_{11} - \mathcal{B}_{22}\mathcal{F}_{21} & 0 \end{bmatrix} \\
&= \max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} s\mathcal{E}_{11} - \mathcal{A}_{11} + \mathcal{B}_{12}\mathcal{B}_{22}^{-1}\mathcal{A}_{21} - (\mathcal{B}_{11} - \mathcal{B}_{12}\mathcal{B}_{22}^{-1}\mathcal{B}_{21})\mathcal{F}_{11} & \mathcal{B}_{11} - \mathcal{B}_{12}\mathcal{B}_{22}^{-1}\mathcal{B}_{21} \\ -\mathcal{A}_{21} - \mathcal{B}_{21}\mathcal{F}_{11} - \mathcal{B}_{22}\mathcal{F}_{21} & 0 \end{bmatrix}.
\end{aligned}$$

Consequently, we have

$$\mu_1 = \max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} s\mathcal{E}_{11} - \mathcal{A}_{11} + \mathcal{B}_{12}\mathcal{B}_{22}^{-1}\mathcal{A}_{21} & \mathcal{B}_{11} - \mathcal{B}_{12}\mathcal{B}_{22}^{-1}\mathcal{B}_{21} \\ -\mathcal{A}_{21} - \mathcal{B}_{21}\mathcal{F}_{11} - \mathcal{B}_{22}\mathcal{F}_{21} & 0 \end{bmatrix}. \tag{55}$$

Because (33) holds and $\mathcal{B}_{22}$ is nonsingular,

$$\begin{aligned} \mu_1 + \tau_2 &= \operatorname{rank} \begin{bmatrix} s\mathcal{E}_{11} - \mathcal{A}_{11} & \mathcal{B}_{11} & \mathcal{B}_{12} \\ -\mathcal{A}_{21} & \mathcal{B}_{21} & \mathcal{B}_{22} \end{bmatrix} \\ &= \operatorname{rank} \begin{bmatrix} s\mathcal{E}_{11} - \mathcal{A}_{11} + \mathcal{B}_{12}\mathcal{B}_{22}^{-1}\mathcal{A}_{21} & \mathcal{B}_{11} - \mathcal{B}_{12}\mathcal{B}_{22}^{-1}\mathcal{B}_{21} & 0 \\ -\mathcal{A}_{21} & \mathcal{B}_{21} & \mathcal{B}_{22} \end{bmatrix} \quad \forall s \in \mathbf{C}, \end{aligned}$$

which gives that

$$\operatorname{rank} \begin{bmatrix} s\mathcal{E}_{11} - \mathcal{A}_{11} + \mathcal{B}_{12}\mathcal{B}_{22}^{-1}\mathcal{A}_{21} & \mathcal{B}_{11} - \mathcal{B}_{12}\mathcal{B}_{22}^{-1}\mathcal{B}_{21} \end{bmatrix} = \mu_1 \quad \forall s \in \mathbf{C}. \tag{56}$$

Thus, by applying Lemma 2(ii) to (55) we get

$$\mathcal{A}_{21} + \mathcal{B}_{21}\mathcal{F}_{11} + \mathcal{B}_{22}\mathcal{F}_{21} = 0. \tag{57}$$

Since $\mathcal{B}_{22}$ is nonsingular, $\mathcal{B}_{21}\mathcal{F}_{12} + \mathcal{B}_{22}\mathcal{F}_{22} = \mathcal{B}_{22}\tilde{\mathcal{F}}_{22}$ with $\tilde{\mathcal{F}}_{22} = \mathcal{B}_{22}^{-1}\mathcal{B}_{21}\mathcal{F}_{12} + \mathcal{F}_{22}$, and

$$\begin{aligned} &\mathcal{C}_{22} \begin{bmatrix} s\mathcal{E}_{22} - \mathcal{A}_{22} - \mathcal{B}_{22}\tilde{\mathcal{F}}_{22} \\ s\mathcal{E}_{32} - \mathcal{A}_{32} \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{B}_{22} \\ 0 \end{bmatrix} \mathcal{H}_{22} \\ &\quad = \begin{bmatrix} 0 & \mathcal{C}_{22} \end{bmatrix} (\mathcal{U}(sE_{11} - A_{11} - \begin{bmatrix} B_{12} & A_{12} \end{bmatrix} \mathcal{F})\mathcal{V})^{-1} \begin{bmatrix} \mathcal{B}_{12} \\ \mathcal{B}_{22} \\ 0 \end{bmatrix} \mathcal{H}_{22} \end{aligned}$$

is diagonal and nonsingular. Hence, the condition (39) of the theorem follows.

. We will prove the sufficiency constructively. Assume that conditions (38) and (39) hold. Since $n_3 = \tilde{n}_3$ implies $n_2 = \tilde{n}_2$, the system (37) is square. From the condition (39) there are matrices $\tilde{\mathcal{F}}_{22}$ and $\mathcal{H}_{22}$ such that

$$\mathcal{T}_{22}(s) = \mathcal{C}_{22} \begin{bmatrix} s\mathcal{E}_{22} - \mathcal{A}_{22} - \mathcal{B}_{22}\tilde{\mathcal{F}}_{22} \\ s\mathcal{E}_{32} - \mathcal{A}_{32} \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{B}_{22} \\ 0 \end{bmatrix} \mathcal{H}_{22} \text{ is diagonal and nonsingular.} \tag{58}$$

Define $(\mathcal{F}, \mathcal{H})$ by

$$\begin{cases} \begin{bmatrix} \mathcal{D}_{11} & 0 \\ \mathcal{B}_{21} & \mathcal{B}_{22} \end{bmatrix} \mathcal{W}^T\mathcal{H}\mathcal{P}^{\mathcal{T}} = \begin{bmatrix} I & 0 \\ 0 & \mathcal{B}_{22}\mathcal{H}_{22} \end{bmatrix}, \\ \begin{bmatrix} \mathcal{D}_{11} & 0 \\ \mathcal{B}_{21} & \mathcal{B}_{22} \end{bmatrix} \mathcal{W}^T\mathcal{F}\mathcal{V} = \begin{bmatrix} -\mathcal{C}_{11} & -\mathcal{C}_{12} \\ -\mathcal{A}_{21} & \mathcal{B}_{22}\tilde{\mathcal{F}}_{22} \end{bmatrix}. \end{cases} \tag{59}$$

and partition $\mathcal{W}^T\mathcal{F}\mathcal{V}$ and $\mathcal{W}^T\mathcal{H}\mathcal{P}^T$ as in (41). A direct calculation yields that

$$\mathcal{D}_{11}\mathcal{H}_{11} = I, \quad \mathcal{B}_{21}\mathcal{F}_{12} + \mathcal{B}_{22}\mathcal{F}_{22} = \mathcal{B}_{22}\tilde{\mathcal{F}}_{22}, \tag{60}$$

and, furthermore,

$$\mathcal{T}_{\mathcal{F},\mathcal{H}}(s) = \mathcal{P}^T \begin{bmatrix} I & 0 \\ 0 & \mathcal{T}_{22}(s) \end{bmatrix} \mathcal{P}$$

is diagonal and nonsingular. Therefore, $\mathcal{F}$ and $\mathcal{H}$ above solve the RRDP for system (15). □

By combining Theorems 1, 8, and 10 we obtain the following result, which presents explicit and numerically verifiable necessary and sufficient solvability conditions for the RRDP of system (1).

THEOREM 11. (1) (9) (32)

(i) (1)

(ii)

$$\text{(a)} \quad n_3 = \tilde{n}_3, \quad E_{23} = 0, \quad E_{33} = 0,$$
$$\text{(b)} \quad \mathcal{D}_{21} = 0, \quad \mathcal{B}_{22} \text{ is nonsingular,}$$

(37)

(iii) (a) (b)

$$\text{(c)} \quad \mathcal{L} \text{ is nonsingular,}$$

$\mathcal{L}$ c_i i $\mathcal{C}_{22}$

$$c_i \left(\begin{bmatrix} \mathcal{E}_{22} \\ \mathcal{E}_{32} \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{A}_{22} \\ \mathcal{A}_{32} \end{bmatrix} \right)^j \left(\begin{bmatrix} \mathcal{E}_{22} \\ \mathcal{E}_{32} \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{B}_{22} \\ 0 \end{bmatrix} \right) \neq 0$$

j

$$l_i = \min\{j \geq 0 : j \text{ is integer satisfying } c_i \left(\begin{bmatrix} \mathcal{E}_{22} \\ \mathcal{E}_{32} \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{A}_{22} \\ \mathcal{A}_{32} \end{bmatrix} \right)^j \left(\begin{bmatrix} \mathcal{E}_{22} \\ \mathcal{E}_{32} \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{B}_{22} \\ 0 \end{bmatrix} \right) \neq 0\};$$

$l_i = \mu_2 - 1$

$$\mathcal{L} = \begin{bmatrix} c_1 \left(\begin{bmatrix} \mathcal{E}_{22} \\ \mathcal{E}_{32} \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{A}_{22} \\ \mathcal{A}_{32} \end{bmatrix} \right)^{l_1} \\ c_2 \left(\begin{bmatrix} \mathcal{E}_{22} \\ \mathcal{E}_{32} \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{A}_{22} \\ \mathcal{A}_{32} \end{bmatrix} \right)^{l_2} \\ \\ c_{m-\nu} \left(\begin{bmatrix} \mathcal{E}_{22} \\ \mathcal{E}_{32} \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{A}_{22} \\ \mathcal{A}_{32} \end{bmatrix} \right)^{l_{m-\nu}} \end{bmatrix} \left(\begin{bmatrix} \mathcal{E}_{22} \\ \mathcal{E}_{32} \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{B}_{22} \\ 0 \end{bmatrix} \right). \tag{61}$$

It is well known that it is ill-conditioned to compute the matrix $\mathcal{L}$ in Theorem 11(iii). Hence, Theorem 11(iii) cannot be used for the purpose of numerical computation [18]. Fortunately, the RRDP for system (37), in which

$$\begin{bmatrix} \mathcal{E}_{22} \\ \mathcal{E}_{32} \end{bmatrix}$$

is nonsingular, has been reinvestigated, and a numerically reliable algorithm has been developed using orthogonal transformations in [14, 16]. As a result, Theorem 11(ii),

the proofs of Theorems 8 and 10, and the work in [14, 16] can be used as a basis for devising a numerically reliable algorithm for solving the RRDP for the descriptor system (1) as follows.

ALGORITHM 1.

Input E, A, B C (1) E

Output (F, H) (1)

1 (9) (a) 11(ii)

„

2 (32). (b) 11(ii)

„

3 (37) $(\tilde{\mathcal{F}}_{22}, \mathcal{H}_{22})$ [16]

4 4 (59) (31) (F, H) (F, H)

In Algorithm 1, Steps 1, 2, and 3 are implemented using only orthogonal transformations, and the equations in Step 4 can be solved by existing reliable methods in MATLAB software. Therefore, Algorithm 1 is numerically reliable.

In the following we present a numerical example to illustrate Algorithm 1. In this example, all calculations were carried out using MATLAB 5.3 on a HP 712/80 workstation with IEEE standard; i.e., the machine accuracy is about $\epsilon \cong 10^{-16}$. For the sake of space limitation, we display only the matrices in systems (1), (15), and (37) and the computed (F, H). But, all other data produced by Algorithm 1 can be obtained from us on request.

1. Given a system of the form (1) with

$$E = \begin{bmatrix} -2.114533471754 & -1.370853194916 & 1.736011048459 & -0.777424723706 & 0.15552253682 & -2.56440325995 \\ 1.981948836367 & 1.026489453419 & -1.105079726788 & 1.060541636692 & 0.15747945272 & 1.83814732774 \\ -1.127692341885 & -0.738768389781 & 0.831143443676 & -0.430676489733 & 0.07468092963 & -1.60807586863 \\ -0.854341367338 & -0.769620562458 & 0.940592064356 & -0.149518153491 & 0.18927466491 & -1.19013292159 \\ 0.959445385414 & 0.382758195679 & -0.738961628069 & 0.400705245192 & -0.09207348992 & 1.47285621736 \\ 1.096197281699 & 0.333111128291 & -0.419336283785 & 0.700588861755 & 0.14595331319 & 1.23474652554 \end{bmatrix},$$

$$A = \begin{bmatrix} -2.496603142235 & -3.027855453825 & 3.468783236849 & 0.841588723375 & -0.389109896897 & -3.758198911069 \\ 2.915229861871 & 2.562626759424 & -2.641568611375 & 0.232761516265 & -0.025600000866 & 4.225601295896 \\ -2.9425385772285 & -3.2106922758844 & 3.4465197574242 & -0.1313124529388 & -0.175539883543 & -3.956516650065 \\ -2.489894380756 & -2.570286952960 & 1.313484388828 & 0.586156437213 & 0.670989681583 & -3.581315152567 \\ -0.021382994553 & -0.390533814618 & 0.523695683833 & 0.011304370558 & 0.011554521246 & -0.807953060870 \\ 1.419644861011 & 0.660544600602 & -0.674155524801 & 0.845044224805 & -0.061255771247 & 1.272523221691 \end{bmatrix},$$

$$B = \begin{bmatrix} -0.2981194205345 & 0.2863134942399 & 0.2343502478597 \\ -0.1599567951551 & -0.0253643526707 & 0.2901408282503 \\ 0.2845183665656 & 0.1741030084725 & -0.2934607414939 \\ -0.4542405355796 & -0.2318865773690 & 0.0001128595941 \\ 0.1872510988007 & -0.0171206916918 & -0.0015501494030 \\ -0.3507039683173 & 0.0264223889745 & 0.4739000261467 \end{bmatrix},$$

$$C = \begin{bmatrix} -0.166889510228 & 0.038161721675 & 0.362202814343 & 0.989416167409 & -0.191104341881 & 0.172703849180 \\ 0.010071486000 & 0.008210117004 & 0.005330380954 & 0.012683631440 & 0.010082426509 & -0.026547375104 \\ -1.212588186646 & -0.619627725178 & 0.990749246465 & -0.455865810296 & 0.152486891422 & -1.845755890228 \end{bmatrix},$$

E is singular with $\operatorname{rank}(E) = 5$.

By performing Step 1 of Algorithm 1 we get the form (9) with

$$n_1 = 5,\ n_2 = \tilde{n}_2 = 1,\ n_3 = \tilde{n}_3 = 0,\ E_{23} \text{ and } E_{33} \text{ are nonexistent,}$$

$$E_{11} = \begin{bmatrix} -1.81412863731296 & 0.42421095468505 & -1.05025682473091 & 3.17662461567576 & -1.81554610479108 \\ -2.28707883576749 & 0.58790381967586 & -0.87731476780594 & 3.11620021773695 & -2.16544496029582 \\ -0.48446826993548 & 0.42267233234963 & -0.15485265545682 & 0.41083294329550 & -0.28637255352499 \\ -0.44261627208646 & -0.22062625954174 & -0.18343402237887 & 1.30497574966312 & -1.04899876010130 \\ 0.73772319004855 & -0.40897839106130 & 0.02906644815678 & -0.44508380911416 & 0.59700805882825 \end{bmatrix},$$

$$A_{11} = \begin{bmatrix} -2.82399053295603 & -1.54190993452611 & -1.18141927967749 & 6.87716695193563 & -2.85898044575291 \\ -3.48328679385907 & -0.57668298994711 & -1.79073168993224 & 7.63482380219169 & -3.24859841311216 \\ 0.68284133476315 & 0.51353758146361 & 0.55962161514123 & -1.87270290373406 & 0.62638380994589 \\ 0.27344305777165 & -0.35875365745341 & -0.07124731006193 & 0.14804286298798 & 0.46447362716188 \\ -0.23342751856190 & -1.00663606035709 & -0.06650585078729 & 1.15561346009253 & -0.39823667919127 \end{bmatrix},$$

$$\left[B_{12} \mid A_{12}\right] = \left[\begin{array}{cc|c} -0.29338446075126 & -0.22855442644477 & -0.66970485705469 \\ 0.02262609006324 & 0.57657906855771 & -0.24241575506484 \\ -0.03965740026352 & -0.07378004229509 & 0.01290081567965 \\ -0.32758200246128 & -0.43860929198782 & -0.44527719790514 \\ -0.14453716656447 & -0.41240236312705 & -0.25082197786380 \end{array}\right],$$

$$C_1 = \begin{bmatrix} -0.49764149376588 & 0.26436703986119 & -0.37975128183368 & -0.52558936731624 & -0.33982643711893 \\ -0.00466385565791 & 0.01770707047205 & -0.01384911053026 & -0.00520420774033 & 0.02470573280247 \\ 1.07461245340293 & -0.12658398822948 & 0.34423979427567 & -1.76367786178704 & 1.44010323440087 \end{bmatrix},$$

$$C_2 = \begin{bmatrix} 0.59356291253968 \\ 0 \\ 0 \end{bmatrix}.$$

So, the condition (16) is true.

Next, we perform Step 2 of Algorithm 1 to get the form (32) with

$$\mu_1 = 1,\ \mu_2 = 4, \tau_2 = 2,\ \tau_3 = 2,\ \nu = 1,\ \mathcal{D}_{21} = 0,\ \mathcal{B}_{22} \text{ is nonsingular},$$

$$\begin{bmatrix} \mathcal{E}_{22} \\ \mathcal{E}_{32} \end{bmatrix} = \begin{bmatrix} 0.18497650851300 & 0.21650255217584 & -1.42163757723914 & 0.06052836892962 \\ 0.14215088563157 & 0.08522385113982 & -6.06459778219956 & 1.67086227506901 \\ 0.00000000000000 & 0.00753734517815 & -0.05947922313868 & 0.05399731930089 \\ -0.13125083366301 & -0.20761568732362 & -1.17186177742605 & -0.20101487098613 \end{bmatrix},$$

$$\begin{bmatrix} \mathcal{A}_{22} \\ \mathcal{A}_{32} \end{bmatrix} = \begin{bmatrix} 1.98285434904403 & 1.32236482538160 & -6.66396698837071 & 2.82024406998733 \\ 2.52659314741932 & 1.78106640040819 & -8.93198215464291 & 3.05066915354195 \\ -0.00000000000000 & 0.01488524911679 & 0.06423512937203 & -0.04810327759609 \\ -0.25920285056388 & -0.41001322791008 & 0.12457919258231 & 0.11493631853526 \end{bmatrix},$$

$$\begin{bmatrix} \mathcal{B}_{22} \\ 0 \end{bmatrix} = \begin{bmatrix} -0.17545104068710 & -0.03097433768157 \\ -0.13911242684627 & 0.26719805271403 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

$$\mathcal{C}_{22} = \begin{bmatrix} 0 & 0 & 0 & 0.03412545630891 \\ 0 & 0 & -2.35658869878193 & 0.95930344402099 \end{bmatrix}.$$

Hence, the condition (38) holds.

Then, by performing Step 3 of Algorithm 1, we get a solution $(\tilde{\mathcal{F}}_{22}, \mathcal{H}_{22})$ to the RRDP for system (37):

$$\tilde{\mathcal{F}}_{22} = \begin{bmatrix} 9.802293309432168 & 5.646580746820156 & -35.08609913450191 & 14.41371812628298 \\ -3.301832971710871 & -3.096022324474101 & 10.20273393604259 & -1.043725946131937 \end{bmatrix}$$

and

$$\mathcal{H}_{22} = \begin{bmatrix} 0.9996702178578495 & -0.7886455744849721 \\ -0.02.567986619979914 & 0.6148480770444584 \end{bmatrix}.$$

Finally, by solving the four linear equations in Step 4 of Algorithm 1, we get

$$H = \begin{bmatrix} -0.17686381243266 & -0.46126177248064 & -0.11010692440574 \\ -5.21439030830574 & -0.29532442684054 & 0.55048173448973 \\ 1.10055215419784 & 0.83667261229188 & -0.82755442430760 \end{bmatrix},$$

$$F = \begin{bmatrix} 2.2179774813366 & 3.178794801661 & -4.860781467075 & 0.199813884087 & 0.281334640565 & 8.543338100611 \\ 6.031618125212 & 7.960378436260 & -7.806500773764 & 4.155869542985 & 0.747155047430 & 9.384264975229 \\ -16.361782614150 & -16.151965525174 & 14.586848000379 & -0.397781332210 & 0.589084350377 & -24.14475367080 \end{bmatrix}.$$

Now we verify that the above pair (F, H) is a solution of the RRDP for system (1). By computing the SVD of E using MATLAB code svd.m we obtain orthogonal matrices $W, \mathcal{W} \in \mathbf{R}^{6\times 6}$ such that $(WE\mathcal{W}, W(A+BF)\mathcal{W}, W(BH), C\mathcal{W})$ is of the following form:

$$WE\mathcal{W} = \begin{matrix} \\ \\ \end{matrix}\!\!\begin{matrix} 5 & 1 \end{matrix}\begin{bmatrix} \Theta_{11} & 0 \\ 0 & 0 \end{bmatrix}\begin{matrix} \}5 \\ \}1 \end{matrix}, \quad W(A+BF)\mathcal{W} = \begin{matrix} \\ \\ \end{matrix}\!\!\begin{matrix} 5 & 1 \end{matrix}\begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{bmatrix}\begin{matrix} \}5 \\ \}1 \end{matrix},$$

$$WBH = \begin{bmatrix} \Psi_1 \\ \Psi_2 \end{bmatrix}\begin{matrix} \}5 \\ \}1 \end{matrix}, \quad C\mathcal{W} = \begin{matrix} 5 & 1 \end{matrix}\!\!\begin{bmatrix} \Upsilon_1 & \Upsilon_2 \end{bmatrix},$$

where Θ_{11} is nonsingular and $\Phi_{22} = -1.00000000000004 \neq 0$. Thus, the pencil $(E, A+BF)$ is regular and of index at most one. Furthermore, a simple calculation by using MATLAB code

$$C(sE-A-BF)^{-1}BH$$
$$= \mathrm{tf}\left(\mathrm{ss}\left(\Theta_{11}^{-1}(\Phi_{11}-\Phi_{12}\Phi_{22}^{-1}\Phi_{21}), \Theta_{11}^{-1}(\Psi_1-\Phi_{12}\Phi_{22}^{-1}\Psi_2), (\Upsilon_1-\Upsilon_2\Phi_{22}^{-1}\Phi_{21}), -\Upsilon_2\Phi_{22}^{-1}\Psi_2\right)\right)$$

yields that

$$C(sE-A-BF)^{-1}BH = \begin{bmatrix} T_{11}(s) & T_{12}(s) & T_{13}(s) \\ T_{21}(s) & T_{22}(s) & T_{23}(s) \\ T_{31}(s) & T_{32}(s) & T_{33}(s) \end{bmatrix}$$

with

$$\begin{bmatrix} T_{11}(s) \\ T_{21}(s) \\ T_{31}(s) \end{bmatrix} = \begin{bmatrix} \frac{s^5-5.466s^4+9.888s^3-5.912s^2-4.865\times 10^{-14}s-2.039\times 10^{-28}}{s^5-5.466s^4+9.888s^3-5.912s^2-3.147\times 10^{-14}s-7.126\times 10^{-28}} \\ \frac{s^3-3.950s^2+3.90s+1.732\times 10^{-14}}{s^5-5.466s^4+9.888s^3-5.912s^2-3.147\times 10^{-14}s-7.126\times 10^{-28}} \times 1.292\times 10^{-15} \\ -\frac{s^3-3.949s^2+3.900s+2.252\times 10^{-14}}{s^5-5.466s^4+9.888s^3-5.912s^2-3.147\times 10^{-14}s-7.126\times 10^{-28}} \times 3.444\times 10^{-14} \end{bmatrix},$$

$$\begin{bmatrix} T_{12}(s) \\ T_{22}(s) \\ T_{32}(s) \end{bmatrix} = \begin{bmatrix} \frac{s^5+1.811s^4-16.65s^3+25.82s^2-15.20s-7.119\times 10^{-14}}{s^5-5.466s^4+9.888s^3-5.912s^2-3.147\times 10^{-14}s-7.126\times 10^{-28}} \times 4.613\times 10^{-16} \\ \frac{0.005079s^4-0.02776s^3+0.05022s^2-0.03003s-2.034\times 10^{-15}}{s^5-5.466s^4+9.888s^3-5.912s^2-3.147\times 10^{-14}s-7.126\times 10^{-28}} \\ \frac{s^3-3.896s^2+5.276s-2.929}{s^5-5.466s^4+9.888s^3-5.912s^2-3.147\times 10^{-14}s-7.126\times 10^{-28}} \times 1.958\times 10^{-14} \end{bmatrix},$$

$$\begin{bmatrix} T_{13}(s) \\ T_{23}(s) \\ T_{33}(s) \end{bmatrix} = \begin{bmatrix} \frac{s^5-8.802s^4-6.047s^3+13.74s^2+3.245\times 10^{-15}s-2.693\times 10^{-30}}{s^5-5.466s^4+9.888s^3-5.912s^2-3.147\times 10^{-14}s-7.126\times 10^{-28}} \times 1.895\times 10^{-16} \\ \frac{s^2-1.841s-0.2652}{s^5-5.466s^4+9.888s^3-5.912s^2-3.147\times 10^{-14}s-7.126\times 10^{-28}} 3.659\times 10^{-15} \\ \frac{0.07654s^4-0.4184s^3+0.7568s^2-0.4526s+2.889\times 10^{-14}}{s^5-5.466s^4+9.888s^3-5.912s^2-3.147\times 10^{-14}s-7.126\times 10^{-28}} \end{bmatrix}.$$

Although the off-diagonal elements of $C(sE-A-BF)^{-1}BH$ are not exactly zero, the numerator coefficients of all off-diagonal terms are of an order of magnitude $\mathbf{O}(10^{-14})$ that are attributed to numerical rounding error. Thus, $C(sE-A-BF)^{-1}BH$ is for practical purposes diagonal and nonsingular. This can also be demonstrated by

a step response analysis: it has been found that the response to steps or sinusoidal inputs applied on each separate control channel of the closed-loop system effectively gives terms on the off-diagonal parts which are about $\mathbf{O}(10^{-14})$. Hence, the above pair (F, H) is a solution of the RRDP for system (1).

Example 2. We can also verify that $C(sE - A - BF)^{-1}BH$ in Example 1 is diagonal as follows.

Let $(BH)_i$ denote the ith column of BH ($i = 1, 2, 3$). We obtain orthogonal matrices W_i and $\mathcal{W}_i$ $(i = 1, 2, 3)$ by computing the controllable staircase forms [20] of the pairs $(sE-A-BF, (BH)_i)$ $(i = 1, 2, 3)$ such that $(W_i(sE-A-BF)\mathcal{W}_i, W_i(BH)_i, C\mathcal{W}_i)$ $(i = 1, 2, 3)$ are of the following forms:

$$\begin{cases} W_1(sE - A - BF)\mathcal{W}_1 = \begin{matrix} \\ \end{matrix} \begin{matrix} 4 & 2 \end{matrix} \\ \quad \begin{bmatrix} s\Theta_{11}^{(1)} - \Phi_{11}^{(1)} & s\Theta_{12}^{(1)} - \Phi_{12}^{(1)} \\ 0 & s\Theta_{22}^{(1)} - \Phi_{22}^{(1)} \end{bmatrix} \begin{matrix} \}4 \\ \}2 \end{matrix}, \\ W_1(BH)_1 = \begin{bmatrix} \Psi_1^{(1)} \\ 0 \end{bmatrix} \begin{matrix} \}4 \\ \}2 \end{matrix}, \quad C\mathcal{W}_1 = \begin{matrix} \\ \end{matrix} \begin{matrix} 4 & 2 \end{matrix} \\ \quad \begin{bmatrix} \Upsilon_{11}^{(1)} & \Upsilon_{12}^{(1)} \\ \Upsilon_{21}^{(1)} & \Upsilon_{22}^{(1)} \\ \Upsilon_{31}^{(1)} & \Upsilon_{32}^{(1)} \end{bmatrix} \begin{matrix} \}1 \\ \}1 \\ \}1 \end{matrix}, \end{cases}$$

$$\begin{cases} W_2(sE - A - BF)\mathcal{W}_2 = \begin{matrix} \\ \end{matrix} \begin{matrix} 4 & 2 \end{matrix} \\ \quad \begin{bmatrix} s\Theta_{11}^{(2)} - \Phi_{11}^{(2)} & s\Theta_{12}^{(2)} - \Phi_{12}^{(2)} \\ 0 & s\Theta_{22}^{(2)} - \Phi_{22}^{(2)} \end{bmatrix} \begin{matrix} \}4 \\ \}2 \end{matrix}, \\ W_2(BH)_2 = \begin{bmatrix} \Psi_2^{(2)} \\ 0 \end{bmatrix} \begin{matrix} \}4 \\ \}2 \end{matrix}, \quad C\mathcal{W}_2 = \begin{matrix} \\ \end{matrix} \begin{matrix} 4 & 2 \end{matrix} \\ \quad \begin{bmatrix} \Upsilon_{11}^{(2)} & \Upsilon_{12}^{(2)} \\ \Upsilon_{21}^{(2)} & \Upsilon_{22}^{(2)} \\ \Upsilon_{31}^{(2)} & \Upsilon_{32}^{(2)} \end{bmatrix} \begin{matrix} \}1 \\ \}1 \\ \}1 \end{matrix}, \end{cases}$$

and

$$\begin{cases} W_3(sE - A - BF)\mathcal{W}_3 = \begin{matrix} \\ \end{matrix} \begin{matrix} 4 & 2 \end{matrix} \\ \quad \begin{bmatrix} s\Theta_{11}^{(3)} - \Phi_{11}^{(3)} & s\Theta_{12}^{(3)} - \Phi_{12}^{(3)} \\ 0 & s\Theta_{22}^{(3)} - \Phi_{22}^{(3)} \end{bmatrix} \begin{matrix} \}4 \\ \}2 \end{matrix}, \\ W_3(BH)_3 = \begin{bmatrix} \Psi_3^{(3)} \\ 0 \end{bmatrix} \begin{matrix} \}4 \\ \}2 \end{matrix}, \quad C\mathcal{W}_3 = \begin{matrix} \\ \end{matrix} \begin{matrix} 4 & 2 \end{matrix} \\ \quad \begin{bmatrix} \Upsilon_{11}^{(3)} & \Upsilon_{12}^{(3)} \\ \Upsilon_{21}^{(3)} & \Upsilon_{22}^{(3)} \\ \Upsilon_{31}^{(3)} & \Upsilon_{32}^{(3)} \end{bmatrix} \begin{matrix} \}1 \\ \}1 \\ \}1 \end{matrix}, \end{cases}$$

where

$$\left\| \begin{bmatrix} \Upsilon_{21}^{(1)} \\ \Upsilon_{31}^{(1)} \end{bmatrix} \right\|_2 / \|C\|_2 = \mathbf{O}(10^{-15}), \quad \left\| \begin{bmatrix} \Upsilon_{11}^{(2)} \\ \Upsilon_{31}^{(2)} \end{bmatrix} \right\|_2 / \|C\|_2 = \mathbf{O}(10^{-15}),$$
$$\left\| \begin{bmatrix} \Upsilon_{11}^{(3)} \\ \Upsilon_{21}^{(3)} \end{bmatrix} \right\|_2 / \|C\|_2 = \mathbf{O}(10^{-15}),$$

and

$$\|\Upsilon_{11}^{(1)}\|/\|C\| = \mathbf{O}(1), \quad \|\Upsilon_{21}^{(2)}\|/\|C\| = \mathbf{O}(1), \quad \|\Upsilon_{31}^{(3)}\|/\|C\| = \mathbf{O}(1).$$

Therefore, $C(sE - A - BF)^{-1}BH$ is diagonal.

4. Conclusions. We have presented necessary and sufficient conditions for the solvability of the RRDP for descriptor systems. A numerical procedure, which is implementable and reliable, has been provided to verify these solvability conditions and compute the solution matrices.

Acknowledgment. The authors are grateful to the anonymous referees and Professor B. Kågström for their invaluable comments and suggestions.

REFERENCES

[1] C. A. Smith, *Automated Continuous Process Control*, Wiley, New York, 2001.
[2] Q. G. Wang, *Decoupling Control*, Springer-Verlag, Berlin, 2003.
[3] C.-T. Chen, *Linear System Theory and Design*, Holt, Rinehart and Winston, New York, 1984.
[4] J. F. Lafay, J. Descusse, and M. Malabre, *Solution to Morgan's problem*, IEEE Trans. Automat. Control, 33 (1988), pp. 732–739.
[5] J. C. Martinez Garcia and M. Malabre, *The simultaneous disturbance rejection and regular row by row decoupling: A geometric approach*, IEEE Trans. Automat. Control, 40 (1995), pp. 365–369.
[6] C. Commault, J. M. Dion, and J. Montoya, *Simultaneous decoupling and disturbance rejection: A structural approach state space system*, Internat. J. Control, 59 (1994), pp. 1325–1344.
[7] L. Dai, *Singular Control Systems*, Lecture Notes in Control and Inform. Sci., Vol. 118, Springer-Verlag, Berlin, 1989.
[8] J. W. 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.
[9] E. Fabian and W. M. Wonham, *Decoupling and disturbance rejection*, IEEE Trans. Automat. Control, 20 (1975), pp. 399–401.
[10] P. L. Falb and W. A. Wolovich, *Decoupling in the design and synthesis of multivariable control systems*, IEEE Trans. Automat. Control, 12 (1967), pp. 651–659.
[11] M. Christodoulou, *Decoupling in the design and synthesis of singular systems*, Automatica, 22 (1986), pp. 245–249.
[12] D. Chu and V. Mehrmann, *Disturbance decoupling for descriptor systems by state feedback*, SIAM J. Control Optim., 38 (2000), pp. 1830–1858.
[13] D. Chu and V. Mehrmann, *Disturbance decoupling for linear time-invariant systems: A matrix pencil approach*, IEEE Trans. Automat. Control, 46 (2001), pp. 802–808.
[14] D. Chu and R. Tan, *Numerically reliable computing for row by row decoupling problem with stability*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 1143–1170.
[15] D. Chu and R. Tan, *Solvability conditions and parameterization of all solutions for triangular decoupling problem*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 1171–1182.
[16] D. Chu and R. Tan, *Numerical Computation for Row by Row Decoupling Problem*, Technical report, National University, Singapore, 2000.
[17] A. Bunse-Gerstner, V. Mehrmann, and N. K. Nichols, *Regularization of descriptor systems by derivative and proportional state feedback*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 46–67.
[18] P. H. Petkov, N. D. Christov, and M. M. Konstantinov, *Computational Methods in Linear Control Systems*, Prentice–Hall, Hertfordshire, UK, 1991.
[19] P. N. Paraskevopulos and F. N. Koumboulis, *The decoupling of generalized state-space systems via state feedback*, IEEE Trans. Automat. Control, 37 (1992), pp. 148–152.
[20] P. Van Dooren, *The generalized eigenstructure problem in linear system theory*, IEEE Trans. Automat. Control, 6 (1981), pp. 111–129.
[21] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach*, 2nd ed., Springer-Verlag, New York, 1985.
[22] R. Byers, T. Geerts, and V. Mehrmann, *Descriptor systems without controllability at infinity*, SIAM J. Control Optim., 35 (1997), pp. 462–479.
[23] D. Chu and Y.S. Hung, *Row by Row Decoupling Problem for Descriptor Systems*, Technical report, National University, Singapore, 2000.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 3, pp. 703–723

STOCHASTIC EQUILIBRIA OF AIMD COMMUNICATION NETWORKS*

F. WIRTH†, R. STANOJEVIĆ†, R. SHORTEN†, AND D. LEITH†

Abstract. In this paper tools are developed to analyse a recently proposed random matrix model of communication networks that employ additive-increase multiplicative-decrease (AIMD) congestion control algorithms. We investigate properties of the Markov process describing the evolution of the window sizes of network users. Using paracontractivity properties of the matrices involved in the model, it is shown that the process has a unique invariant probability, and the support of this probability is characterized. Based on these results we obtain a weak law of large numbers for the average distribution of resources between the users of a network. This shows that under reasonable assumptions such networks have a well-defined stochastic equilibrium. ns2 simulation results are discussed to validate the obtained formulae. (The simulation program *ns*2, or *network simulator*, is an industry standard for the simulation of Internet dynamics.)

Key words. positive matrices, infinite products of positive matrices, AIMD congestion control, communication networks, Markov e-chain, law of large numbers

AMS subject classifications. 15A60, 68M12, 15A52

DOI. 10.1137/040620953

1. Introduction.

The dynamics of communication networks have attracted increased attention in recent years. Networks of devices that employ [illegible] (AIMD) congestion control algorithms, such as the widely deployed [illegible] (TCP), have become the focus of much of this activity. Typically, the approach adopted by the community is to model such networks by means of a fluid analogy and to employ techniques from control theory and convex optimization in their analysis; see the recent book by Srikant [27] and the references therein for an overview of this work. Recently, several authors have proposed an alternative model of TCP dynamics using products of random matrices [2, 3, 24]. The basic approach followed in these papers is to use ideas from hybrid systems theory to model the dynamics of AIMD networks as a switched, or time-varying, discrete time linear system. The approach adopted in [24] allows for techniques from the theory of nonnegative matrices and Markov chains to be employed in the analysis of these networks. The application of these techniques to the study of such networks and the mathematical analysis of the model are the principal contributions of this paper.

Networks of unsynchronized sources and drop-tail queues have been the subject of several other studies [1, 3, 5, 12, 16], and it has been documented by many authors that networks of many AIMD flows exhibit extremely complex behavior. Consequently, it is convenient to analyze such networks from a probabilistic viewpoint, as we shall do in section 4. The novelty of our approach lies in the fact that we use positive matrices to model network behavior. We shall see that this will enable us to use results from

*Received by the editors December 16, 2004; accepted for publication (in revised form) by R. Nabben March 22, 2006; published electronically September 19, 2006. This work was supported by the European Union–funded research training network *Multi-Agent Control*, HPRN-CT-1999-00107, by the Enterprise Ireland grant SC/2000/084/Y, and by the Collaborative Research Center 637 "Autonomous Logistic Processes—A Paradigm Shift and its Limitations," funded by the German Research Foundation.

http://www.siam.org/journals/simax/28-3/62095.html

†NUI Maynooth, Hamilton Institute, Maynooth, Co. Kildare, Ireland (fabian.wirth@nuim.ie, rade.stanojevic@nuim.ie, robert.shorten@nuim.ie, doug.leith@nuim.ie).

the theory of positive matrices to be employed to make predictions concerning the behavior of AIMD networks.

Fluid analogy approaches to the modeling of networks of unsynchronized sources have been the subject of wide study in the TCP community; see [6, 13, 14, 18, 19, 20, 21, 22, 15, 28, 17] and the accompanying references for further details. However, several authors have recently developed hybrid system models of networks with a single bottleneck link which employ AIMD congestion control mechanisms, most notably Hespanha [11] and Baccelli and Hong [2]. We note that the model derived in [2] is similar to the model presented here. However, whereas the model derived by Baccelli and Hong is also a random matrix model, it has an affine structure. The corresponding homogeneous (linear) part is characterized by matrices without any nonnegativity structure. In [25, 24] the same model as the one presented here is discussed. The paper [24] deals with the derivation of expected average throughputs and with the question of model validation. In [25] implications of the model for network responsiveness and network fairness are discussed, and the model validation is carried one step further in that the effects of background traffic are analyzed.

In section 2 we begin our discussion by giving an overview of AIMD congestion control and by briefly reviewing the random matrix model of AIMD network dynamic first derived in [24]. In section 3 a number of basic results are presented relating to the set of matrices used in the model. It is shown that on a jointly invariant subspace the matrices are paracontractive, which is used to show that with probability one, left products of the matrices approach the set of rank-1 column stochastic matrices. This ergodicity property plays a vital role in all the subsequent considerations. Section 4 is devoted to the analysis of the Markov chain model of the AIMD process. It is shown that the chain in question is an e-chain. Using the results of section 3 we obtain that this chain has positive and aperiodic states. From this we obtain the unique existence of an invariant probability and weak law of large number statements. Finally, the support of the invariant probability is characterized. In section 5 we collect and derive a number of results that are useful in characterizing the stochastic equilibria of various types of communication networks that employ AIMD congestion control mechanisms. In section 6 we apply these results to the study of networks employing TCP congestion control. It is shown that the model is able to predict the average behavior of TCP flows very accurately.

2. Column stochastic matrices and AIMD congestion control. A communication network consists of a number of sources and sinks connected together via links and routers. In this paper we assume that these links can be modeled as a constant propagation delay together with a queue, that the queue is operating according to a drop-tail discipline, and that all of the sources are operating an AIMD-like congestion control algorithm. In AIMD congestion control each source maintains an internal variable w_i (the window size) which tracks the number of sent unacknowledged packets that can be in transit at any time. When the window size is exhausted, the source must wait for an acknowledgment before sending a new packet. Congestion control is achieved by dynamically adapting the window size according to an additive-increase multiplicative-decrease law. Roughly speaking, the source gently probes the network for spare capacity by increasing the rate at which packets are inserted into the network, and backs off rapidly the number of packets transmitted through the network when congestion is detected through the loss of data packets. More specifically, an individual source sends packets of data through the network to a destination, and the transmission is deemed complete if an acknowledgment issued

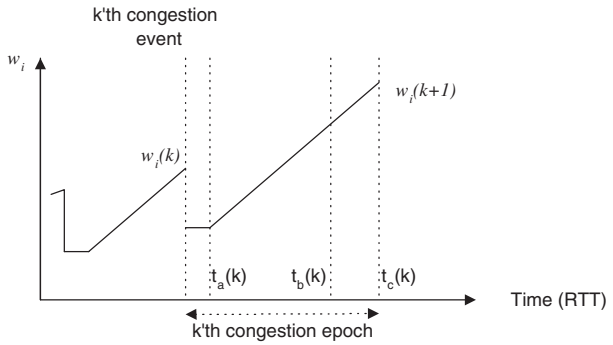


FIG. 2.1. *Evolution of window size.*

by the destination upon receipt of the packet is received by the source. As long as transmission is successful, that is, as long as all acknowledgments are received, the source increments $w_i(t)$ by a fixed amount α_i upon receipt of an acknowledgment. If an acknowledgment for a certain packet does not arrive at the sender, it is assumed that there has been a packet loss due to congestion in the network. As a consequence, the variable $w_i(t)$ is reduced in multiplicative fashion to $\beta_i w_i(t)$, where $0 < \beta_i < 1$.

2.1. A model for AIMD dynamics. In [26] a model has been presented which assumes that (i) at congestion every source experiences a packet drop; and (ii) each source has the same round-trip time (RTT).[1] In [24] this model has been extended to a random model of unsynchronized networks, where sources have different RTTs. We briefly describe the derivation of the model here. A standing assumption of the model is that all sources compete for the capacity of a single bottleneck router, and if packets are lost, this happens because the queue of that router is overflowing.

By a [illegible] we describe the situation that more packets arrive at a router than can be serviced and the queue of the router is already full. In this case, necessarily some packets are lost. Without the assumption of synchronization, at a congestion event not all sources are necessarily informed of this congestion. For the moment uniform RTT is still assumed; we will weaken this assumption later on. Let $w_i(k)$ denote the congestion window size of source i immediately before the kth network congestion event is detected by the source.

Over the kth congestion epoch as depicted in Figure 2.1 three important events can be discerned: $t_a(k)$, $t_b(k)$, and $t_c(k)$. The time $t_a(k)$ denotes the instant at which the number of unacknowledged packets in flight equals $\beta_i w_i(k)$; $t_b(k)$ is the time at which the bottleneck queue is full; and $t_c(k)$ is the time at which packet drop is detected by some of the sources, where time is measured in units of RTT.[2] It follows from the definition of the AIMD algorithm that the window evolution is completely defined over all time instants by knowledge of the $w_i(k)$ and the event times $t_a(k)$, $t_b(k)$, and $t_c(k)$ of each congestion epoch. We therefore only need to investigate the behavior of these quantities.

[1] One RTT is the time between sending a packet and receiving the corresponding acknowledgment when there are no packet drops.

[2] Note that measuring time in units of RTT results in a linear rate of increase for each of the congestion window variables between congestion events.

We assume that sources that lose a package at congestion are informed of this loss one RTT after the queue at the bottleneck link becomes full; that is, $t_c(k)-t_b(k)=1$. Also,

$$w_i(k)\geq 0 \quad \text{and} \quad \sum_{i=1}^{n} w_i(k) = P + \sum_{i=1}^{n} \alpha_i \ \forall k>0, \tag{2.1}$$

where P is the maximum number of packets which can be in transit in the network at any time; P is usually equal to $q_{max}+BT_d$, where q_{max} is the maximum queue length of the congested link, B is the service rate of the congested link in packets per second, and T_d is the RTT when the queue is empty. At the $(k+1)$th congestion event

$$w_i(k+1) = \begin{cases} \beta_i^s w_i(k) + \alpha_i[t_c(k)-t_a(k)] & \text{if source } i \text{ experiences congestion,} \\ w_i(k) + \alpha_i[t_c(k)-t_a(k)] & \text{else,} \end{cases} \tag{2.2}$$

and we set

$$\beta_i(k) \in \{\beta_i^s, 1\}, \tag{2.3}$$

corresponding to whether the source experiences a packet loss or not. Then summing the equations in (2.2) and using (2.1) we obtain

$$t_c(k)-t_a(k) = \frac{1}{\sum_{i=1}^{n}\alpha_i}\left[P - \sum_{i=1}^{n}\beta_i(k)w_i(k)\right] + 1, \tag{2.4}$$

and using (2.2)–(2.4), it follows that

$$w_i(k+1) = \beta_i(k)w_i(k) + \frac{\alpha_i}{\sum_{j=1}^{n}\alpha_j}\left[\sum_{j=1}^{n}(1-\beta_j(k))w_j(k)\right]. \tag{2.5}$$

Thus the dynamics of an entire network of such sources is given by

$$w(k+1) = A(k)w(k), \tag{2.6}$$

where $w^T(k)=[w_1(k),\dots,w_n(k)]$, and, writing $D(\beta(k)) = \operatorname{diag}(\beta_1(k),\dots,\beta_n(k))$,

$$A(k) = D(\beta(k)) + \frac{1}{\sum_{j=1}^{n}\alpha_j}\begin{bmatrix}\alpha_1\\ \alpha_2\\ \cdots\\ \alpha_n\end{bmatrix}\begin{bmatrix} 1-\beta_1(k) & \cdots & 1-\beta_n(k)\end{bmatrix}. \tag{2.7}$$

As the entries of $w(k)$ are nonnegative for all $k\geq 0$ the equations (2.6) define a positive linear system [4]. Using $b_i(s)\in(0,1], i=1,\dots,n$, we also see that all possible matrices that appear are column stochastic. In what follows we will call column stochastic matrices of the form (2.7)

So far we have worked with the assumption of uniform RTT, which is quite restrictive (although it may, for example, be valid in some long-distance networks [29]). We now extend our approach to more general network conditions. As we will see, the model that we obtain shares many structural and qualitative properties of the model described above. To distinguish variables, the nominal parameters of the

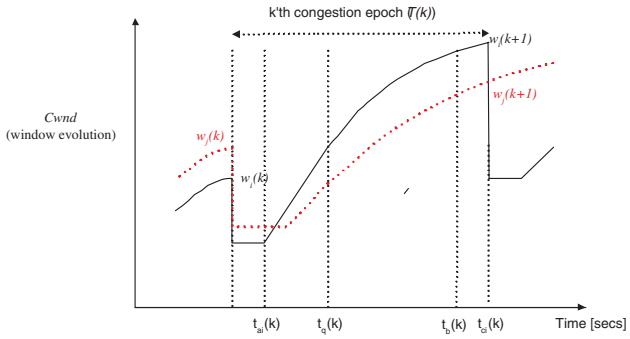


FIG. 2.2. *Evolution of window size over a congestion epoch. $T(k)$ is the length of the congestion epoch in seconds.*

sources used in the previous section are now denoted by α_i^s, β_i^s, $i = 1, \ldots, n$. Here the index s may remind the reader that these are the parameters that are chosen by each

Consider the general case of a number of sources competing for shared bandwidth in a generic dumbbell topology (where sources may have different RTTs and drops need not be synchronized). The evolution of the window size w_i of a typical source as a function of time, over the kth congestion epoch, is depicted in Figure 2.2. As before a number of important events may be discerned, where we now measure time in seconds, rather than units of RTT. Denote by $t_{ai}(k)$ the time at which the number of packets in flight belonging to source i is equal to $\beta_i^s w_i(k)$; $t_q(k)$ is the time at which the bottleneck queue begins to fill; $t_b(k)$ is the time at which the bottleneck queue is full; and $t_{ci}(k)$ is the time at which the ith source is informed of congestion. In this case the evolution of the ith congestion window variable does not evolve linearly with time after t_q seconds due to the effect of the bottleneck queue filling and the resulting variation in RTT; namely, the RTT of the ith source increases according to $RTT_i(t) = T_{d_i} + q(t)/B$ after t_q, where T_{d_i} is the RTT of source i when the bottleneck queue is empty and $0 \le q(t) \le q_{\max}$ denotes the number of packets in the queue. Note also that we do not assume that every source experiences a drop when congestion occurs. For example, a situation is depicted in Figure 2.2 where the ith source experiences congestion at the end of the epoch, whereas the jth source does not.

Given these general features it is clear that the modeling task is more involved than in the synchronized case. Nonetheless, it is possible to relate $w_i(k)$ and $w_i(k+1)$ using a similar approach to the synchronized case by accounting for the effect of nonuniform RTTs and unsynchronized packet drops as follows.

Due to the variation in RTT, the congestion window of a flow does not evolve linearly with time over a congestion epoch. Nevertheless, we may relate $w_i(k)$ and $w_i(k+1)$ linearly by defining an average rate $\alpha_i(k)$ depending on the kth congestion epoch:

$$\alpha_i(k) := \frac{w_i(k+1) - \beta_i(k) w(k)}{T(k)}, \tag{2.8}$$

where $T(k)$ is the duration of the kth epoch measured in seconds. Equivalently we have

$$w_i(k+1) = \beta_i(k) w_i(k) + \alpha_i(k) T(k)\,. \tag{2.9}$$

In the case when $q_{max} \ll BT_{d_i}$, $i = 1, \ldots, n$, the average α_i are (almost) independent of k and given by $\alpha_i(k) \approx \alpha_i^s/T_{d_i}$ for all $k \in \mathbb{N}, i = 1, \ldots, n$. The situation when

$$\alpha_i \approx \frac{\alpha_i^s}{T_{d_i}}, \quad i = 1, \ldots, n, \tag{2.10}$$

is of considerable practical importance and such networks are the principal concern of this paper. See [24] for a discussion of networks where this assumption is reasonable.

In view of (2.3) and (2.9) a convenient representation of the network dynamics is obtained as follows. At congestion the bottleneck link is operating at its capacity B, i.e.,

$$\sum_{i=1}^{n} \frac{w_i(k) - \alpha_i}{RTT_{i,max}} = B, \tag{2.11}$$

where $RTT_{i,max}$ is the RTT experienced by the ith flow when the bottleneck queue is full. Note that $RTT_{i,max}$ is independent of k. Setting $\gamma_i := (RTT_{i,max})^{-1}$ we have that

$$\sum_{i=1}^{n} \gamma_i w_i(k) = B + \sum_{i=1}^{n} \gamma_i \alpha_i \,. \tag{2.12}$$

Using steps similar to the ones performed in (2.2)–(2.4) we obtain the model

$$w_i(k+1) = \beta_i(k) w_i(k) + \frac{\alpha_i}{\sum_{j=1}^{n} \gamma_j \alpha_j} \left(\sum_{j=1}^{n} \gamma_j (1 - \beta_j(k)) w_j(k) \right), \tag{2.13}$$

and the dynamics of the entire network of sources at the kth congestion event are again described by $w(k+1) = A(k)w(k)$, where

$$A(k) = D(\beta(k)) + \frac{1}{\sum_{j=1}^{n} \gamma_j \alpha_j} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdots \\ \alpha_n \end{bmatrix} [\gamma_1(1 - \beta_1(k)), \ldots, \gamma_n(1 - \beta_n(k))], \tag{2.14}$$

and where $\beta_i(k)$ is either 1 or β_i^s. The nonnegative matrices $A_2, \ldots, A_m$ are constructed by taking the matrix A_1,

$$A_1 = \begin{bmatrix} \beta_1^s & 0 & \cdots & 0 \\ 0 & \beta_2^s & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \beta_n^s \end{bmatrix} + \frac{1}{\sum_{j=1}^{n} \gamma_j \alpha_j} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdots \\ \alpha_n \end{bmatrix} \begin{bmatrix} \gamma_1(1 - \beta_1^s), & \ldots, & \gamma_n(1 - \beta_n^s) \end{bmatrix},$$

and setting some, but not all, of the β_i to 1. This gives rise to $m = 2^n - 1$ matrices associated with the system (2.13) that correspond to the different combinations of source drops that are possible. These matrices are not AIMD matrices in the sense we have defined above. However, by a small transformation we come back to our original situation.

By considering the evolution of $w_\gamma^T(k) = [\gamma_1 w_1(k), \gamma_2 w_2(k), \ldots, \gamma_n w_n(k)]$ we obtain the following description of the network dynamics:

$$w_\gamma(k+1) = \bar{A}(k) w_\gamma(k) \tag{2.15}$$

with $\bar{A}(k) \in \bar{\mathcal{A}} = \{\bar{A}_1, \ldots, \bar{A}_m\}$, $m = 2^n - 1$, and where the $\bar{A}_i$ are obtained by the diagonal similarity transformation associated with the change of variables. As before the nonnegative matrices $\bar{A}_2, \ldots, \bar{A}_m$ are constructed by taking the matrix $\bar{A}_1$ and setting some, but not all, of the β_i^s to 1. It is easy to see that all of the matrices in the set $\bar{\mathcal{A}}$ are now AIMD matrices; for convenience we use this representation of the network dynamics to prove the main mathematical results presented in this paper. Note furthermore that the similarity transformation used to bring the matrices in AIMD form depends only on the round-trip times RTT_i and not on the α_i^s, β_i^s.

2.2. Networks of flows whose parameters vary in time. Before proceeding with our analysis we note that for some applications it is convenient to allow the parameters of the matrix $A(k)$ to vary in more general a manner than that described in the previous two sections. Our model may be extended trivially to model networks whose AIMD parameters vary with time: $\alpha_i(k)$; $\beta_i(k)$. Such situations may arise in applications where the protocol adapts its parameters to reflect prevailing network conditions or in applications where variations in network delays lead to a consequent variation in the AIMD parameters (for example, due to routing changes or in wireless networks) [22]; in fact a number of AIMD networks of this type have recently been proposed by a number of authors in the context of high-speed long-distance networks [29]. We account for such behavior in this paper by defining the set $\mathcal{M}$ to be the union of a finite number of matrix sets $\bar{\mathcal{A}}_j$, each of which is defined as above but which corresponds to fixed AIMD parameters $\{\alpha_1^j, \ldots, \alpha_n^j\}$ and $\{\beta_1^j, \ldots, \beta_n^j\}$, $1 \le j \le h$, with $\mathcal{M} = \bigcup_{j=1}^h \bar{\mathcal{A}}_j$, where h is some fixed integer.

3. Preliminaries. The principal objective of this paper is to collect and develop analytic tools to analyze models of the form derived in section 2. We will see in section 5 that it is possible to characterize the stochastic behavior of the random variable $w(k)$ under certain assumptions. The derivation of these results is somewhat technical, and to ease exposition we introduce here a number of definitions and preliminary results.

3.1. Basic notation. The following results are based on the theory of nonnegative matrices. A matrix A or a vector x is said to be nonnegative if each of its entries is a nonnegative real number and matrices or vectors are called positive if all their entries are positive. We write $A \succ B$ or $A \succeq B$ if $A - B$ is positive, respectively, nonnegative. The set of nonnegative matrices in $\mathbb{R}^{n\times n}$ is denoted by $\mathbb{R}_+^{n\times n}$. The componentwise absolute value of $A = (a_{ij}) \in \mathbb{R}^{n\times m}$ is defined by $|A| := (|a_{ij}|) \in \mathbb{R}_+^{n\times m}$.

A special subset of $\mathbb{R}_+^{n\times n}$ are the column stochastic matrices. A matrix $A \in \mathbb{R}_+^{n\times n}$ is called column stochastic if for each of its columns the sum of the corresponding elements is equal to 1. Denoting $e := [1, 1, \ldots, 1]^T$, it follows that e^T is a left eigenvector of a column stochastic matrix corresponding to the eigenvalue 1. We denote by $\mathcal{R} \subset \mathbb{R}^{n\times n}$ the set of all column stochastic matrices of rank-1 and the distance between a matrix $P \in \mathbb{R}^{n\times n}$ and the set $\mathcal{R}$ by $\operatorname{dist}(P, \mathcal{R}) = \inf\{\|P - C\| : C \in \mathcal{R}\}$, where $\|\cdot\|$ is the induced l_1-norm. Finally, the standard jth unit vector is denoted by e_j, so that $e = \sum_{j=1}^n e_j$.

3.2. Basic assumptions. Our basic objective is to model the evolution of the vector $w(k)$ for networks of AIMD flows. We consider a set of AIMD matrices $\mathcal{M} = \{M_1, \ldots, M_\mu\}, \mu \ge 1$. Associated to this set we consider the deterministic system

$$x(k+1) \in \{Mx(k) \mid M \in \mathcal{M}\} \tag{3.1}$$

and a Markov chain model

$$w(k+1) = A(k)w(k)\,, \tag{3.2}$$

where for each k the $A(k)$ is a random variable with values in $\mathcal{M}$. We recall that by (2.1) the sum $\sum_i w_i(k)$ is a constant. We may thus restrict our attention to the simplex

$$\Sigma := \left\{ x \in \mathbb{R}^n_+ \mid e^T x = \sum_{i=1}^n x_i = 1 \right\},$$

and we will study the evolution of (3.2) on Σ. We assume that the random variables $A(k), k = 0, 1, \ldots$, are independent and identically distributed (i.i.d.) and denote

$$P(A(k) = M_i) = \rho_i\,, \quad i = 1, \ldots, \mu\,.$$

As we are dealing with probabilities, necessarily, we assume $\sum_i \rho_i = 1$. With this setup the sequence $\{w(k)\}_{k\in\mathbb{N}}$ is a Markov process. The random variable of a product of length k is denoted by $\Pi(k) = A(k)A(k-1)\ldots A(0)$.

Clearly, $w(k) = \Pi(k)w(0)$, and consequently the behavior of $w(k)$, as well as the network fairness and convergence properties, are governed by the asymptotic properties of the matrix product $\Pi(k)$ as $k \to \infty$.

ASSUMPTION 3.1. $\mathcal{M} = \{M_1, \ldots, M_\mu\}$ (2.7) $A(k) = M_i$ (3.2) k $\rho_i > 0$

COMMENT 3.2. 3.1 $w(k+1) = A(k)w(k), A(k) = M_i$ k ρ_i $A(k), k \in \mathbb{N}$ $\mathcal{M}$ (3.2) (3.1) 0 $\mathcal{M}$

Given the probabilities ρ_i for $M_i \in \mathcal{M}$, one may then define the probability λ_j that source j experiences a backoff at the kth congestion event as follows:

$$\lambda_j = \sum \rho_i\,,$$

where the summation is taken over those i which correspond to a matrix in which the jth source sees a drop. To put it another way, the summation is over those indices i for which the matrix M_i is defined with a value of $\beta_j \neq 1$.

ASSUMPTION 3.3. $\mathcal{M} = \{M_1, \ldots, M_\mu\}$ (3.2) $P(A(k) = M_i) = \rho_i, i = 1, \ldots, \mu$ $\lambda_j > 0$ $j \in \{1, \ldots, n\}$

Simply stated, by Assumption 3.3 all flows must see a drop almost surely at some time (provided that they live for a long enough time).

3.3. Column stochastic matrices. Column stochastic matrices will play a central role in the discussion in section 5. We begin by collecting some results. The following two are immediate consequences of the definition of a column stochastic matrix.

LEMMA 3.4. [illegible] $A \in \mathbb{R}_+^{n\times n}$ [illegible] $e^T A = e^T$ [illegible]

It is sometimes convenient to consider the subspace orthogonal to e, which we denote by

$$S := \{z \in \mathbb{R}^n \mid e^T z = 0\}\,.$$

The subspace S is an invariant subspace for all column stochastic matrices. Given a column stochastic matrix A we denote by $\tilde{A} : S \to S$ the linear operator obtained by restricting A to S. Furthermore, we denote by $\|\cdot\|$ the 1-norm and the corresponding induced matrix norm.

LEMMA 3.5. [illegible] A [illegible] $\|A\| = 1$ [illegible] $\|\tilde{A}\| \le 1$ [illegible] A [illegible] $\|\tilde{A}\| < 1$

[illegible]. The first claim is immediate from the standard characterization of the induced 1-norm as the column-sum norm. The second claim follows as $\|\tilde{A}\| \le \|A\|$ using the definition of induced norms. Finally, if A is positive, then for a vector $z \in S, \|z\| = 1$ it holds that $-A|z| \prec |Az| \prec A|z|$ as z has positive and negative entries due to $e^T z = 0$. This implies for $z \in S, \|z\| = 1$ that

$$\|\tilde{A}z\| = \|Az\| = \|\,|Az|\,\| < \|A|z|\,\| = 1\,.$$

This shows the assertion. ∎

A feature in the proof of our main results is the observation that products of our AIMD matrices converge to a certain compact subset of the rank-1 idempotent matrices (in the sense that the distance to this set goes to zero). We use the following lemma to estimate the distance of a matrix product from the set $\mathcal{R}$ defined at the beginning of this section.

LEMMA 3.6. [illegible] $A \in \mathbb{R}_+^{n\times n}$ [illegible] $\operatorname{dist}(A, \mathcal{R}) \le 2\|\tilde{A}\|$

[illegible]. Let $A_1 = A - Aee^T/n$. Note that Aee^T/n is a rank-1 column stochastic matrix. Then $\operatorname{dist}(A, R) = \inf\{\|A - C\| :\ C \in \mathcal{R}\} \le \|A - Aee^T/n\| = \|A_1\|$. We are proving that $\|A_1\| \le 2\|\tilde{A}\|$. So let $x = z + te$, where $z \in S, t \in \mathbb{R}$ are arbitrary. Then

$$A_1 x = (A - Aee^T/n)(z + te) = Az = \tilde{A}z\,,$$

so

$$\|A_1 x\| \le \|\tilde{A}z\| \le \|\tilde{A}\|\|z\|\,.$$

To complete the proof we show that $\|z\| \le 2\|z + te\|$. Indeed, if $z_1, z_2, \ldots, z_n$ are the components of z ordered such that $z_1 \ge z_2 \ge \cdots \ge z_r \ge 0 > z_{r+1} \ge \cdots \ge z_n$, then $\|z\| = |z_1| + |z_2| + \cdots + |z_n| = 2(|z_1| + |z_2| + \cdots + |z_r|)$. On the other hand for $t \ge 0$,

$$\|z + te\| = \sum_{j=1}^{n} |z_j + t| \ge \sum_{j=1}^{r} |z_j + t| \ge \sum_{j=1}^{r} |z_j| = \frac{1}{2}\|z\|\,,$$

thus $\|z\| \le 2\|z + te\|$. For $t < 0$ a similar argument applies. ∎

Recall that the similarity transformation described to obtain (2.15) is applied simultaneously to the matrices from (2.14). Thus each matrix $M \in \mathcal{M}$ can be written in the form

$$\operatorname{diag}(\beta_1, \beta_2, \ldots, \beta_n) + \frac{1}{\sum_{j=1}^{n} \alpha_j \gamma_j}[\alpha_1\gamma_1, \ldots, \alpha_n\gamma_n]^T[(1-\beta_1), \ldots, (1-\beta_n)]\,, \tag{3.3}$$

where $\alpha_j(M)$ are positive, and all $\beta_j(M)$ are positive and not greater than 1. The parameters γ_j are also positive and independent of $M \in \mathcal{M}$, as they are determined by the RTTs of the sources; see (2.15). Thus the matrices in $\mathcal{M}$ are column stochastic. Note that if the jth column of $M \in \mathcal{M}$ is not strictly positive, then that column is equal to e_j. Using the assumptions given in section 3.2, we now aim to prove certain convergence results for the restriction of $A(k)A(k-1)\cdots A(1)$ to S. To this end we employ the notion of paracontractivity [7, 10] from the theory of nonhomogeneous matrix products. A linear operator A on $\mathbb{R}^n$ is called [illegible] with respect to the norm $\|\cdot\|$ if

$$Ax \neq x \Rightarrow \|Ax\| < \|x\|. \tag{3.4}$$

We will employ the following three results to show that almost surely products of matrices from $\mathcal{M}$ converge to the set $\mathcal{R}$. The following result is proved in [7].

THEOREM 3.7. [illegible] $\|\cdot\|$ [illegible] $\mathbb{R}^n$ [illegible] $\mathcal{F} \subset \mathbb{R}^{n\times n}$ [illegible] $\|\cdot\|$ [illegible] $\{A_k\}_{k\in\mathbb{N}} \subset \mathcal{F}^{\mathbb{N}}$ [illegible] $\{A_k A_{k-1}\dots A_1\}_{k\in\mathbb{N}}$ [illegible]

The second result shows that all matrices from $\mathcal{M}$ are paracontractive with respect to the 1-norm on S.

LEMMA 3.8. [illegible] $A \in \mathcal{M}$ [illegible] $\tilde{A}$ [illegible] S [illegible] 1 [illegible]

[illegible]. As before, let $\|\cdot\|$ denote the 1-norm. For $x \in S$ we want to show (3.4). We know that any matrix from $\mathcal{M}$ can be written in the form (3.3), where $\beta_i \in (0,1], i = 1,\dots,n$, and $\beta_j < 1$ for some $j \in \{1,\dots,n\}$. Also $\alpha_i > 0$ and $\gamma_i > 0$ for $i = 1,2,\dots,n$. Without loss of generality, assume that $\beta_1 = \beta_2 = \cdots = \beta_q = 1$ for $q < n$ and $\beta_i < 1, i = q+1,\dots,n$. In this case our matrix A is of the form

$$A = \begin{bmatrix} I_q & A_{12} \\ 0 & A_{22} \end{bmatrix},$$

where I_q is the identity matrix of order q and where $A_{12}, A_{22} \succ 0$ are such that the elements of each column of A sums to 1. Pick $x \in S$. If we partition $x = \begin{bmatrix} z_1^T & z_2^T \end{bmatrix}^T$ accordingly, we have

$$Ax = \begin{bmatrix} z_1 + A_{21}z_2 \\ A_{22}z_2 \end{bmatrix}.$$

By Lemma 3.5 it follows that $\|Ax\| \le \|x\|$. If $\|Ax\| = \|x\|$, then in each entry of Ax the summands have the same sign, because otherwise $\|Ax\| < \||A|x|\| \le \||x|\| = \|x\|$, a contradiction. For $1 \le j \le q$, this implies that for $(Ax)_j = x_j + a_{jq+1}x_{q+1} + \cdots + a_{jn}x_n$ the signs of the summands coincide. Similarly for $q+1 \le j \le n$ the signs of the summands of $(Ax)_j = a_{jq+1}x_{q+1} + \cdots + a_{jn}x_n$ coincide. This implies $x_i x_j \ge 0$ for all $i = 1,2,\dots,n$ and all $j = q+1, q+2,\dots,n$. If we fix $j \ge q+1$, we have

$$0 = x_j e^T x = x_j(x_1 + \cdots + x_n) \ge x_j^2. \tag{3.5}$$

We conclude that $z_2 = 0$, which also means that $Ax = x$. Thus for $x \in S$, we have $\|Ax\| \le \|x\|$ with equality if and only if $Ax = x$, as desired. ∎

Our third result is purely technical and is stated as a separate lemma to aid exposition of Theorem 3.10.

COROLLARY 3.9. [illegible] $A \in \mathcal{M}$ [illegible] $i_1, i_2, \dots, i_q$ [illegible] $x \in S$ [illegible] $Ax = x$ [illegible] x [illegible] $e_{i_1}, e_{i_2}, \dots, e_{i_q}$

. This follows from the previous proof, as we have seen that $Ax = x$ implies that $x_j = 0$ for $j = q+1, \dots, n$. In other words, $x \in \operatorname{span}\{e_1, \dots, e_q\}$. The general statement follows by permutation. □

Given the three previous results it is now possible to show that almost all products of matrices from $\mathcal{M}$ approach the set $\mathcal{R}$.

THEOREM 3.10. $\{A_k\}_{k\in\mathbb{N}}$ $\mathcal{M}$ $i \in \{1, 2, \dots, n\}$ $T_i \in \mathcal{M}$ i $\{A_k\}_{k\in\mathbb{N}}$

$$\lim_{k\to\infty} \{\tilde{A}_k \tilde{A}_{k-1} \cdots \tilde{A}_1\} = 0\,.$$

3.3 $\{A(k)\}_{k\in\mathbb{N}}$ $\lim_{k\to\infty} \tilde{A}(k)\tilde{A}(k-1)\cdots\tilde{A}(0) = 0$

. By Lemma 3.8, the matrices $\tilde{A}_k$, $k \in \mathbb{N}$, are paracontractive with respect to $\|\cdot\|$. Using Theorem 3.7 it follows that $\{\tilde{A}_k \tilde{A}_{k-1} \cdots \tilde{A}_1\}_{k\in\mathbb{N}}$ is convergent. To prove that the limit is 0 let $s \in S$. Then there exist $y \in S$ such that $y = \lim_{k\to\infty} A_k A_{k-1} \cdots A_1 s$. We will prove that $y = 0$ from which the first assertion follows. For fixed i let $\{A_{n_k}\}_{k\in\mathbb{N}}$ be a subsequence of $\{A_k\}_{k\in\mathbb{N}}$ with $A_{n_k} = T_i$. Then

$$y = \lim_{k\to\infty} A_{n_k} A_{n_k-1} \cdots A_1 s = T_i \lim_{k\to\infty} A_{n_k-1} \cdots A_1 s = T_i y.$$

Thus $T_i y = y \in S$ since $s \in S$. By Corollary 3.9 the ith coordinate of y is zero. Since i is arbitrary, it follows that $y = 0$.

By Assumption 3.3 for each $j \in \{1, \dots, n\}$ the probability that matrices with positive jth column occur infinitely often in a realization of the process is equal to 1. Thus $\lim_{k\to\infty} \tilde{A}(k)\tilde{A}(k-1)\cdots\tilde{A}(0) = 0$ with probability 1. □

The next result shows that the expected distance between $A(k)A(k-1)\cdots A(1)$ and $\mathcal{R}$ decreases exponentially; a fact of independent interest.

PROPOSITION 3.11. $\{A(k)\}_{k\in\mathbb{N}}$ 3.1 3.3 $d(k) := E(\operatorname{dist}(A(k)A(k-1)\cdots A(1), \mathcal{R}))$ $\eta < 1$ $C \geq 1$ k

$$d(k) \leq C\eta^k. \tag{3.6}$$

. Let $\theta = 1 - \min_{j=1,n} \lambda_j < 1$ and let l be an integer such that $1 > n\theta^l$.

At first, note that the jth column of the product of several matrices from $\mathcal{M}$ is positive if and only if one of these matrices has positive jth column, otherwise it is equal to e_j. Consider the products of length l: $\Pi(l) = A(l)A(l-1)\cdot A(1)$. The probability that the jth column of $\Pi(l)$ is not strictly positive is $o_j := (1-\lambda_j)^l \leq \theta^l$. For the probability q_l that at least one column of $\Pi(l)$ is not strictly positive, we have that $q_l \leq o_1 + o_2 + \cdots + o_n \leq n\theta^l$. Thus the probability p_l that $\Pi(l)$ is positive satisfies $p_l = 1 - q_l \geq 1 - n\theta^l > 0$. Let $k = dl + r$, where $0 \leq r < l$. We can split the product $\Pi(k) = A(k)A(k-1)\cdots A(1)$ into the product of the first r terms $D_0 = A(k)A(k-1)\cdots A(k-r+1)$ and the product of d blocks of length l: $D_i = A(il)A(il-1)\cdots A(l(i-1)+1)$ for $i = 1, 2, \dots, d$. So $\Pi(k) = D_0 D_d \cdots D_1$. Note that for all $i = 0, 1, \dots, d$, D_i, as a product of column stochastic matrices, is column stochastic, and therefore $\|D_i\| = 1$ and $\|\tilde{D}_i\| \leq 1$. With this notation we have

$$\operatorname{dist}(\Pi(k), \mathcal{R}) \leq 2\|\tilde{\Pi}(k)\| = 2\|\tilde{D}_0 \tilde{D}_d \cdots \tilde{D}_1\| \leq 2\|\tilde{D}_d \cdots \tilde{D}_1\|\,.$$

Define

$$\delta := \max\{\|\tilde{T}\| :\ T = A_l A_{l-1} \dots A_1 > 0, A_1, A_2, \dots, A_l \in \mathcal{M}\} < 1\,. \tag{3.7}$$

Since the set in (3.7) is finite, the maximum exists and is strictly less than 1 by Lemma 3.5. For any $j \in \{0, 1, 2, \dots, d\}$ the probability that exactly j of the matrices $D_1, D_2, \dots, D_d$ are positive is equal to $z_j = \binom{d}{j} p_l^j (1-p_l)^{d-j}$. We also know that if j of matrices $D_1, D_2, \dots, D_d$ are positive, then $\|(D_d D_{d-1} \cdots D_1)\tilde{}\,\| = \|\tilde{D}_d \cdots \tilde{D}_1\| \leq \|\tilde{D}_d\| \cdots \|\tilde{D}_1\| \leq \delta^j$. Thus we obtain

$$\begin{aligned} d(k) &\leq 2E(\|\tilde{D}_d\| \cdots \|\tilde{D}_1\|) \leq 2 \sum_{j=0}^{d} z_j \delta^j \\ &= 2 \sum_{j=0}^{d} \binom{d}{j} (p_l \delta)^j (1-p_l)^{d-j} = 2(1 + p_l\delta - p_l)^d \leq C\eta^k \,, \end{aligned}$$

where for the last inequality we choose

$$\eta := (1 - p_l + p_l\delta)^{1/l} < 1 \quad \text{and} \quad C := 2/\eta^l \,. \tag{3.8}$$

This shows the assertion. □

4. Invariant measures. In this section we study the existence of invariant measures of the Markov process $\{w(k)\}_{k\in\mathbb{N}}$. Throughout we assume that Assumptions 3.1 and 3.3 are satisfied. Our considerations are based on the results presented in [23], to which we refer the reader for further background material. We briefly present some basic properties for the Markov chain $\{w(k)\}_{k\in\mathbb{N}}$ on the simplex Σ. By $\mathcal{B}(\Sigma)$ we denote the Borel σ-algebra of Σ.

Associated with our Markov chain there is a transition kernel $P(x, X)$ for $x \in \Sigma, X \in \mathcal{B}(\Sigma)$, which gives the probability to reach the set X from the point x. This transition kernel acts on continuous functions $h : \Sigma \to \mathbb{R}$ through

$$Ph(x) = \int_\Sigma h(y) P(x, dy) = \sum_{i=1}^{\mu} \rho_i h(M_i x) \,. \tag{4.1}$$

It is obvious that Ph is continuous for continuous h, so that P is [illegible]. Furthermore we have $\|A_i\| \leq 1, i = 1, \dots, \mu$, so that $\|A_i(x-y)\| \leq \|x-y\|$. Using the uniform continuity of h it follows that for any continuous function $h : \Sigma \to \mathbb{R}$, the sequence

$$P^k h \,, \quad k \in \mathbb{N} \,,$$

defined inductively through repeated application of (4.1), is equicontinuous. Markov chains whose transition kernel have this property are called [illegible]; see [23].

An important notion in the study of Markov chains are invariant probabilities. Recall that a probability measure π is called [illegible] for a Markov process if

$$\pi(X) = \int_\Sigma P(x, X) d\pi(x) \quad \forall X \in \mathcal{B}(\Sigma),$$

that is, intuitively, the distribution of mass on Σ given by the probability measure π is not changed if it is rearranged according to the evolution of the Markov process.

As we are considering an e-chain, we obtain from [23, Theorem 12.0.1] that an invariant probability exists in our case. We aim to show its uniqueness. To this end

we first study the possible support of invariant measures. We introduce the set of sequences

$$\mathcal{L} := \{\{A_k\}_{k\in\mathbb{N}} \in \mathcal{M}^{\mathbb{N}} \mid \{A_k\}_{k\in\mathbb{N}} \text{ satisfies the conditions of Theorem 3.10}\}.$$

By Theorem 3.10 we know that the left products of a sequence $\{A_k\}_{k\in\mathbb{N}} \in \mathcal{L}$ approach the set of rank-1 column stochastic matrices. We define the set of limit points of such sequences by

$$\mathcal{R}_{\mathcal{L}} := \{R \in \mathcal{R} \mid \exists \{A_k\}_{k\in\mathbb{N}} \in \mathcal{L}, k_l \to \infty : \lim_{l\to\infty} \Pi(k_l) = R\}.$$

As the matrices $R \in \mathcal{R}$ are column stochastic and of rank 1 they can be represented in the form $R = ze^T$, where $z \succeq 0$ and $\|z\| = 1$. Thus the set $\mathcal{R}_{\mathcal{L}}$ naturally defines a subset of the simplex Σ by

$$\mathcal{C} := \{z \in \Sigma \mid ze^T \in \mathcal{R}_{\mathcal{L}}\}. \tag{4.2}$$

We note the following properties of $\mathcal{C}$.

PROPOSITION 4.1. [illegible] $\mathcal{M}$ [illegible] (3.1) [illegible] (3.2) [illegible] $\mathcal{C}$ [illegible] (4.2) [illegible]

(i) $\mathcal{C}$ [illegible] (3.1).

(ii) [illegible] $\{x(k)\}_{k\in\mathbb{N}}$ $x(0) \in \Sigma$ [illegible] (3.1) [illegible]

$$\operatorname{dist}(x(k), \mathcal{C})$$

(iii) [illegible] $z \in \mathcal{C}$ [illegible] $U \subset \Sigma$ [illegible] z [illegible] $k_0 > 0$ [illegible] $P^k(x, U) > \delta > 0$ [illegible] $k \geq k_0$ [illegible] $x \in \Sigma$.

(iv) [illegible] $w_0 \in \Sigma$ [illegible]

$$\lim_{k\to\infty} \operatorname{dist}(w(k), \mathcal{C}) = 0.$$

[illegible]. (i) Let $x \in \mathcal{C}, B \in \mathcal{M}$. By definition there exists a sequence $\{A_k\}_{k\in\mathbb{N}} \in \mathcal{L}$ and $k_l \to \infty$ such that

$$\Pi(k_l) = A_{k_l} A_{k_l-1} \dots A_1 \to ze^T.$$

We write $\Pi(k_l) = ze^T + \Delta_k$, where $\|\Delta_k\| \to 0$. Now we define a new sequence by repeating our initial sequence and inserting B, i.e., we consider the sequence

$$\{A_1, A_2, \dots, A_{k_1}, B, A_1, A_2, \dots, A_{k_2}, B, A_1, \dots, A_{k_3}, B, A_1, \dots\}.$$

Denoting products of length l of this sequence by $\Psi(l)$ we have

$$\begin{aligned}\Psi\left(l + \sum_{j=1}^{l} k_j\right) = B\Pi(k_l)\Psi\left((l-1) + \sum_{j=1}^{l-1} k_j\right) &= B(ze^T + \Delta_k)\Psi\left((l-1) + \sum_{j=1}^{l-1} k_j\right) \\ &= Bze^T + B\Delta_k\Psi\left((l-1) + \sum_{j=1}^{l-1} k_j\right),\end{aligned}$$

where we have used that all matrices are column stochastic in the last step. As $\|\Delta_k\| \to 0$, this implies that $\Psi(l + \sum_{j=1}^{l} k_j) \to Bze^T$ as $l \to \infty$. The constructed sequence clearly lies in $\mathcal{L}$ so that $Bz \in \mathcal{C}$, which is what we wanted to show.

(ii) Let $x \in \Sigma$. Pick a $z \in \operatorname{cl}\mathcal{C}$ such that $\operatorname{dist}(x, \mathcal{C}) = \|x - z\|$. Then for $A \in \mathcal{M}$ it follows using (i) that

$$\operatorname{dist}(Ax, \mathcal{C}) \leq \|Ax - Az\| \leq \|x - z\| = \operatorname{dist}(x, \mathcal{C}).$$

This shows the assertion.

(iii) Fix $z \in \mathcal{C}$ and let $U \subset \Sigma$ be an open neighborhood of z. Then we may choose $\epsilon > 0$ such that $x \in \Sigma, \|x - z\| < \epsilon$ implies $x \in U$. By definition of $\mathcal{C}$ there exists a k_0 and a product $\Pi(k_0)$ such that $\|\Pi(k_0) - ze^T\| < \epsilon$. This implies for any $x \in \Sigma$ that

$$\|\Pi(k_0)x - z\| = \|(\Pi(k_0) - ze^T)x\| < \epsilon,$$

so that $\Pi(k_0)x \in U$ and, consequently, $P^{k_0}(x, U) > \delta > 0$ for all $x \in \Sigma$. As this probability is independent of x we see in particular that $P^k(z, U) > \delta > 0$ for all $k \geq k_0$ by considering the transition from $k - k_0$ to k.

(iv) This is an immediate consequence of Theorem 3.10. □

In the terminology of Markov chains, we have proved in Proposition 4.1(iii) that each $z \in \mathcal{C}$ is [illegible] and [illegible] for the Markov chain $\{w(k)\}_{k\in\mathbb{N}}$. For a general definition of positive and aperiodic states of an e-chain, see [23, pp. 456, 459]. Using the existence of positive and aperiodic states we obtain the following fundamental statement from [23, Theorem 18.0.2] and [8].

THEOREM 4.2. [illegible] $\mathcal{M}$ [illegible] (3.2) [illegible]

(i) [illegible] π.

(ii) [illegible] $x \in \Sigma$ [illegible] $h : \Sigma \to \mathbb{R}$ [illegible] $w(0) = x$ [illegible]

$$\lim_{k\to\infty} \frac{1}{k} \sum_{j=0}^{k-1} h(w(j)) = \int_\Sigma h(y) d\pi(y) \quad \text{[illegible]}$$

(iii) [illegible] $x \in \Sigma$ [illegible] $h : \Sigma \to \mathbb{R}$ [illegible]

$$\int_\Sigma h(y) P^k(x, dy) \to \int_\Sigma h(y) d\pi(y) \quad \text{[illegible]}\ k \to \infty.$$

The previous result can be sharpened by using the special structure of the set of AIMD matrices $\mathcal{M}$.

THEOREM 4.3. [illegible] $\mathcal{M}$ [illegible] (3.2) [illegible] π [illegible]

$$\operatorname{supp}\pi = \operatorname{cl}\mathcal{C}.$$

[illegible]. We first show that $\mathcal{C} \subset \operatorname{supp}\pi$. Assume to the contrary that $x \in \mathcal{C} \setminus \operatorname{supp}\pi$. Then there exists an open neighborhood V of x with $V \cap \operatorname{supp}\pi = \emptyset$. By Proposition 4.1(iii) it follows for all $y \in \operatorname{supp}\pi$ that $P^k(y, V) > 0$ for some k large enough, which contradicts $x \notin \operatorname{supp}\pi$.

To show $\operatorname{supp}\pi \subset \operatorname{cl}\mathcal{C}$, let $\epsilon > 0$ and consider the set

$$U_\epsilon := \{x \in \Sigma \mid \operatorname{dist}(x, \mathcal{C}) > \epsilon\}.$$

As the distance of $w(k)$ to $\mathcal{C}$ is nonincreasing for every sample path by Proposition 4.1(ii), this shows that $P(x, U_\epsilon) > 0$ implies $x \in U_\epsilon$. Thus

$$\pi(U_\epsilon) = \int_\Sigma P(x, U_\epsilon) d\pi(x) = \int_{U_\epsilon} P(x, U_\epsilon) d\pi(x)\,.$$

If $\pi(U_\epsilon) > 0$, this shows that with probability 1 any evolution starting in U_ϵ stays in U_ϵ. This is a contradiction to $\operatorname{dist}(w(k), \mathcal{C}) \to 0$ with probability 1, which we know by Proposition 4.1(iv). This shows $\pi(U_\epsilon) = 0$, and as $\epsilon > 0$ was arbitrary, we obtain the assertion. □

The interesting point of the previous result is that the support of the invariant probability π is determined by the set of matrices $\mathcal{M}$, and only the distribution of mass on that set changes under variation of the probabilities ρ_i. In the next section we show that in some cases the expected values of the average can be elegantly expressed in terms of the data, without the knowledge of the invariant probability π.

5. Long term averages. From a practical point of view, we now present the main results of the paper. For the system defined in section 3.2 we know that the stochastic process $\{w(k)\}$ satisfies the strong law of large numbers. An important consequence of this result is that the vector of window sizes $w(k)$, averaged over time, converges in probability to a well-defined stochastic equilibrium. It is of interest to know what this equilibrium is given the data of the system.

Recall that $\Pi(k)$ is the random variable defined by $\Pi(k) = A(k-1)A(k-2)\dots A(0)$. It is prudent at this point to note that it follows from the discussion that the expectation of the random variable $A(k)$ is independent of k and is equal to

$$E(A(k)) = E(A(1)) = \sum_{i=1}^{\mu} \rho_i M_i. \tag{5.1}$$

Given Assumption 3.3, this immediately implies that matrix $E(A(1))$ is a positive column stochastic matrix and consequently has a unique Perron[3] eigenvector x_p given by $E(A(1))x_p = x_p$, $x_p^T y = 1$. Using the independence of the random variables $A(k)$, this shows the following statement.

PROPOSITION 5.1. [illegible] $\mathcal{M}$ [illegible] $\{A(k)\}_{k\in\mathbb{N}}$ [illegible] 3.1 [illegible] 3.3 [illegible] $\Pi(k)$ [illegible]

$$E(\Pi(k)) = \left(\sum_{i=1}^{\mu} \rho_i M_i\right)^k, \quad \text{[illegible]} \quad \lim_{k\to\infty} E(\Pi(k)) = x_p e^T, \tag{5.2}$$

[illegible] $x_p \succ 0$ [illegible]

$$\left(\sum_{i=1}^{\mu} \rho_i M_i\right) x_p = x_p\,, \quad e^T x_p = 1\,. \tag{5.3}$$

We are now interested in the long-term average of the window size. To this end we define the random variable $\overline{w}(k)$ by

$$\overline{w}(k) := \frac{1}{k+1}\sum_{i=0}^{k} w(i) = \left(\frac{1}{k+1}\sum_{i=0}^{k} \Pi(i)\right) w(0) = \overline{\Pi(k)} w(0)\,.$$

[3]Recall that for any column stochastic matrix $V \succ 0$ with Perron eigenvector x_p, it holds that $\lim_{k\to\infty} V^k = x_p e^T$ [4].

COROLLARY 5.2. [illegible] $\mathcal{M}$ [illegible] $\{A(k)\}_{k\in\mathbb{N}}$ [illegible] 3.1 [illegible] 3.3 [illegible] $\overline{w}(k)$ [illegible]

$$E(\overline{w}(k)) = \frac{1}{k+1}(I + E(A(1)) + E(A(1))^2 + \cdots + E(A(1))^k)w(0),$$

[illegible] x_p [illegible] (5.3) [illegible]

$$\lim_{k\to\infty} E(\overline{w}(k)) = x_p e^T w(0)\,.$$

[illegible]. This follows since $E(A(1))^k \to x_p e^T$ as $k \to \infty$. □

The following theorem shows how the average distribution of network capacities can be characterized.

THEOREM 5.3. [illegible] $\mathcal{M}$ [illegible] (3.2) [illegible] 3.1 [illegible] 3.3 [illegible]

$$\lim_{k\to\infty} \overline{w}(k) = x_p e^T w(0), \tag{5.4}$$

[illegible] x_p [illegible] (5.3)

[illegible]. This is a consequence of Theorem 4.2 and Corollary 5.2. To be precise, by Theorem 4.2(ii) we have that if $w(0) \in \Sigma$, then

$$\overline{w}(k) \to \int_\Sigma w d\pi(w) =: E_\pi(w)$$

almost surely. (To obtain the desired result for vectors from the scalar results presented in Theorem 4.2, it suffices to consider the projections onto each coordinate.) If $w(0) \succeq 0$ is not in Σ, this equation scales by $e^T w(0)$ by linearity. Thus in particular $E(\overline{w}(k)) \to E_\pi(w) e^T w(0)$. As by Corollary 5.2 we have $E(\overline{w}(k)) \to x_p e^T w(0)$, which implies (5.4). □

To summarize, the previous result says that the average distribution of the resources of the network is given by the vector x_p, which can be simply obtained by finding the dominant eigenvalue of $\sum \rho_i M_i \succ 0$.

5.1. Stochastic equilibria of AIMD networks. Proposition 5.1 and Theorem 5.3 provide remarkable insights into the behavior of communication networks employing AIMD congestion control. In principle, they relate the asymptotic properties of such networks to the Perron eigenvector of $E(A(1))$. Since $E(A(1))$ is easily computable, it is possible not only to predict but also to control the asymptotic properties of such networks through judiciously manipulating the AIMD parameters and/or the probabilities ρ_i. In this context it is natural to ask whether the Perron eigenvector of $E(A(1))$ can be directly related to the AIMD parameters of the network. We now discuss some examples, where the calculation of $E(A(1))$ is particularly simple.

(i) [illegible]. By this we mean that the network parameters cannot change in time and that there is a unique set of AIMD parameters $((\alpha_1, \ldots, \alpha_n), (\beta_1, \ldots, \beta_n))$ that is used in the construction of all matrices $M \in \mathcal{M}$. In this case it is readily shown that

$$E(A(1)) = \operatorname{diag}(\delta_1, \ldots, \delta_n) + \frac{1}{\sum_{i=1}^n \alpha_i \gamma_i}[\alpha_1\gamma_1, \ldots, \alpha_n\gamma_n]^T[1-\delta_1, \ldots, 1-\delta_n], \tag{5.5}$$

where $\delta_i = 1 - \lambda_i(1-\beta_i)$. Further, it follows directly by inspection that the Perron eigenvector of E(A(1)) is given by

$$x_p = \left[\frac{\alpha_1\gamma_1}{\lambda_1(1-\beta_1)} \ , \ \dots \ , \ \frac{\alpha_n\gamma_n}{\lambda_n(1-\beta_n)} \right]^T .$$

Consequently, the network convergence properties and the rates of convergence of $E(w(k))$ can be controlled directly by manipulating the network parameters $(\alpha_i, \beta_i, \rho_i)$. Clearly, such networks are of great interest since most practical wireline networks (including those employing TCP) fall into this category. A more detailed discussion of such network types can be found in [24].

(ii) [illegible]. Here we assume that there is a finite set of AIMD parameters $((\alpha_1^l, \dots, \alpha_n^l), (\beta_1^l, \dots, \beta_n^l))$, $l = 1, \dots, m$, and all matrices in $M \in \mathcal{M}$ are constructed as an AIMD matrix corresponding to one of these parameters. In this case it is convenient to consider two cases: (a) networks where the $\alpha_i = \alpha_i^l$ are independent of l and the β_i^l vary; and (b) networks where both α_i^l and β_i^l vary.

In the first case it is again readily shown that

$$E(A(1)) = \operatorname{diag}(\delta_1, \dots, \delta_n) + \frac{1}{\sum_{i=1}^n \alpha_i\gamma_i}[\alpha_1\gamma_1, \dots, \alpha_n\gamma_n]^T[1-\delta_1, \dots, 1-\delta_n], \tag{5.6}$$

where $\delta_i = E(\beta_i) < 1$. As before x_p can be found by inspection and is given by

$$x_p = \left[\frac{\alpha_1\gamma_1}{1-\delta_1} \ , \ \dots \ , \ \frac{\alpha_n\gamma_n}{1-\delta_n} \right]^T . \tag{5.7}$$

In the more general case it appears to be difficult to derive explicit formulae for x_p. One simplification occurs when the following situation prevails. The matrix $E(A(1))$ can be written as

$$E(A(1)) = \sum_{j=1}^{h} \sum_{M_i \in \bar{\mathcal{A}}_j} \rho_i M_i = \sum_{j=1}^{h} Z_j. \tag{5.8}$$

In the case when the Z_j are positive matrices with a common Perron eigenvector x_p, it follows that x_p is also the Perron eigenvector of $E(A(1))$ and the stochastic equilibria of the corresponding communication network is defined by x_p. Hence, it follows that time-varying networks constructed by switching between networks with a common equilibrium results in a constituent network with the same equilibrium state (although the rate of convergence to this equilibrium is difficult to bound).

6. Experimental results. The mathematical results derived in section 5 are surprisingly simple when one considers the potential mathematical complexity of the unsynchronized network model (2.6). The simplicity of these results is a direct consequence of Assumptions 3.1 and 3.3. The objective of this section is therefore twofold: (i) to validate the unsynchronized model (2.6) in a general context; and (ii) to validate the analytical predictions of the model and thereby confirm that the aforementioned assumptions are appropriate in practical situations.

6.1. Networks of two unsynchronized flows: Ensemble averages. We first consider the behavior of two TCP flows in the dumbbell topology shown in

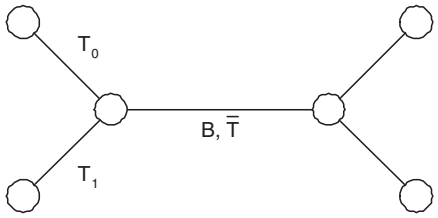


FIG. 6.1. *Dumbbell topology.*

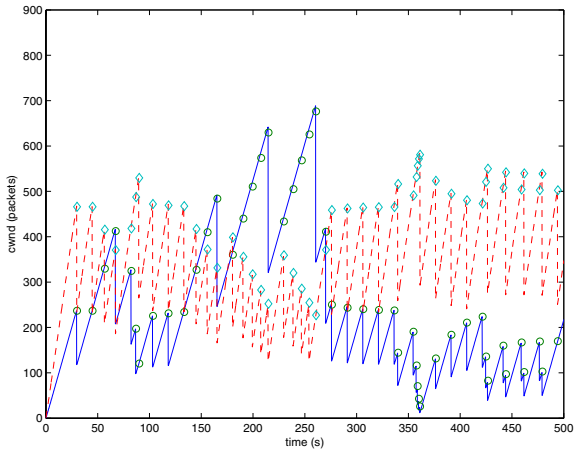


FIG. 6.2. *Evolution of window size: Predictions of the network model compared with packet-level ns2 simulation results. Key:* $\circ$ *= flow 1 (model),* $\diamond$ *= flow 2 (model), dashed line = flow 1 (ns2), solid line = flow 2 (ns2). Network parameters:* $B = 100Mb$*,* $q_{max} = 80$ *packets,* $\bar{T} = 20ms$*,* $T_0 = 102ms$*;* $T_1 = 42ms$*; no background web traffic.*

Figure 6.1. Our analytic results are based upon two fundamental assumptions: (i) that the dynamics of the evolution of the source congestion windows can be accurately modeled by (2.6); and (ii) that the allocation of packet drops among the sources at congestion can be described by random variables. We consider each of these assumptions in turn.

(i) [illegible]. A comparison of the predictions made by the model (2.6) against the output of a packet-level ns2 simulation is depicted in Figure 6.2. Here, the pattern of packet drops observed in the simulation is used to select the appropriate matrix $A(k)$ from the set $\mathcal{M}$ at each congestion event when evaluating (2.6). As can be seen, the model output is very accurate. In Figure 6.3 we also plot the evolution of the linear combination $\sum_{i=1}^{n} \gamma_i w_i$, where the γ_i are defined in (2.12). It can be seen that $\sum_{i=1}^{n} \gamma_i w_i$ has the same value at each congestion event thereby validating the constraint (2.12) used in the model.

(ii) [illegible]. It is well known that networks of TCP flows with drop-tail queues can exhibit a rich variety of deterministic drop-behaviors [9]. However, most real networks carry at least a small amount of web traffic. It is shown in [25] that already a small amount of background web traffic is enough to disrupt the coherent structure associated with phase effects and other complex phenomena previously observed in simulations of unsynchronized networks [9]. This is confirmed by

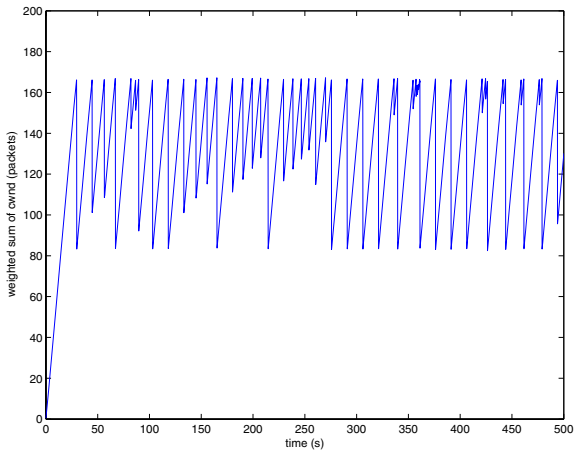


FIG. 6.3. *Evolution of* $\sum_{i=1}^{n} \gamma_i w_i$. *Network parameters:* $B = 100Mb$, $q_{max} = 80$ *packets,* $\bar{T} = 20ms$, $T_0 = 102ms$; $T_1 = 42ms$; *no background web traffic.*

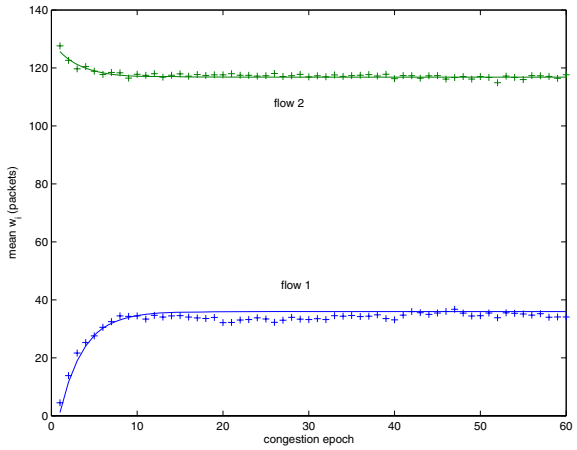


FIG. 6.4. *Variation of ensemble mean* $w_i(k)$ *with congestion epoch in dumbbell topology of Figure* 6.1. *Key:* + = ns2 *simulation result (average over* 200 *runs); solid line = Proposition* 5.1. *Network parameters:* $B = 50Mb$, $q_{max} = 50$ *packets,* $\bar{T} = 20ms$, $T_0 = 102ms$, $T_1 = 2ms$; *approximately* 0.5% *bidirectional background web traffic.*

statistical tests of this measured data, which confirm the validity of Assumptions 3.1 and 3.3.

By performing repeated packet-level simulations with different random seed values for the web traffic generator, the ensemble average congestion window can be estimated. We can also determine from the simulation results the proportion of congestion events corresponding to both flows simultaneously seeing a packet drop, flow 1 seeing a drop only, and flow 2 seeing a drop only. Using these estimates of the probabilities ρ_i, the ensemble average congestion window can also be estimated from Proposition 5.1. An example of the resulting estimates are shown in Figure 6.4. Here, we run simulations for 250 seconds with one flow started at 0 seconds and a second TCP flow started after 50 seconds (giving the first flow the opportunity to reach its

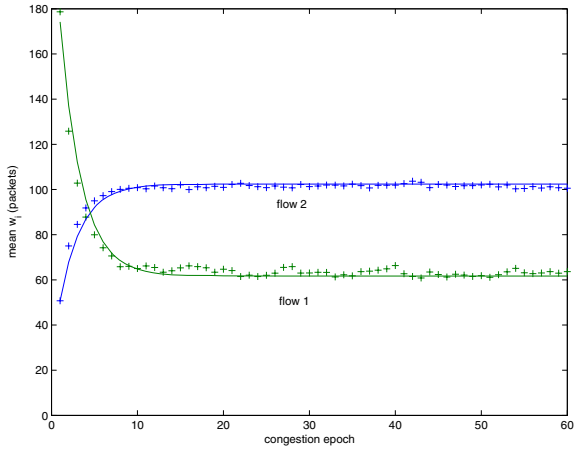


FIG. 6.5. *Variation of ensemble mean $w_i(k)$ with congestion epoch in dumbbell topology of Figure* 6.1. *Key:* $+ =$ ns2 *simulation result (average over* 200 *runs); solid line Proposition* 5.1. *Network parameters:* $B = 50Mb$, $q_{max} = 50$ *packets,* $\bar{T} = 20ms$, $T_0 = 2ms$, $T_1 = 42ms$; *approximately* 0.5% *bidirectional background web traffic.*

steady state). A small amount of bidirectional background web traffic is also included and slow-start is switched off to allow us to focus on the congestion avoidance behavior. The average congestion window evolution, estimated from 200 runs of the simulation, is plotted in Figure 6.4 together with the predictions of Proposition 5.1. It can be seen that the agreement is remarkably good. Not only is the long-term average accurately captured, but so is the manner in which the flows converge to this long-term average. That is, the model accurately describes the dynamic evolution over time, on average, of the TCP flows and thereby is useful for the analysis of both short and long-lived flows. The results shown in Figure 6.4 are for a single choice of network conditions, but the model remains accurate for other conditions; see, for example, Figure 6.5. As can be seen from the figures, the predictions of Proposition 5.1 and the ns2 simulations are consistently in close agreement.

Acknowledgment. The authors wish to thank John Foy for useful discussions.

REFERENCES

[1] E. ALTMAN, T. JIMÉNEZ, AND R. NÚÑEZ-QUEIJA, *Analysis of two competing TCP/IP connections*, Perform. Evaluation, 49 (2002), pp. 43–55.

[2] F. BACCELLI AND D. HONG, *AIMD, fairness and fractal scaling of TCP traffic*, in Proceedings of the IEEE INFOCOM 2002, New York, 2002, pp. 229–238.

[3] F. BACCELLI AND D. HONG, *Interaction of TCP flows as billiards*, in Proceedings of the IEEE INFOCOM 2003, San Francisco, 2003, pp. 841–853.

[4] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, SIAM, Philadelphia, 1994.

[5] P. BROWN, *Resource sharing of TCP connections with different round trip times*, in Proceedings of the IEEE INFOCOM 2000, Tel Aviv, Israel, 2000, pp. 1734–1741.

[6] Y. CHAIT, C. V. HOLLOT, V. MISRA, H. HAN, AND Y. HALEVI, *Dynamic analysis of congested TCP networks*, in Proceedings of American Control Conference, San Diego, 1999, pp. 2430–2435.

[7] L. ELSNER, I. KOLTRACHT, AND M. NEUMANN, *On the convergence of asynchronous paracontractions with applications to tomographic reconstruction from incomplete data*, Linear Algebra Appl., 130 (1990), pp. 65–82.

[8] J. H. Elton, *An ergodic theorem for iterated maps*, Ergodic Theory Dynam. Systems, 7 (1987), pp. 481–488.
[9] S. Floyd and V. Jacobson, *Traffic phase effects in packet-switched gateways*, J. Internetworking: Practice and Experience, 3 (1992), pp. 115–156.
[10] D. Hartfiel, *Nonhomogeneous Matrix Products*, World Scientific, Singapore, 2002.
[11] J. Hespanha, *Stochastic hybrid systems: Application to communication networks*, in Hybrid Systems, Computation, and Control, R. Alur and G. J. Pappas, eds., Proceedings of HSCC 2004, Lecture Notes in Comput. Sci. 2993, Springer, Heidelberg, 2004, pp. 387–401.
[12] J. Hespanha, S. Bohacek, K. Obraczka, and J. Lee, *Hybrid modeling of TCP congestion control*, in Hybrid Systems, Computation, and Control, M. D. Di Benedetto and A. L. Sangiovanni-Vicentelli, eds., Proceedings of the HSCC 2001, Lecture Notes in Comput. Sci. 2034, Springer, Heidelberg, 2001, pp. 291–304.
[13] C. V. Hollot and Y. Chait, *Nonlinear stability analysis of a class of TCP/AQM networks*, in Proceedings of the 40th IEEE Conference on Decision and Control, Orlando, FL, 2001, pp. 2309–2314.
[14] C. V. Hollot, V. Misra, D. Towsley, and W. Gong, *A control theoretic analysis of RED*, in Proceedings of the IEEE INFOCOM, Anchorage, AK, 2001, pp. 1510–1519.
[15] C. V. Hollot, V. Misra, D. Towsley, and W. Gong, *Analysis and design of controllers for AQM routers supporting TCP flows*, IEEE Trans. Automat. Control, 47 (2002), pp. 945–959.
[16] D. Hong and D. Lebedev, *Many TCP User Asymptotic Analysis of the AIMD Model*, Tech. report INRIA 4229, INRIA Rocquencourt, 2001.
[17] R. Johari and D. Tan, *End-to-end congestion control for the internet: Delays and stability*, IEEE/ACM Trans. Networking, 9 (2001), pp. 818–832.
[18] F. P. Kelly, *Mathematical modelling of the internet*, in Proceedings of the ICIAM 99 (Edinburgh), 4th International Congress of Industrial Applied Mathematics, Oxford University Press, Oxford, UK, 2000, pp. 105–116.
[19] S. S. Kunniyur and R. Srikant, *Stable, scalable, fair congestion control, and AQM schemes that achieve high utilization in the internet*, IEEE Trans. Automat. Control, 48 (2003), pp. 2024–2029.
[20] S. Low, F. Paganini, and J. Doyle, *Internet congestion control*, IEEE Control Systems Mag., 32 (2002), pp. 28–43.
[21] S. Mascolo, *Congestion control in high speed communication networks using the Smith principle*, Automatica, 35 (1999), pp. 1921–1935.
[22] L. Massoulie, *Stability of distributed congestion control with heterogeneous feedback delays*, IEEE Trans. Automat. Control, 47 (2002), pp. 895–902.
[23] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Comm. Control Engrg. Ser., Springer, London, 1993.
[24] R. Shorten, F. Wirth, and D. Leith, *A positive systems model of TCP-like congestion control: Asymptotic results*, IEEE/ACM Trans. Networking, 14 (2006), pp. 616–629.
[25] R. Shorten, C. King, F. Wirth, and D. Leith, *Modelling TCP congestion control dynamics in droptail environments*, Automatica, to appear.
[26] R. N. Shorten, D. J. Leith, J. Foy, and R. Kilduff, *Analysis and design of AIMD congestion control algorithms in communication networks*, Automatica, 41 (2005), pp. 725–730.
[27] R. Srikant, *Internet Congestion Control*, Control Theory 14, Birkhäuser Boston, Boston, MA, 2004.
[28] G. Vinnicombe, *On the stability of networks operating TCP-like congestion control*, in Proceedings of the 15th IFAC World Congress on Automatic Control, Barcelona, Spain, 2002.
[29] L. Xu, K. Harfoush, and I. Rhee, *Binary increase congestion control (BIC) for fast long-distance networks*, in Proceedings of IEEE INFOCOM 2004, Hong Kong, 2004, pp. 2514–2524.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 3, pp. 724–748

FAST SOLUTION OF TOEPLITZ- AND CAUCHY-LIKE LEAST-SQUARES PROBLEMS*

G. RODRIGUEZ†

Abstract. The least-squares solution of overdetermined linear systems with Toeplitz- or Cauchy-like structure is studied with an "augmented matrix" approach. A fast algorithm for the computation of the pseudoinverse in the full-rank case is developed, based on the displacement properties of the matrices involved, and the parameters on which the algorithm depends are determined optimally. Finally, the performance of the method is tested through numerical experimentation.

Key words. least squares, displacement structure, Toeplitz, Cauchy, generalized Schur algorithm, augmented matrix

AMS subject classifications. 65F05, 65F20, 15A09

DOI. 10.1137/050629148

1. Displacement and reconstructibility. The idea of displacement structure was introduced in [12, 26] in connection to Toeplitz matrices and later extended in many subsequent papers and reviews (see, e.g., [6, 23, 28, 29, 34]). In this setting a structured matrix is characterized by a low , meaning that its image under a certain has low rank. The displacement structure of generalized inverse matrices and of pseudoinverses was investigated in [17, 18].

There are basically two types of displacement operators, defined on the linear space of complex $m \times n$ matrices. The first one is the so-called Stein displacement operator (or),

$$\nabla_{\{U,V\}}(A) = A - UAV,$$

where U and V are square matrices of dimension m and n, respectively. In this paper we adopt the Sylvester (or) displacement operator, introduced for the first time in [23]. We say that a matrix A of size $m \times n$ satisfies a (Sylvester) displacement equation for two given $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ if

$$\Delta_{\{U,V\}}(A) := UA - AV = GH^*,$$

where $G \in \mathbb{C}^{m\times\delta}$, $H \in \mathbb{C}^{n\times\delta}$, and H^* denotes the conjugate transpose of H. The matrices G and H, in general not unique, are called the of A, and

$$\operatorname{rank}_\Delta(A) := \operatorname{rank}(\Delta_{\{U,V\}}(A)) \le \delta$$

is its . If G and H are full-rank matrices, then $\delta = \operatorname{rank}_\Delta(A)$. In the cases of interest δ depends only on the structure of the matrix A and not on its dimension.

*Received by the editors April 13, 2005; accepted for publication (in revised form) by D. A. Bini April 4, 2006; published electronically September 19, 2006. This work was supported by MIUR under the COFIN grant 2004015437 and by INdAM-GNCS.
http://www.siam.org/journals/simax/28-3/62914.html
†Dipartimento di Matematica e Informatica, Università di Cagliari, viale Merello 92, 09123 Cagliari, Italy (rodriguez@unica.it).

The main advantage in considering the displacement structure of a matrix is that, unlike other classical structures, it is inherited by its inverse, by its Schur complements of any order, and by products of structured matrices. This property led to the development of a family of fast algorithms (order $O(n^2)$ for an $n \times n$ matrix) for the LU factorization of a structured matrix, based on the recursive computation of Schur complements by accessing and modifying only the displacement matrices and generators of the given matrix.

The [illegible] displacement structure, characterized by diagonal displacement matrices, is of particular interest. In fact, as this structure is [illegible], i.e., it is preserved through rows and/or column permutations, it allows the employment of pivoting procedures to improve the steadiness of LU factorization [13, 20, 22]. Moreover, the displacement matrices typical of some classical structures (Toeplitz, Hankel, Vandermonde, etc.) can be diagonalized by fast transforms, so that it is possible to convert structured matrices belonging to these classes into Cauchy-like matrices by suitably modifying their generators [13, 20, 21].

When the displacement operator Δ is injective, the information contained in the displacement matrices and generators of a given matrix is sufficient to reconstruct the matrix itself. On the contrary, when the kernel of Δ is nontrivial it is necessary to store extra data, besides the generators, to recover all the elements of a matrix. In this case, the matrix is called [illegible].

In [27], using results from [14], partially reconstructible matrices are treated by splitting the space into a direct sum of the kernel of the displacement operator, whose dimension is generally small, and its orthogonal complement. In this paper, since the adopted displacement matrices (and so the kernel of the operator) are going to change during the computation, we use a different, more pragmatic approach to deal with partially reconstructible matrices (see also [22]).

The aim of this paper is to apply some of the results quoted above to an idea originally introduced in [25], where the authors showed that many [illegible] matrices, like T_1^{-1}, T_1T_2, $T_1 - T_2T_3^{-1}T_4$, even though they do not share the Toeplitz structure with their factors T_i, can be expressed as Schur complements of certain augmented matrices and stored by means of their displacement matrices and generators. In [25], in particular, it is suggested to express the solution

$$\mathbf{x} = (A^T W^{-1} A)^{-1} A^T W^{-1} \mathbf{b}$$

of the generalized least-squares problem [1]

$$\min_{\mathbf{x}} \|B^{-1}(A\mathbf{x} - \mathbf{b})\|_2,$$

where A is a full-rank matrix and $W = BB^T$ is positive definite, as the Schur complement of the submatrix

$$\begin{bmatrix} -W & A \\ A^T & 0 \end{bmatrix}$$

into the larger matrix

$$\begin{bmatrix} -W & A & -\mathbf{b} \\ A^T & 0 & 0 \\ 0 & I & 0 \end{bmatrix}.$$

In the following, continuing research started in [35], we develop this idea by converting a full-rank Toeplitz least-squares problem into a Cauchy-like one, and by employing the generalized Schur algorithm for its solution. We show that it is possible to apply this approach only to a particular class of Cauchy-like matrices and that, by assigning suitable values to some parameters, the matrix resulting from the above conversion falls into this class. The stability of the algorithm is enhanced by a pivoting strategy and by choosing optimally the constants on which the method depends. The performance of the method is finally illustrated by the results of numerical experiments.

There are many references about fast and superfast algorithms for Toeplitz least-squares problems; see, e.g., [2, 7, 8, 10, 11, 36]. In [7], a superfast method is proposed for the solution of a Toeplitz least-squares linear system, which operates on the displacement representation of a particular augmented matrix by a divide-and-conquer version of the generalized Schur algorithm.

Among the most recent papers, we mention [5], [16], and [40]. In [5], applying some results from [6], a fast algorithm for the solution of a square Toeplitz-like linear system is developed, which is proved to be backward stable. Since the algorithm is based on a modified fast QR factorization, it probably could be adapted for least-squares problems. In [16], using an augmented matrix approach, but a procedure different from the one proposed here, the linear system is previously transformed into a Cauchy-like one and then solved by a variation of the generalized Schur algorithm; moreover, an approximation of the total pivoting strategy is proposed, which gives good stability properties without enlarging the complexity of the algorithm. Finally, in [40] the authors describe a superfast method ($O((m+n)\log^2(m+n))$ floating point operations for an $m \times n$ matrix), based on the extension of the Toeplitz matrix to a circulant one and on the successive conversion of the least-squares linear system into an interpolation problem; the algorithm is stabilized by a particular technique.

Obviously, direct methods are not the only possible approach for solving least-squares problems, and iterative methods (see, e.g., [4]) often outperform them. Anyway, as noted above, there is still much interest in the study of direct algorithms, and we think that the method developed here can lead to interesting followups, as outlined in section 8, in Tikhonov regularization, in particular for what concerns multiparameter regularization and the estimation of the optimal regularization parameter.

2. Cauchy-like least squares. A [illegible] (or [illegible]) matrix C of size $m \times n$ is a matrix which satisfies, for given complex vectors $\mathbf{t} \in \mathbb{C}^m$ and $\mathbf{s} \in \mathbb{C}^n$, the displacement equation

$$\Delta_{\{D_{\mathbf{t}}, D_{\mathbf{s}}\}}(C) = D_{\mathbf{t}} C - C D_{\mathbf{s}} = G_C H_C^*, \tag{2.1}$$

where the displacement matrices are diagonal,

$$D_{\mathbf{t}} = \operatorname{diag}(t_1, \ldots, t_m), \quad D_{\mathbf{s}} = \operatorname{diag}(s_1, \ldots, s_n),$$

the generators are

$$G_C^* = \begin{bmatrix} \boldsymbol{\phi}_1 & \cdots & \boldsymbol{\phi}_m \end{bmatrix}, \quad H_C^* = \begin{bmatrix} \boldsymbol{\psi}_1 & \cdots & \boldsymbol{\psi}_n \end{bmatrix},$$

$\boldsymbol{\phi}_i, \boldsymbol{\psi}_j \in \mathbb{C}^\delta$ are column vectors and, usually, $\delta \ll n$.

When the condition $t_i \neq s_j$ is verified for any (i, j), the elements of the matrix C can be explicitly written in the form

$$C_{ij} = \frac{\boldsymbol{\phi}_i^* \cdot \boldsymbol{\psi}_j}{t_i - s_j}, \qquad i = 1, \ldots, m, \ j = 1, \ldots, n.$$

This formula loses its significance if $t_k = s_\ell$ for (k, ℓ) in some subset $\mathcal{I}_0 \subset \{1, \dots, m\} \times \{1, \dots, n\}$. When this happens, the corresponding Cauchy-like matrix is [illegible], in the sense of the definition given in section 1. This means that the elements C_{ij} with $(i, j) \notin \mathcal{I}_0$ are uniquely determined by the displacement matrices and generators, while it is necessary to store extra information to recover the elements indexed in $\mathcal{I}_0$. In this case, the kernel of the displacement operator consists of all the matrices whose entries are zero whenever their indexes are not in $\mathcal{I}_0$, so that the dimension of the null space agrees with the cardinality of $\mathcal{I}_0$.

If for each k there is at most one index ℓ such that $t_k = s_\ell$, like, for example, when $\mathbf{t}$ is a permutation of $\mathbf{s}$ and $s_i \neq s_j$ for $i \neq j$, a possible workaround consists of storing the vector

$$\mathbf{u} = C\mathbf{e}, \quad \mathbf{e} = (1, \dots, 1)^T,$$

and reconstructing C by the formula

$$C_{ij} = \begin{cases} \frac{\boldsymbol{\phi}_i^* \cdot \boldsymbol{\psi}_j}{t_i - s_j}, & t_i \neq s_j, \\ u_i - \sum_{k \neq j} C_{ik}, & t_i = s_j. \end{cases}$$

Let us consider the linear system

$$C\mathbf{x} = \mathbf{b}, \tag{2.2}$$

where C is an $m \times n$ complex Cauchy-like matrix with $m \geq n$, $\operatorname{rank}(C) = n$, $\mathbf{x} \in \mathbb{C}^n$, and $\mathbf{b} \in \mathbb{C}^m$. Computing its least-squares solution [1] means solving the optimization problem

$$\min_{\mathbf{x} \in \mathbb{C}^n} \|C\mathbf{x} - \mathbf{b}\|_2, \tag{2.3}$$

whose minimizer is the solution of the system of normal equations

$$C^*C\mathbf{x} = C^*\mathbf{b}. \tag{2.4}$$

When a matrix A of dimension $m \times n$ is partitioned into blocks as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

with $A_{11} \in \mathbb{C}^{r \times r}$ nonsingular, we define its [illegible] as

$$\mathcal{S}_r(A) := A_{22} - A_{21}A_{11}^{-1}A_{12}.$$

The computation of $\mathcal{S}_r(A)$ is equivalent to the application of the first r steps of Gauss reduction. Anyway, when A has displacement structure, the generalized Schur algorithm [28, 29] can perform this task in $O(\alpha m r)$ operations, with α independent on m and r, operating only on the displacement matrices and generators of A.

Now, consider the augmented matrix

$$M_C = \begin{bmatrix} I_m & C & 0 \\ C^* & 0 & C^* \\ 0 & I_n & 0 \end{bmatrix} \tag{2.5}$$

of dimension $(m+2n)\times(2m+n)$, where I_m is the identity matrix of dimension m. A matrix which can be partitioned into structured blocks, like (2.5), is sometimes referred to as a [illegible] [19].

The [illegible] (or [illegible]) of C [1], that is, the solution operator of problem (2.3), is given by the Schur $(m+n)$-complement of M_C; in fact

$$\mathcal{S}_{m+n}(M_C) = -\begin{bmatrix}0 & I_n\end{bmatrix}\begin{bmatrix}I_m & C\\ C^* & 0\end{bmatrix}^{-1}\begin{bmatrix}0\\ C^*\end{bmatrix} = (C^*C)^{-1}C^*.$$

We will compute $\mathcal{S}_{m+n}(M_C)$ by a fast implementation of the generalized Schur algorithm and use it to evaluate the solution

$$\mathbf{x} = (C^*C)^{-1}C^*\mathbf{b}$$

to the normal equations (2.4). To do so, we must first investigate the displacement structure of the mosaic matrix (2.5).

From (2.1) we obtain the following equation for C^*:

$$D_{\mathbf{s}}^*C^* - C^*D_{\mathbf{t}}^* = -H_CG_C^*; \tag{2.6}$$

in order to satisfy the displacement equations of both C and C^* we define the displacement structure of M_C with respect to

$$\mathcal{D}_L = D_{\mathbf{t}} \oplus D_{\mathbf{s}}^* \oplus D_{\mathbf{s}},$$
$$\mathcal{D}_R = D_{\mathbf{t}}^* \oplus D_{\mathbf{s}} \oplus D_{\mathbf{t}}^*.$$

Since the displacement matrices $\mathcal{D}_L$ and $\mathcal{D}_R$ are diagonal, M_C is itself Cauchy-like.

By writing explicitly $\Delta_{\{\mathcal{D}_L,\mathcal{D}_R\}}(M_C)$ we observe that, while the identity matrix in the block with coordinates $(3,2)$ satisfies the equation

$$D_{\mathbf{s}}I_n - I_nD_{\mathbf{s}} = 0,$$

in accordance with its trivial Cauchy structure ($\operatorname{rank}_{\Delta_{\{D_{\mathbf{s}},D_{\mathbf{s}}\}}}(I_n) = 0$ for any diagonal matrix $D_{\mathbf{s}}$), the $(1,1)$ block reads

$$D_{\mathbf{t}}I_m - I_mD_{\mathbf{t}}^*,$$

leading to the unacceptable consequence of considering I_m as a Cauchy-like matrix with displacement rank m when $\mathbf{t}$ is a vector with complex entries. This shows that the above approach is favorable only when $D_{\mathbf{t}} = D_{\mathbf{t}}^*$, in particular when C is a real Cauchy-like matrix. When this happens, the displacement rank of M_C is 2δ if $\delta = \operatorname{rank}_\Delta(C)$, and a pair of generators is given by

$$G_M = \begin{bmatrix}G_C & 0\\ 0 & -H_C\\ 0 & 0\end{bmatrix}, \quad H_M = \begin{bmatrix}0 & G_C\\ H_C & 0\\ 0 & G_C\end{bmatrix}.$$

However, a nice displacement structure for M_C can be recovered also when $D_{\mathbf{t}} \neq D_{\mathbf{t}}^*$, but $|t_i| = |s_j| = 1$, $i = 1,\dots,m$, $j = 1,\dots,n$. In fact, under this assumption $D_{\mathbf{t}}$ and $D_{\mathbf{s}}$ are unitary matrices, and multiplying (2.6) times $D_{\mathbf{s}}$ on the left and $D_{\mathbf{t}}$ on the right leads to

$$D_{\mathbf{s}}C^* - C^*D_{\mathbf{t}} = G_{C^*}H_{C^*}^*, \tag{2.7}$$

with $G_{C^*} = D_{\mathbf{s}} H_C$ and $H_{C^*} = D_{\mathbf{t}}^* G_C$. In this case M_C is Cauchy-like with rank 2δ, displacement matrices

$$\begin{aligned} \mathcal{D}_L &= D_{\mathbf{t}} \oplus D_{\mathbf{s}} \oplus D_{\mathbf{s}}, \\ \mathcal{D}_R &= D_{\mathbf{t}} \oplus D_{\mathbf{s}} \oplus D_{\mathbf{t}}, \end{aligned} \tag{2.8}$$

and generators

$$G_M = \begin{bmatrix} G_C & 0 \\ 0 & G_{C^*} \\ 0 & 0 \end{bmatrix}, \quad H_M = \begin{bmatrix} 0 & H_{C^*} \\ H_C & 0 \\ 0 & H_{C^*} \end{bmatrix}.$$

Note that this procedure can be applied also when $|t_i| = |s_j| = \alpha \neq 0$ for all i, j by simply rescaling the initial linear system (2.2).

It is important to note that if we consider the displacement operator $\Delta_{\{\mathcal{D}_L, \mathcal{D}_R\}}$ with displacement matrices (2.8), then the blocks of M_C with coordinates $(1,1)$, $(1,3)$, $(2,2)$, and $(3,2)$, marked here with a gray background,

$$\begin{bmatrix} I_m & C & 0 \\ C^* & 0 & C^* \\ 0 & I_n & 0 \end{bmatrix},$$

are partially reconstructible, since their displacement matrices are equal. Under the assumption that $t_i \neq t_j$ and $s_i \neq s_j$ for $i \neq j$, the kernel of the corresponding displacement operator consists of all diagonal matrices.

3. Toeplitz-like least squares. A Toeplitz matrix $T \in \mathbb{C}^{m \times n}$ is characterized by the property

$$T_{ij} = t_{i-j}, \quad i = 1, \ldots, m, \ j = 1, \ldots, n.$$

A matrix $A \in \mathbb{C}^{m \times n}$ is called [illegible] if, for fixed $\xi, \eta \in \mathbb{C} \setminus \{0\}$, it satisfies the displacement equation

$$\Delta_{\xi,\eta}(A) := Z_{\xi,m} A - A Z_{\eta,n} = G_A H_A^*, \tag{3.1}$$

where $G \in \mathbb{C}^{m \times \delta}$, $H \in \mathbb{C}^{n \times \delta}$, δ is a fixed integer ($\delta \ll n$), and each of the displacement matrices is a ϕ-[illegible], defined by

$$Z_{\phi,k} = \begin{pmatrix} 0 & 0 & \cdots & 0 & \phi \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \in \mathbb{C}^{k \times k}.$$

Toeplitz-like matrices include Toeplitz matrices as a subclass. In fact any Toeplitz matrix T satisfies (3.1) with displacement rank $\delta = 2$, regardless of its dimension. A pair of generators for T is given by

$$G_A = \begin{pmatrix} -\eta t_0 & 1 \\ t_{-n+1} - \eta t_1 & 0 \\ t_{-n+2} - \eta t_2 & 0 \\ \vdots & \vdots \\ t_{-n+m-1} - \eta t_{m-1} & 0 \end{pmatrix}, \quad H_A = \begin{pmatrix} 0 & \overline{\xi t_{m-1} - t_{-1}} \\ 0 & \overline{\xi t_{m-2} - t_{-2}} \\ \vdots & \vdots \\ 0 & \overline{\xi t_{m-n+1} - t_{-n+1}} \\ 1 & \overline{\xi t_{m-n}} \end{pmatrix}.$$

When $m = n$, the operator $\Delta_{\xi,\eta}(A)$ is noninvertible if and only if $\xi = \eta$. In this case, its kernel consists of all the ξ-circulant matrices [14, Theorem 1.1]. As will be shown later, for computational purposes it is convenient to take $|\xi| = |\eta| = 1$.

In analogy with section 2, we consider the overdetermined linear system

$$A\mathbf{x} = \mathbf{b}, \tag{3.2}$$

where A is an $m \times n$ complex Toeplitz-like matrix, $m \geq n$, $\text{rank}(A) = n$, $\mathbf{x} \in \mathbb{C}^n$, and $\mathbf{b} \in \mathbb{C}^m$, and solve it in the least-squares sense. The [illegible] of A,

$$A^\dagger = (A^*A)^{-1}A^*,$$

can be computed as the Schur $(m+n)$-complement of the mosaic matrix

$$M_A = \begin{bmatrix} I_m & A & 0 \\ A^* & 0 & A^* \\ 0 & I_n & 0 \end{bmatrix} \tag{3.3}$$

of dimension $(m+2n) \times (2m+n)$.

Fixing an arbitrary pair of complex numbers (ξ, η), the Toeplitz structure of A^* leads to the displacement equation

$$\Delta_{\eta,\xi}(A^*) = Z_{\eta,n}A^* - A^*Z_{\xi,m} = G_{A^*}H_{A^*}^*, \tag{3.4}$$

where G_{A^*} and H_{A^*}, in principle, could be computed analogously to G_A and H_A. Moreover, for the identity matrix we have

$$\Delta_{\xi,\xi}(I_m) = Z_{\xi,m}I_m - I_mZ_{\xi,m} = 0 \tag{3.5}$$

for any $\xi \in \mathbb{C}$. It is then immediate to observe that M_A satisfies the displacement equation

$$\mathcal{Z}_L M_A - M_A \mathcal{Z}_R = G_{M_A} H_{M_A}^*$$

with

$$\mathcal{Z}_L = Z_{\xi,m} \oplus Z_{\eta,n} \oplus Z_{\eta,n},$$
$$\mathcal{Z}_R = Z_{\xi,m} \oplus Z_{\eta,n} \oplus Z_{\xi,m},$$

and

$$G_{M_A} = \begin{bmatrix} G_A & 0 \\ 0 & G_{A^*} \\ 0 & 0 \end{bmatrix}, \quad H_{M_A} = \begin{bmatrix} 0 & H_{A^*} \\ H_A & 0 \\ 0 & H_{A^*} \end{bmatrix}.$$

When A is a Toeplitz matrix, we have $\text{rank}_\Delta(M_A) \leq 4$. However, since the displacement matrices $\mathcal{Z}_L$ and $\mathcal{Z}_R$ are not shift matrices, M_A is not Toeplitz-like itself.

For any $\xi \in \mathbb{C}$, we take its mth complex roots $(t_j)^m = \xi$, $j = 1, \ldots, m$, ordered by increasing phase, and introduce the matrices of dimension m:

$$D_{\xi,m} = \text{diag}(t_1, \ldots, t_m),$$
$$\mathcal{F}_{\xi,m} = \frac{1}{\sqrt{m}} \begin{bmatrix} 1 & t_1 & \cdots & t_1^{m-1} \\ 1 & t_2 & \cdots & t_2^{m-1} \\ \vdots & \vdots & & \vdots \\ 1 & t_m & \cdots & t_m^{m-1} \end{bmatrix}.$$

The matrix $\mathcal{F}_{\xi,m}$ is connected to the normalized Fourier matrix $\mathcal{F}_m = \mathcal{F}_{1,m}$ by means of the relation

$$\mathcal{F}_{\xi,m} = \mathcal{F}_m \cdot \mathrm{diag}(1, t_1, t_1^2, \ldots, t_1^{m-1}).$$

If we choose ξ so that $|\xi| = 1$, the matrix $\mathcal{F}_{\xi,m}$ is unitary and we can apply the factorizations

$$Z_{\xi,m} = \mathcal{F}_{\xi,m}^* D_{\xi,m} \mathcal{F}_{\xi,m}, \quad Z_{\eta,n} = \mathcal{F}_{\eta,n}^* D_{\eta,n} \mathcal{F}_{\eta,n} \tag{3.6}$$

in (3.1), (3.4), and (3.5) in order to convert the Toeplitz-like blocks A, A^*, I_m, and I_n of M_A to Cauchy-like matrices [13, 21]. The corresponding displacement equations are summarized in Table 1.

TABLE 1
Cauchy-like matrices and their displacement features.

	Displacement	Generators	
$C = \mathcal{F}_{\xi,m} A \mathcal{F}_{\eta,n}^*$	$\{D_{\xi,m}, D_{\eta,n}\}$	$G_C = \mathcal{F}_{\xi,m} G_A$	$H_C = \mathcal{F}_{\eta,n} H_A$
$C^* = \mathcal{F}_{\eta,n} A^* \mathcal{F}_{\xi,m}^*$	$\{D_{\eta,n}, D_{\xi,m}\}$	$G_{C^*} = \mathcal{F}_{\eta,n} G_{A^*}$	$H_{C^*} = \mathcal{F}_{\xi,m} H_{A^*}$
I_m	$\{D_{\xi,m}, D_{\xi,m}\}$	$G_{I_m} = 0$	$H_{I_m} = 0$
I_n	$\{D_{\eta,n}, D_{\eta,n}\}$	$G_{I_n} = 0$	$H_{I_n} = 0$

Then, since the diagonal entries of $D_{\xi,m}$ and $D_{\eta,n}$ are complex numbers of modulus 1, we are under the assumptions which led to the displacement equation (2.7) for C^* and can conclude that the mosaic matrix

$$M_C = \begin{bmatrix} I_m & C & 0 \\ C^* & 0 & C^* \\ 0 & I_n & 0 \end{bmatrix},$$

made up with the Cauchy-like blocks resulting from Table 1, satisfies the displacement equation

$$\mathcal{D}_L M_C - M_C \mathcal{D}_R = G_{M_C} H_{M_C}^*$$

with

$$\begin{aligned} \mathcal{D}_L &= D_{\xi,m} \oplus D_{\eta,n} \oplus D_{\eta,n}, \\ \mathcal{D}_R &= D_{\xi,m} \oplus D_{\eta,n} \oplus D_{\xi,m}, \end{aligned} \tag{3.7}$$

and

$$G_{M_C} = \begin{bmatrix} G_C & 0 \\ 0 & G_{C^*} \\ 0 & 0 \end{bmatrix}, \quad H_{M_C} = \begin{bmatrix} 0 & H_{C^*} \\ H_C & 0 \\ 0 & H_{C^*} \end{bmatrix}. \tag{3.8}$$

Moreover, the displacement ranks of M_C and of the original matrix M_A are the same; in particular, when the matrix A is Toeplitz we have $\mathrm{rank}_\Delta(M_C) \le 4$. We stress the fact that, also in this case, the blocks of M_C with coordinates $(1,1)$, $(1,3)$, $(2,2)$, and $(3,2)$ are partially reconstructible Cauchy-like matrices.

We observe that, by (2.7), we have

$$G_{C^*} = D_{\eta,n} H_C \quad \text{and} \quad H_{C^*} = D_{\xi,m}^* G_C,$$

so that a pair of generators for A^* is given by

$$G_{A^*} = Z_{\eta,n} H_A \quad \text{and} \quad H_{A^*} = Z^*_{\xi,m} G_A.$$

From the computational point of view, we also remark that the transformations described above will turn a real matrix A into a complex Cauchy-like matrix C, forcing us to switch to complex arithmetic.

4. Choosing the values of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$. The values of ξ and η chosen in (3.1) determine the displacement structure of the Toeplitz matrix A and, consequently, of the corresponding Cauchy-like matrix C (see Table 1).

The only restriction on the choice of these values, for the moment, is that ξ and η should be complex numbers of unit modulus. This assumption allows us both to employ factorizations (3.6) to transform M_A into the Cauchy-like matrix M_C by a fast and stable computation and to make use of the displacement equation (2.7) for the adjoint matrix C^*.

Anyway, for a Cauchy-like matrix C to be totally reconstructible from its displacement matrices and generators, it is necessary that all the diagonal entries of the first displacement matrix be different from those of the second one. Since we are free to fix ξ and η to suit our needs, we will choose them in order that the matrices $D_{\xi,m}$ and $D_{\eta,n}$ in Table 1 satisfy this condition.

Moreover, since under this assumption all the elements of C are given by

$$C_{ij} = \frac{\boldsymbol{\phi}^*_i \cdot \boldsymbol{\psi}_j}{t_i - s_j}, \qquad i = 1, \dots, m,\ j = 1, \dots, n,$$

with $\boldsymbol{\phi}^*_i$, $\boldsymbol{\psi}^*_j$ being the rows of the generators G_C, H_C and t_i, s_j being the diagonal entries of $D_{\xi,m}$, $D_{\eta,n}$, respectively, we intend to increase steadiness and avoid overflows in the computation by ensuring that the minimum value assumed by the denominator in this expression is as large as possible.

For what follows, it is not restrictive to set $\xi = 1$ and $\eta = \mathrm{e}^{\mathrm{i}\pi\varphi}$, with $\varphi \in (0, 1]$, so that

$$\begin{aligned} t_{k+1} &= \mathrm{e}^{\mathrm{i}\pi \frac{2k}{m}}, \qquad k = 0, \dots, m-1, \\ s_{j+1} &= \mathrm{e}^{\mathrm{i}\pi \frac{\varphi+2j}{n}}, \quad j = 0, \dots, n-1. \end{aligned} \tag{4.1}$$

Our objective is to take the value of φ, which solves the following optimization problem:

$$\max_{\varphi} \min_{k,j} \left| t_{k+1} - s_{j+1} \right|. \tag{4.2}$$

We will show that it is possible to solve (4.2) by the same computational cost required for the evaluation of $\gcd(m, n)$ (the greatest common divisor between m and n), that is, $O(\log m)$ using Euclid's algorithm [30, section 4.5.2, Algorithm X], which is implemented in the `gcd` function of MATLAB [33].

We have

$$\left| t_{k+1} - s_{j+1} \right| = \left| \mathrm{e}^{\frac{\mathrm{i}\pi}{m}[\alpha\varphi - 2(k - j\alpha)]} - 1 \right| = 2 \left| \sin\left(\theta_{(\varphi,k,j)}\right) \right|,$$

where $\alpha = \frac{m}{n} \geq 1$ and

$$\theta_{(\varphi,k,j)} = \frac{\pi}{m} \left[\frac{\alpha\varphi}{2} - (k - j\alpha) \right] \in (-\pi, \pi)$$

for $k = 0, \dots, m-1$ and $j = 0, \dots, n-1$.

For a given φ, the quantity $|t_{k+1} - s_{j+1}|$ reaches its minimum when the angle $\theta_{(\varphi,k,j)}$ is closest to any of the angles $-\pi$, 0, π. The angles $\theta_{(\varphi,k,j)}$ which are closest to $-\pi$ and π are, respectively,

$$\theta_{(\varphi,m-1,0)} = -\pi + \frac{\pi}{m}\left(1 + \frac{\alpha\varphi}{2}\right),$$

$$\theta_{(\varphi,0,n-1)} = \pi - \frac{\pi}{n}\left(1 - \frac{\varphi}{2}\right).$$

The optimal value of φ is obtained by equating the sines of these two angles, obtaining

$$\bar{\varphi} = \frac{m-n}{m} = 1 - \frac{1}{\alpha}.$$

The minimum of (4.2) corresponding to $\bar{\varphi}$ is $2\sin\left(\pi\frac{m+n}{2mn}\right)$.

Let us now consider how to determine φ in order that the value of $\theta_{(\varphi,k,j)}$ closest to zero is as large as possible. When α is an integer, the sequence $\{k - j\alpha\}$ takes integer values and it is immediate to observe that in this case the optimal φ is $\varphi^* = \alpha^{-1}$. For $m = 2n$, for example, we obtain $\varphi^* = \frac{1}{2}$ and $\eta = \mathrm{i}$, while we get $\eta = -1$ for $m = n$ (this is the choice made in [13]).

When α is noninteger, to maximize the angle $\theta_{(\varphi,k,j)}$ for each pair (k, j) we must impose the condition that the ratio $\frac{\alpha\varphi}{2}$ is one-half of the minimum nonzero value taken by $|k - j\alpha|$.

THEOREM 4.1. [illegible] $|k - j\alpha|$ [illegible] $k = 0, \ldots, m-1$ [illegible] $j = 0, \ldots, n-1$ [illegible]

$$\beta = \frac{\gcd(m,n)}{n}.$$

[illegible]. Let $g = \gcd(m,n)$ and $\bar{n} = n/g$. Since $j\alpha$ is noninteger for $j = 1, \ldots, \bar{n}-1$, it is sufficient to consider $j < \bar{n}$.

We have $m \geq n$, so there exist natural numbers r and s such that $m = rn + s$. Moreover, $\ell = s/g$ is integer, so we can write

$$k - j\alpha = k - jr - j\frac{s}{n} = k - jr - \frac{j\ell}{\bar{n}}.$$

Since

$$\{j\ell \bmod \bar{n} : j \in \mathbb{Z}_{\bar{n}}\} = \mathbb{Z}_{\bar{n}},$$

where $\mathbb{Z}_{\bar{n}}$ is the additive group of integers modulo $\bar{n}$ [31], it is immediate to observe that the minimum nonzero value of $|k - j\alpha|$ is

$$\beta = \frac{1}{\bar{n}} = \frac{g}{n}. \qquad \square$$

This result leads to

$$\varphi^* = \frac{\beta}{\alpha} = \frac{\gcd(m,n)}{m}$$

and the corresponding minimum of (4.2) is given by $2\sin\left(\frac{\pi\beta}{2m}\right)$.

Since $\beta \leq 1 \leq \alpha$, letting $\theta^* = \frac{\pi\beta}{2m}$ it is immediate to observe that

$$|\sin(\theta^*)| < \left|\sin\left(\theta_{(\varphi^*,m-1,0)}\right)\right|, \qquad |\sin(\theta^*)| \leq \left|\sin\left(\theta_{(\varphi^*,0,n-1)}\right)\right|,$$

and

$$|\sin(\theta^*)| < \left|\sin\left(\theta_{(\bar{\varphi},m-1,0)}\right)\right| = \left|\sin\left(\theta_{(\bar{\varphi},0,n-1)}\right)\right|,$$

which implies that the optimum of (4.2) is attained for $\varphi = \varphi^*$.

Theorem 4.1 shows that the worst-case scenario is when $\gcd(m, n) = 1$, that is, when m and n are mutually prime numbers. In this case the minimum of (4.2) is $2\sin\left(\frac{\pi}{2mn}\right)$.

5. Schur complementation. The [illegible] is a fast method for computing the LU factorization of a matrix A with displacement structure, through recursive Schur complementation, which operates only on the displacement matrices and generators of A.

In this paper we are not directly interested in LU factorization, but only in the computation of the Schur $(m + n)$-complement of the mosaic matrices M_C (2.5) and M_A (3.3). The outline of the Schur algorithm, when applied to M_C, is the following:

for $k = 1, \ldots, m + n$
  extract the kth column of M_C from the displacement data
  extract the kth row of M_C from the displacement data
  update the left generator G_{M_C}
  update the right generator H_{M_C}

In principle, this procedure could be applied directly to M_C. Anyway, by exploiting its particular mosaic structure and the large number of null entries, it is possible to optimize the algorithm in order to reduce the computational load.

Moreover, as already pointed out in [13, 20], it is simple to employ partial pivoting to improve the stability of the Schur algorithm, when applied to Cauchy-like matrices, as this displacement structure is pivoting-invariant. However, since the entries of the Schur complement $\mathcal{S}_r(A)$ of a given matrix A are, in general, modified by rows pivoting, we must restrict its action to the first r rows of A. In fact, this procedure is equivalent to the matrix product

$$PA = \begin{bmatrix} P_r & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} P_r A_{11} & P_r A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \tag{5.1}$$

where P_r is a permutation matrix of dimension r, and it is immediate to verify that $\mathcal{S}_r(PA) = \mathcal{S}_r(A)$.

As to M_A, its displacement structure seems too complicated for a direct application of the Schur algorithm and, in any case, the structure would be destroyed by pivoting. For these reasons, following the idea introduced in [13, 20], we will first convert M_A into a Cauchy-like matrix by the technique described in section 3, and then apply our optimized algorithm to the resulting matrix.

The algorithm takes as input the displacement matrices (3.7) and the generators (3.8) of M_C and returns the displacement matrices and generators of the

for $i = 1, \dots, m+n$
  $p_i = i$
for $i = 1, \dots, n$
  $d_i = 0$
  $\delta_i = \frac{\boldsymbol{\phi}^*_{m+i} \cdot \boldsymbol{\psi}_{m+n+1}}{t_{m+i} - s_{m+n+1}}$
  $v_i = 1$

FIG. 1. *Initialization of the algorithm.*

Schur complement $\mathcal{S}_{m+n}(M_C)$. We will represent the input matrices in the form

$$\begin{aligned} \mathcal{D}_L &= \operatorname{diag}(t_1, \dots, t_{m+2n}), \\ \mathcal{D}_R &= \operatorname{diag}(s_1, \dots, s_{2m+n}), \\ G^*_{M_C} &= \begin{bmatrix} \boldsymbol{\phi}_1 & \cdots & \boldsymbol{\phi}_{m+2n} \end{bmatrix}, \\ H^*_{M_C} &= \begin{bmatrix} \boldsymbol{\psi}_1 & \cdots & \boldsymbol{\psi}_{2m+n} \end{bmatrix}, \end{aligned} \tag{5.2}$$

where $\boldsymbol{\phi}_i, \boldsymbol{\psi}_j \in \mathbb{C}^{2\delta}$ and δ is the displacement rank of C. Notice that, for the sake of simplicity, we use the symbols t_i, s_j, $\boldsymbol{\phi}_i$, and $\boldsymbol{\psi}_j$ to refer to the displacement structure of the mosaic matrix M_C, and not of the matrix C as before.

Since some of the blocks of M_C are partially reconstructible, it is necessary to use some additional vectors for its storage. The diagonal entries of the block $(2,2)$ will be stored in the vector $\mathbf{d} = (d_1, \dots, d_n)^T$. Their value is initially zero, but pivoting will cause it to change. On the contrary, there is no need to store additional information for the other partially reconstructible blocks, namely, the blocks with coordinates $(1,1)$, $(1,3)$, and $(3,2)$.

A particular treatment is requested for block $(2,3)$. In fact, while at the start of the process it can be totally reconstructed from the generators, as a consequence of pivoting some of the diagonal components of its left displacement matrix may become equal to one of the entries of the right one. In order to be able to reconstruct this block, we use a vector $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)^T$ to store one element for each of its rows,

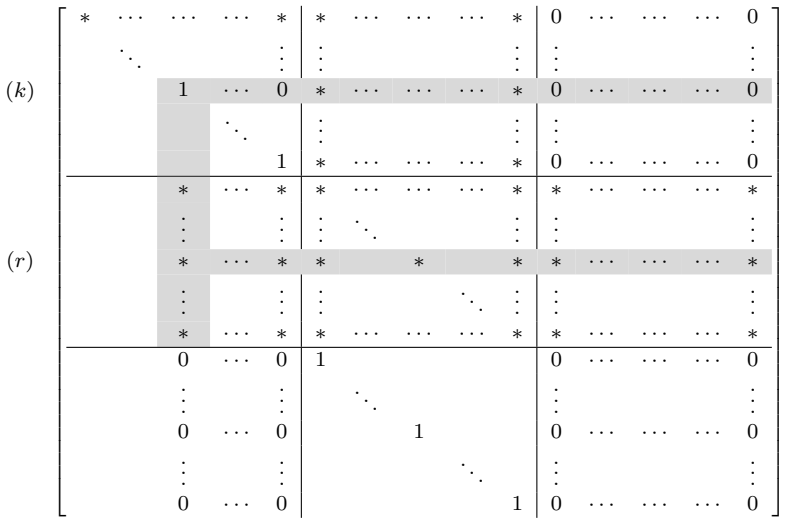


FIG. 2. *kth step of first phase of the algorithm.*

and a vector of indexes $\mathbf{v} = (v_1, \ldots, v_n)^T$ to keep track of their position on each row. Namely, the variable δ_i contains the element $(M_C)_{m+i,m+n+v_i}$.

The initialization of the algorithm, reported in Figure 1, consists of setting up a vector $\mathbf{p}$, which records the permutation introduced by pivoting, and assigning initial values to the vectors $\mathbf{d}$, $\boldsymbol{\delta}$, and $\mathbf{v}$. As can be seen, $\boldsymbol{\delta}$ is initialized with the first column of block $(2, 3)$.

The rest of the algorithm can be divided in two phases. In the first phase only the first m columns of M_C are processed. The kth step is represented in Figure 2, while the algorithm is described in Figure 3. First of all, the nonzero entries of the kth column are computed and the element of maximum modulus is located. Notice that, in this phase, pivoting is naturally limited to the first $m + n$ rows, as the last n components of the kth column are always zero when $k = 1, \ldots, m$.

If the maximum is already in the right position, the kth row can be immediately constructed. If, on the contrary, we are going to exchange the kth and the rth rows,

for $k = 1, \ldots, m$
  $\ell_k = 1$
  for $i = m + 1, \ldots, m + n$
    $\ell_i = \frac{\boldsymbol{\phi}_i^* \cdot \boldsymbol{\psi}_k}{t_i - s_k}$
  find r such that $|\ell_r| = \max_{i=k,m+1,\ldots,m+n} |\ell_i|$
  if $r = k$
    $u_k = 1$
    for $j = k + 1, \ldots, m, m + n + 1, \ldots, 2m + n$
      $u_j = 0$
    for $j = m + 1, \ldots, m + n$
      $u_j = \frac{\boldsymbol{\phi}_k^* \cdot \boldsymbol{\psi}_j}{t_k - s_j}$
  else
    if $p_r = r$
      $w = r$
      $u_r = d_{r-m}$
    else
      $w = m + n + v_{r-m}$
      $u_w = \delta_{r-m}$
    for $j = k, \ldots, w - 1, w + 1, \ldots, 2m + n$
      $u_j = \frac{\boldsymbol{\phi}_r^* \cdot \boldsymbol{\psi}_j}{t_r - s_j}$
    $d_{r-m} = \frac{\boldsymbol{\phi}_k^* \cdot \boldsymbol{\psi}_r}{t_k - s_r}$
    $\delta_{r-m} = 0$
    $v_{r-m} = k$
    swap (ℓ_k, ℓ_r)
    swap $(\boldsymbol{\phi}_k, \boldsymbol{\phi}_r)$
    swap (t_k, t_r)
    swap (p_k, p_r)
  for $i = m + 1, \ldots, m + n$
    $\ell_i = \ell_i / \ell_k$
    $\boldsymbol{\phi}_i = \boldsymbol{\phi}_i - \ell_i \boldsymbol{\phi}_k$
    $d_{i-m} = d_{i-m} - \ell_i u_i$
    $\delta_{i-m} = \delta_{i-m} - \ell_i u_{m+n+v_{i-m}}$
  for $j = k + 1, \ldots, 2m + n$
    $\boldsymbol{\psi}_j = \boldsymbol{\psi}_j - \frac{u_j}{\ell_k} \boldsymbol{\psi}_k$

FIG. 3. *First phase of the algorithm.*

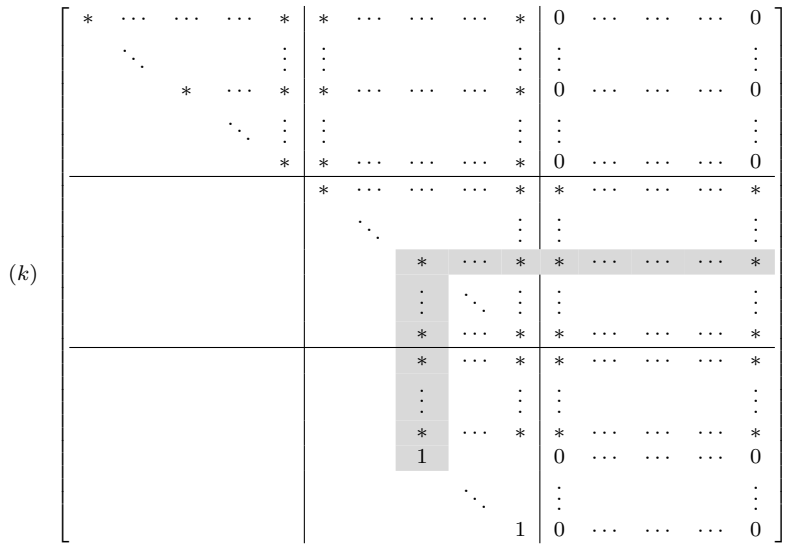


FIG. 4. *kth step of second phase of the algorithm.*

different procedures for computing the new rth row must be adopted depending on whether the rth row has already been moved at a previous step or not. When this is done, we can perform pivoting by swapping the kth and the rth rows of the kth column $\boldsymbol{\ell}$ of M_C, the left generator G_{M_C}, and the vectors $\mathbf{t} = (t_1, \ldots, t_{m+2n})^T$ and $\mathbf{p}$. Pivoting also causes the auxiliary vectors $\mathbf{d}$, $\boldsymbol{\delta}$, and $\mathbf{v}$ to be updated and, in general, changes the $(1,1)$-block from an identity into a generic upper triangular matrix. Finally, Gauss reduction is carried out by modifying the generators of M_C and, again, updating $\mathbf{d}$ and $\boldsymbol{\delta}$.

In the second phase, the iteration index k ranges from $m+1$ to $m+n$ (see Figures 4 and 5). This part of the algorithm shows some small differences with respect to the first phase, due to the need of keeping account of the particular structure of the blocks that are being modified. Notice also that, in this phase, pivoting must be restricted to the first $m+n$ rows of M_C (see (5.1)).

If we denote by $t_i^{(k)}$ and $s_j^{(k)}$ the diagonal entries of the displacement matrices of M_C at the kth iteration, and by $\boldsymbol{\phi}_i^{(k)}$ and $\boldsymbol{\psi}_j^{(k)}$ the columns of $G_{M_C}^*$ and $H_{M_C}^*$ at the same iteration, then at the end of the process the displacement matrices and the generators of the Schur complement $\mathcal{S}_{m+n}(M_C)$ are given by

$$
\begin{aligned}
D_1 &= \operatorname{diag}\left(t_{m+n+1}^{(m+n)}, \ldots, t_{m+2n}^{(m+n)}\right), \\
D_2 &= \operatorname{diag}\left(s_{m+n+1}^{(m+n)}, \ldots, s_{2m+n}^{(m+n)}\right), \\
G_S^* &= \begin{bmatrix} \boldsymbol{\phi}_{m+n+1}^{(m+n)} & \cdots & \boldsymbol{\phi}_{m+2n}^{(m+n)} \end{bmatrix}, \\
H_S^* &= \begin{bmatrix} \boldsymbol{\psi}_{m+n+1}^{(m+n)} & \cdots & \boldsymbol{\psi}_{2m+n}^{(m+n)} \end{bmatrix}.
\end{aligned}
\tag{5.3}
$$

The above algorithm, at worst, involves $3\gamma(m+n)^2 + (2m+n)n$ additions, $\frac{7}{2}(\gamma+1)(m+n)^2 + (\gamma-1)m(m+n) - m^2$ multiplications, and $\frac{1}{2}(m+n)^2$ complex modulus computations, γ being the displacement rank of M_C, i.e., the dimension of

```
for k = m+1, ..., m+n
    ℓ_k = d_{k-m}
    ℓ_{n+k} = 1
    for i = k+1, ..., n+k-1
        ℓ_i = (φ_i^* · ψ_k) / (t_i - s_k)
    find r such that |ℓ_r| = max_{i=k,...,m+n} |ℓ_i|
    if r = k
        if p_k = k
            u_k = d_{k-m}
            for j = k+1, ..., 2m+n
                u_j = (φ_k^* · ψ_j) / (t_k - s_j)
        else
            w = m+n+v_{k-m}
            u_w = δ_{k-m}
            for j = k, ..., w-1, w+1, ..., 2m+n
                u_j = (φ_k^* · ψ_j) / (t_k - s_j)
    else
        if p_r = r
            w = r
            u_r = d_{r-m}
        else
            w = m+n+v_{r-m}
            u_w = δ_{r-m}
        for j = k, ..., w-1, w+1, ..., 2m+n
            u_j = (φ_r^* · ψ_j) / (t_r - s_j)
        d_{r-m} = (φ_k^* · ψ_r) / (t_k - s_r)
        δ_{r-m} = δ_{k-m}
        v_{r-m} = v_{k-m}
        swap (ℓ_k, ℓ_r)
        swap (φ_k, φ_r)
        swap (t_k, t_r)
        swap (p_k, p_r)
    for i = k+1, ..., n+k
        ℓ_i = ℓ_i / ℓ_k
        φ_i = φ_i - ℓ_i φ_k
    for i = k+1, ..., m+n
        d_{i-m} = d_{i-m} - ℓ_i u_i
        δ_{i-m} = δ_{i-m} - ℓ_i u_{m+n+v_{i-m}}
    for j = k+1, ..., 2m+n
        ψ_j = ψ_j - (u_j / ℓ_k) ψ_k
```

FIG. 5. *Second phase of the algorithm.*

vectors $\boldsymbol{\phi}_i$ and $\boldsymbol{\psi}_j$ in (5.2). This means that its complexity as measured in [illegible],[1] considering 2 flops for a complex sum, 6 for a product, and 4 for each modulus, is

$$O\left((27\gamma + 23)(m+n)^2 + 6(\gamma - 1)m(m+n) + 2(2mn + n^2 - 3m^2)\right).$$

This yields $143m^2 + 284mn + 133n^2$ flops when $\gamma = 4$, that is, when M_C comes from the Toeplitz system (3.2).

[1] By a *flop* we mean a real floating point operation of any kind.

6. Computation of the least-squares solution. When A is a Toeplitz matrix, by exploiting the factorizations reported in Table 1, we obtain the following expression for the least-squares solution of (3.2):

$$\mathbf{x}_{LS} = (A^*A)^{-1}A^*\mathbf{b} = \mathcal{F}^*_{\eta,n}(C^*C)^{-1}C^*\mathcal{F}_{\xi,m}\mathbf{b} = \mathcal{F}^*_{\eta,n}\mathcal{S}_{m+n}(M_C)\mathcal{F}_{\xi,m}\mathbf{b}, \tag{6.1}$$

which also shows the relation between the Schur complements of the two augmented matrices M_A and M_C,

$$\mathcal{S}_{m+n}(M_A) = \mathcal{F}^*_{\eta,n}\mathcal{S}_{m+n}(M_C)\mathcal{F}_{\xi,m}.$$

In (6.1), the products times the scaled Fourier matrices can be accomplished in FFT time, while $\mathcal{S}_{m+n}(M_C)$, stored by means of its displacement matrices (D_1, D_2) and generators (G_S, H_S) (see (5.3)), is obtained by the algorithm exposed in the previous section.

Letting $S = \mathcal{S}_{m+n}(M_C)$, the matrix-vector product $S\mathbf{z}$, for any $\mathbf{z} \in \mathbb{C}^m$, can be expressed as the Schur complement $\mathcal{S}_m(B)$, where

$$B = \begin{bmatrix} -I_m & \mathbf{z} \\ S & 0 \end{bmatrix} \in \mathbb{C}^{(m+n)\times(m+1)}.$$

The matrix B is Cauchy-like (with displacement rank 5 if we start from the Toeplitz system (3.2)) as it satisfies the displacement equation

$$\Delta_1 B - B\Delta_2 = G_B H_B^*,$$

with

$$\Delta_1 = D_2 \oplus D_1, \qquad \Delta_2 = D_2 \oplus 0,$$
$$G_B = \begin{bmatrix} 0 & D_2\mathbf{z} \\ G_S & 0 \end{bmatrix}, \quad H_B = \begin{bmatrix} H_S & 0 \\ 0 & 1 \end{bmatrix},$$

so it is particularly convenient to evaluate $\mathbf{x}_{LS}$ by the Schur algorithm. This computation can be performed by the standard version of the algorithm, which we do not report here for the sake of brevity. Notice that this time, given the structure of B, pivoting is not applicable.

REMARK 6.1. [illegible] A [illegible] $\mathbf{b}_i$ $i = 1, \ldots, K$ [illegible] $\mathcal{S}_{m+n}(M_C)$ [illegible] (6.1) [illegible] $(m + 2n) \times (m + n + 1)$ [illegible]

$$N_A = \begin{bmatrix} I_m & A & 0 \\ A^* & 0 & A^*\mathbf{b} \\ 0 & I_n & 0 \end{bmatrix}$$

[illegible] $\mathbf{x}_{LS} = \mathcal{S}_{m+n}(N_A)$ [illegible] pseudoinverse $A^\dagger$ [illegible] $(m + n)$ [illegible]

$$\begin{bmatrix} I_m & A & -I_m \\ A^* & 0 & 0 \\ 0 & I_n & 0 \end{bmatrix}.$$

M_A G_{M_A}

$$H = \begin{bmatrix} 0 & H_{A^*} \\ H_A & 0 \\ 0 & 0 \end{bmatrix}.$$

M_A

7. Numerical results. The algorithm described in the previous sections, which in the following will be denoted TLLS (for), has been implemented in MATLAB [33] in two versions: with and without partial pivoting. We start comparing its performance to two classical direct methods for the least-squares solution of unstructured overdetermined linear systems, namely, the QR factorization of the matrix A in (3.2), followed by the solution of the resulting triangular system, and the solution of the normal equations by means of Cholesky factorization [1, 15]. This choice is motivated by the fact that we consider of primary importance to preliminarily ascertain to which extent the structured approach is favorable with respect to the standard solution methods as to stability, speed of computation, and storage requirements.

Figure 6 shows a comparison between the theoretical computational complexity of TLLS, Householder QR factorization ($O(2n^2(m - \frac{1}{3}n))$), and Cholesky factorization ($O(n^2(m + \frac{1}{3}n))$, including the construction of A^TA), with respect to the variation of n, taking $m = 2n$ and $m = 10n$. In the first case the new algorithm has a lower complexity than both unstructured methods for $n \gtrsim 600$, in the second for $n \gtrsim 2000$. These results were confirmed experimentally through the `flops` counter of MATLAB 5 [32]. A comparison based on a time measurement is actually infeasible since our implementation, being written in the MATLAB programming language, is incomparably slower than the algorithms directly coded in the MATLAB kernel. This hindrance will be overcome as soon as our program is translated into C language and linked to MATLAB through the MEX ( ) interface library [33].

It is clear that the complexity for Householder QR and Cholesky factorizations is linear in m, while it is quadratic for TLLS. This means that for fixed n, when we

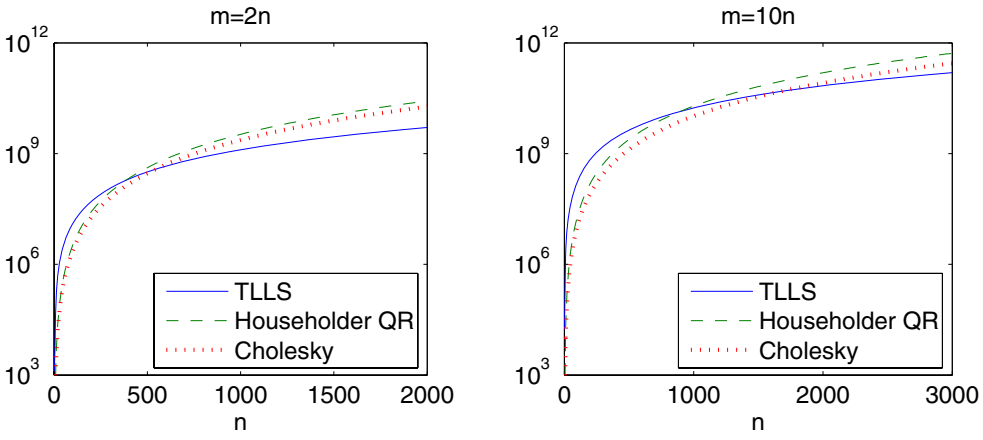


FIG. 6. *Comparison of computational complexity.*

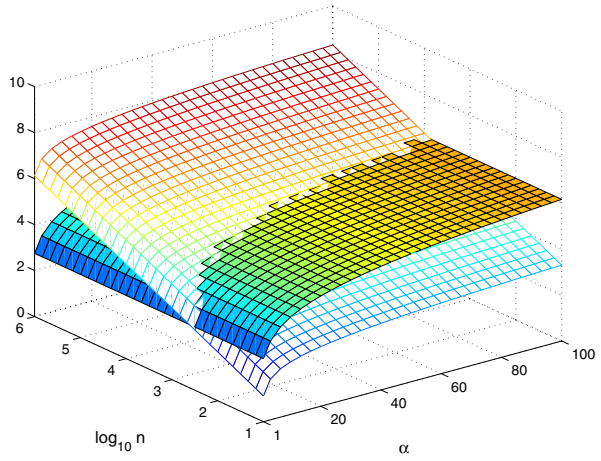


FIG. 7. *Variation of computational complexity with respect to* $\alpha = \frac{m}{n}$ *and* $\log_{10} n$ *for TLLS (shaded surface) and Householder QR (wiregrid surface).*

increase m there will always be a point where QR and Cholesky factorizations are more efficient than TLLS. Setting $\alpha = \frac{m}{n}$, the complexities of Householder QR and TLLS are, respectively,

$$n^2 \left(2n\alpha - \frac{2}{3}n \right) \quad \text{and} \quad n^2(143\alpha^2 + 284\alpha + 133).$$

In Figure 7 we plot these two quantities in logarithmic scale, neglecting the common factor n^2, with respect to the variation of α and n. The graph confirms the predicted behavior for the complexity, but, at the same time, it shows that the crossover grows rapidly with n. For example, when $n > 10000$ the QR factorization can be more efficient than TLLS only if $m \gg 100n$.

Regarding the storage needed by the TLLS method, which from this point of view is obviously far more convenient than any unstructured approach, it requires order $m+n$ floating point variables (the larger arrays used, i.e., the two generators of M_A, take $4(m+2n)$ and $4(2m+n)$ complex variables, respectively).

To illustrate the accuracy of the algorithm, we report some results concerning the solution in the least-squares sense of overdetermined Toeplitz linear systems

$$A\mathbf{x} = \mathbf{b},$$

where the right-hand side $\mathbf{b}$ corresponds to the exact solution $\mathbf{e} = (1, 1, \ldots, 1)^T$. The matrix $A = (a_{i-j})$ belongs to a well-known class of Toeplitz test matrices, namely, the [illegible] [24, 39]

$$a_{i-j} = \rho^{|i-j|}, \qquad i = 1, \ldots, m,\ j = 1, \ldots, n,$$

depending on the parameter $\rho \in (0, 1)$. These matrices are positive definite when $m = n$ and ill-conditioned when $\rho \simeq 1$.

Each test problem has been solved by the four methods considered, that is, the TLLS algorithm, TLLS without pivoting, QR factorization, and the solution of the

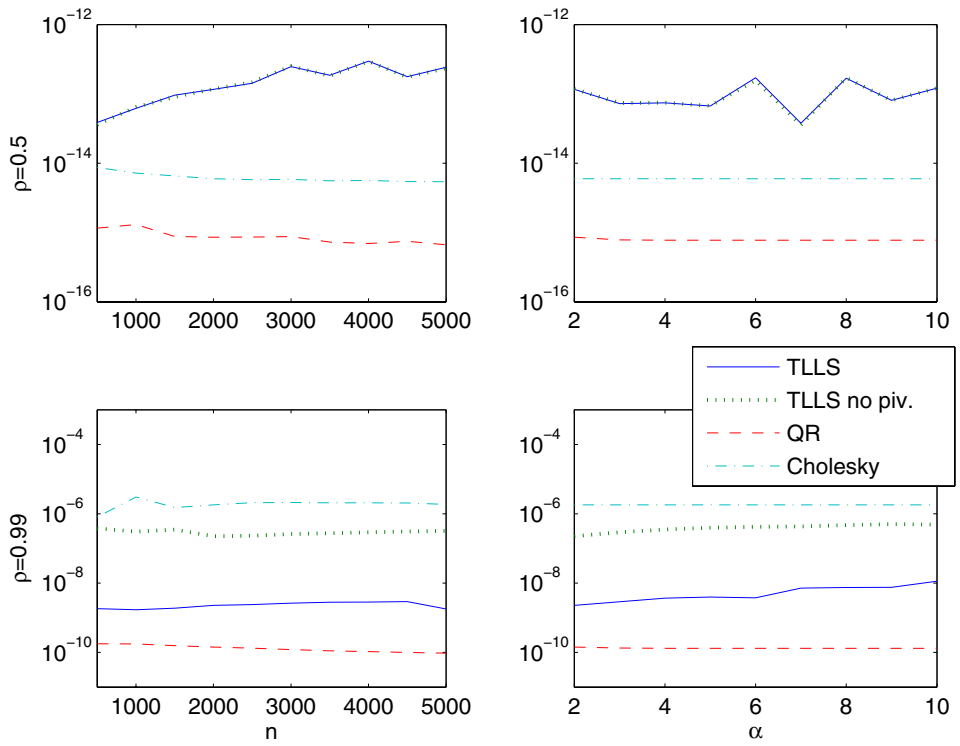


FIG. 8. *Relative error for the KMS matrix: top row,* $\rho = 0.5$*; bottom row,* $\rho = 0.99$*; left column,* $n = 500, 1000, \ldots, 5000$ *and* $m = 2n$*; right column,* $n = 2000$ *and* $\alpha = \frac{m}{n} = 2, 3, \ldots, 10$.

normal equations by means of Cholesky factorization. To measure the accuracy of the results, we adopted the relative error

$$\frac{\|\mathbf{e} - \mathbf{x}\|_2}{\|\mathbf{e}\|_2}$$

between the computed solution $\mathbf{x}$ and the exact solution $\mathbf{e}$.

Figure 8 shows the relative error corresponding to the KMS matrix with $\rho = 0.5$ or $\rho = 0.99$ (the condition numbers are about 9 and $3.7 \cdot 10^4$, respectively, when $m = n = 1000$). The results in the left column are obtained by letting n range from 500 to 5000 with $m = 2n$; in the right column we fixed $n = 2000$ and let $\alpha = \frac{m}{n} = 2, 3, \ldots, 10$. The new algorithm does not appear to be very sensitive to the changes in the dimension or to the "rectangularity" of the matrix A, i.e., the ratio between the number of rows and columns. At the same time, as one would expect, when the problem is well-conditioned the effect of pivoting is negligible and the structured algorithm is slightly less accurate than the unstructured approaches. On the contrary, when the conditioning of the problem gets worse, pivoting is essential and the TLLS algorithm is more accurate than Cholesky and comparable with QR factorization.

The influence of conditioning on the performance of the method is further investigated in the left graph of Figure 9, where the four methods are tested on a problem

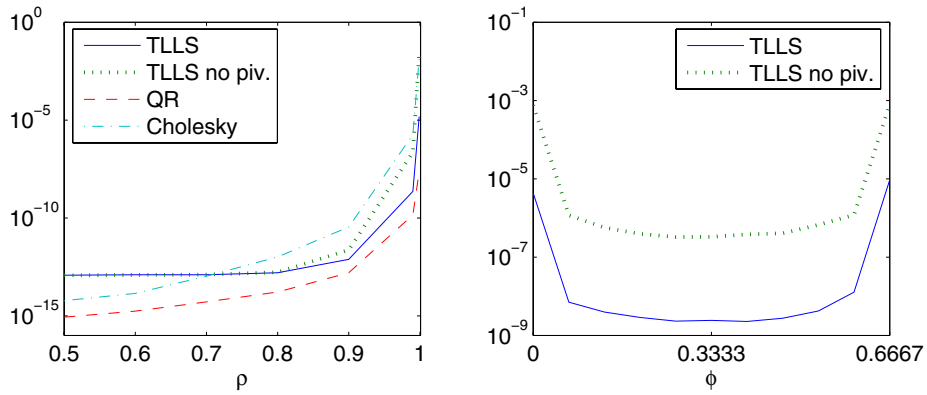


FIG. 9. *On the left, variation of the relative error with respect to the KMS matrix parameter:* $m = 4000$, $n = 2000$, $\rho = 0.5, 0.6, \ldots, 0.9, 0.99, 0.999$. *On the right, influence of the parameter* φ *on the performance of the method:* $m = 3000$, $n = 1000$, $\varphi^* = \frac{1}{3}$.

of dimension $(m, n) = (4000, 2000)$ with ρ varying on the interval $(0.5, 1)$ and the condition number ranging from 9 to $2.5 \cdot 10^6$. Again, we observe that when the value of ρ approaches 1, while all the methods lose accuracy because of ill-conditioning, the effect of pivoting is more appreciable and the results furnished by TLLS and QR factorization get closer.

In the right graph of Figure 9 we illustrate the effect on the quality of the results of the parameter φ, which determines the displacement structure of the matrix A (see (3.1) and (4.1)). In this case the TLLS method is applied with different values of φ to a problem with $(m, n) = (3000, 1000)$ and $\rho = 0.99$ (the condition number is $3.7 \cdot 10^4$). The optimal value of φ, whose computation is described in section 4, is $\varphi^* = \frac{1}{3}$, while for $\varphi = 0$ or $\varphi = \frac{2}{3}$ the Cauchy-like matrix C in which the matrix A is transformed (see Table 1) is partially reconstructible and our method is not applicable. We see that the relative errors increase when φ is near the two endpoints of the interval, while the minimum is reached for $\varphi \simeq \varphi^*$.

The 2-norm condition number which describes the sensitivity of the solution of the least-squares problem $\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|$ with respect to a perturbation in A is

$$\kappa_{LS} = \kappa(A) + \frac{\kappa(A)^2 \tan\theta}{\eta},$$

where $\kappa(A) = \|A\| \cdot \|A^\dagger\|$, $\eta = \frac{\|A\| \cdot \|\mathbf{x}\|}{\|A\mathbf{x}\|} \in [1, \kappa(A)]$, and

$$\theta = \arccos \frac{\|A\mathbf{x}\|}{\|\mathbf{b}\|} \in \left[0, \frac{\pi}{2}\right]$$

is the angle between the right-hand side $\mathbf{b}$ and its projection $A\mathbf{x}$ on the range of A [38, Theorem 18.1]. The solution computed by Householder QR factorization, which is a backward stable algorithm, is influenced by κ_{LS}, so it depends on the two parameters θ and η. On the contrary, the solution computed by Cholesky factorization is affected by the condition number for normal equations, that is, $\kappa(A)^2$, independently on θ and η.

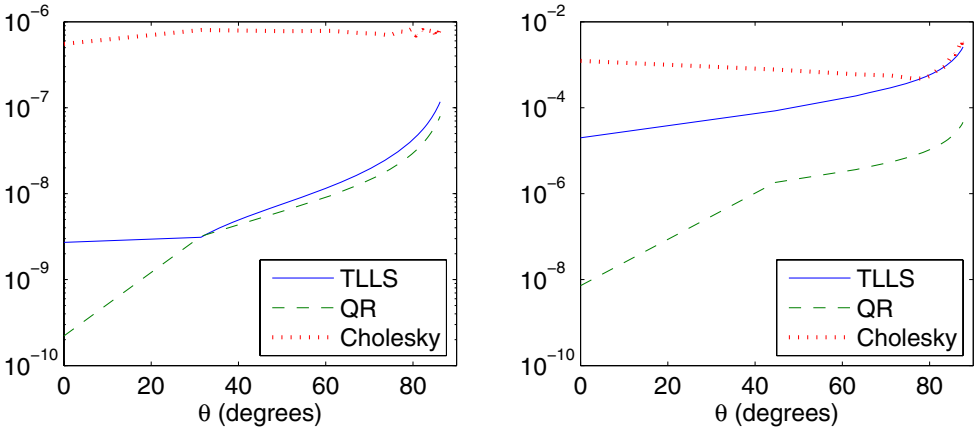


FIG. 10. *Effect of the conditioning, resulting from nonzero residuals, on the relative error. KMS matrix,* $m = 1500$, $n = 500$, $\rho = 0.99$ *(on the left) and* $\rho = 0.999$ *(on the right). The meaning of* θ *is explained in the text.*

To investigate the behavior of our algorithm to this respect, with a test matrix A fixed, we construct a vector $\mathbf{q}$ orthogonal to the range of A by applying the modified Gram–Schmidt orthogonalization method [1] to the augmented matrix $[A|\mathbf{e}_m]$, with $\mathbf{e}_m = (1, \ldots, 1)^T \in \mathbb{R}^m$, selecting the last column of the result and normalizing it. Then, for $\mathbf{y} = A\mathbf{e}_n$, we compute $\mathbf{b} = \mathbf{y} + \tau\mathbf{q}$, letting τ take different values in order to vary the angle θ between $\mathbf{b}$ and $\mathbf{y}$ while preserving the same solution $\mathbf{e}_n$ of the least-squares problem. Figure 10 shows the relative error in the solution when A is the KMS matrix with $(m, n) = (1500, 500)$ and $\rho = 0.99, 0.999$ (the value of $\kappa(A)$ is $3.3 \cdot 10^4$ and $1.1 \cdot 10^6$, respectively). In this test $\eta \simeq 1$ and we let the parameter τ vary so that the angle θ ranges between 0 and 86 degrees. It is clear that the two nonstructured methods behave as predicted. In fact, the error obtained by QR factorization grows with θ and approaches the error coming from the normal equations, which is worse uniformly in θ. The TLLS method appears to produce an error which is sensitive on θ and, for $\rho = 0.99$, behaves essentially the same as QR. In the second case, where the condition number of A is larger, TLLS is much less accurate than QR and comparable with Cholesky, but the trend of the error is still remarkable.

This result, as well as the error graphs reported in Figures 8 and 9, suggests that the TLLS method, even if formally equivalent to the solution of the normal equations, does not inherit from them the bad behavior connected to the squaring of the condition number. A possible explanation for the better performance of TLLS may consist of the fact that it solves the normal equations implicitly without effectively computing the matrix A^*A.

It would be desirable to compare the TLLS algorithm to other existing fast and superfast methods. Unfortunately, almost no software based on the algorithms cited in section 1 seems to be publicly available. The only computer program I was able to get is about the method described in [40]. Marc Van Barel, in fact, was so kind to send me copies of the MATLAB functions used to test this superfast algorithm.

Figure 11 shows the results obtained by applying TLLS and the superfast algorithm to the solution of a linear system with a random Toeplitz matrix of dimension $m \times n$, with $n = 2^k$, $k = 5, \ldots, 13$, and $m = 2n$. The graphs suggest that even

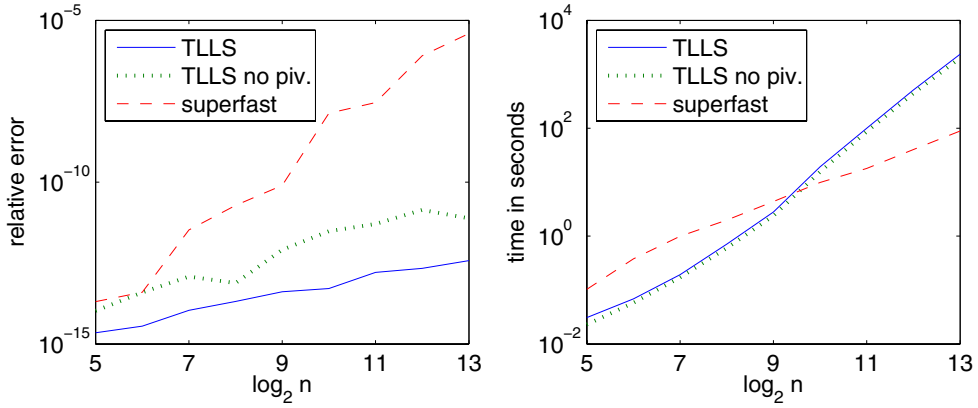


FIG. 11. *Comparison of TLLS and a superfast method.*

though the superfast algorithm is faster than TLLS for $n > 2^9$, at the same time it appears to be less stable, at least on this example. This confirms the fact that the choice of an algorithm cannot be an absolute decision, but it should be guided by the peculiarities of each given problem (dimension, storage, required accuracy, need for fast computation, etc.).

It must be stressed that the implementation of the superfast method is research software, probably not fully optimized and not intended to be officially released, so no real conclusion can be deduced from this numerical experiment. Moreover both programs are widely based on , . loops and written in the MATLAB programming language, so that they are intrinsically slow. It should also be added that one or two steps of iterative refinement could improve the accuracy of both of the algorithms with low computational cost. More work would be required to implement these and other methods in a compiled language and to perform a wide numerical simulation on them.

8. Extensions and future work. The performance of the TLLS method for the least-squares solution of overdetermined Toeplitz linear systems looks promising, as the numerical results show. The next step in its development is to implement the algorithm either in C or Fortran and to compare it with other fast and superfast methods for structured least-squares problems. Moreover, the algorithm could be easily modified to deal with matrices which have a displacement structure different from Toeplitz-like, but which can be converted into Cauchy-like matrices [13, 21].

In the future, we plan to ascertain if it is possible to improve the algorithm in terms of accuracy and speed by using real transforms to convert a matrix from Toeplitz-like to Cauchy-like, instead of complex ones [21]. To be able to solve problems of huge dimension it would also be very important to study the application of total pivoting, or at least an approximation of it which does not increase excessively the complexity of the algorithm.

However, the extension which is of particular interest to us is the application to Tikhonov regularization [37], that is, the solution of the minimum problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \|A\mathbf{x} - \mathbf{b}\|^2 + \lambda^2 \|H\mathbf{x}\|^2 \right\}, \tag{8.1}$$

where $A \in \mathbb{R}^{m\times n}$, $H \in \mathbb{R}^{p\times n}$ $(m \geq n \geq p)$, and λ is a regularization parameter. In fact, it has already been observed in [25] that the solution operator of (8.1) is the Schur $(m+n+p)$-complement of the augmented matrix

$$\left[\begin{array}{ccc|c} I_m & 0 & A & 0 \\ 0 & I_p & \lambda H & 0 \\ A^* & \lambda H^* & 0 & A^* \\ \hline 0 & 0 & I_n & 0 \end{array}\right]$$

(the black lines highlight how the matrix is partitioned).

There are, anyway, other computations connected to Tikhonov regularization which can be expressed as Schur complements. Generalized cross validation (GCV) [9] is a technique used to estimate the optimal value of the regularization parameter. It consists of minimizing the function

$$V(\lambda) = \frac{\frac{1}{m}\|(I - A(\lambda))\mathbf{b}\|^2}{\left[\frac{1}{m}\operatorname{trace}(I - A(\lambda))\right]^2},$$

where

$$A(\lambda) = A(A^*A + \lambda^2 H^*H)^{-1}A^*$$

is called the [illegible]. The computation of the GCV function, which must be repeated many times during the minimum search, is generally performed by employing the [illegible] (GSVD) of the matrix pair (A, H) [15], but this approach cannot be applied when the dimensions of A are large. An alternative algorithm could be constructed by expressing $A(\lambda)$ as the Schur $(m+n+p)$-complement of

$$\left[\begin{array}{ccc|c} I_m & 0 & A & 0 \\ 0 & I_p & \lambda H & 0 \\ A^* & \lambda H^* & 0 & A^* \\ \hline 0 & 0 & A & 0 \end{array}\right].$$

Finally, in multiparameter regularization [3] one or more regularization terms are added to the Tikhonov function, like in

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{\|A\mathbf{x} - \mathbf{b}\|^2 + \lambda^2\|H\mathbf{x}\|^2 + \mu^2\|K\mathbf{x}\|^2\right\},$$

with $A \in \mathbb{R}^{m\times n}$, $H \in \mathbb{R}^{p\times n}$, and $K \in \mathbb{R}^{q\times n}$. Again, the solution

$$\mathbf{x} = \left(A^*A + \lambda^2 H^*H + \mu^2 K^*K\right)^{-1} A^*\mathbf{b}$$

can be computed as the Schur $(m+n+p+q)$-complement of

$$\left[\begin{array}{cccc|c} I_m & 0 & 0 & A & 0 \\ 0 & I_p & 0 & \lambda H & 0 \\ 0 & 0 & I_q & \mu K & 0 \\ A^* & \lambda H^* & \mu K^* & 0 & A^*\mathbf{b} \\ \hline 0 & 0 & 0 & I_n & 0 \end{array}\right].$$

Acknowledgments. This paper is dedicated to the memory of Georg Heinig, a great man without whom this research (and many others) wouldn't have been possible. I would like to thank the three referees, whose comments and corrections led to an improvement in the quality of the paper and an extension of its content. I am particularly grateful to Marc Van Barel for sending me the software developed in [40] and allowing me to use it for numerical experiments.

REFERENCES

[1] A. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996.
[2] A. Bojanczyk, R. P. Brent, and F. de Hoog, *QR factorization of Toeplitz matrices*, Numer. Math., 49 (1986), pp. 81–94.
[3] C. Brezinski, M. Redivo-Zaglia, G. Rodriguez, and S. Seatzu, *Multi-parameter regularization techniques for ill-conditioned linear systems*, Numer. Math., 94 (2003), pp. 203–228.
[4] R. H. Chan, J. G. Nagy, and R. J. Plemmons, *Circulant preconditioned Toeplitz least squares iterations*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 80–97.
[5] S. Chandrasekaran and A. H. Sayed, *A fast stable solver for nonsymmetric Toeplitz and quasi-Toeplitz systems of linear equations*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 107–139.
[6] J. Chun and T. Kailath, *Displacement structure for Hankel, Vandermonde, and related (derived) matrices*, Linear Algebra Appl., 151 (1991), pp. 199–227.
[7] J. Chun and T. Kailath, *Divide-and-conquer solutions of least-squares problems for matrices with displacement structure*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 128–145.
[8] J. Chun, T. Kailath, and H. Lev-Ari, *Fast parallel algorithms for QR and triangular factorization*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 899–913.
[9] P. Craven and G. Wahba, *Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross validation*, Numer. Math., 31 (1979), pp. 377–403.
[10] G. Cybenko, *A general orthogonalization technique with applications to time series analysis and signal processing*, Math Comp., 40 (1983), pp. 323–336.
[11] G. Cybenko, *Fast Toeplitz orthogonalization using inner products*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 734–740.
[12] B. Friedlander, M. Morf, T. Kailath, and L. Ljung, *New inversion formulas for matrices classified in terms of their distance from Toeplitz matrices*, Linear Algebra Appl., 27 (1979), pp. 31–60.
[13] I. Gohberg, T. Kailath, and V. Olshevsky, *Fast Gaussian elimination with partial pivoting for matrices with displacement structure*, Math. Comp., 64 (1995), pp. 1557–1576.
[14] I. Gohberg and V. Olshevsky, *Circulants, displacements and decompositions of matrices*, Integral Equations Operator Theory, 15 (1992), pp. 730–743.
[15] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[16] M. Gu, *New fast algorithms for structured linear least squares problems*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 244–269.
[17] G. Heinig, *Displacement structure of generalized inverse matrices*, Linear Algebra Appl., 211 (1994), pp. 67–83.
[18] G. Heinig, *Displacement structure of pseudoinverses*, Linear Algebra Appl., 197/198 (1994), pp. 623–649.
[19] G. Heinig, *Generalized inverses of Hankel and Toeplitz mosaic matrices*, Linear Algebra Appl., 216 (1995), pp. 43–59.
[20] G. Heinig, *Inversion of generalized Cauchy matrices and other classes of structured matrices*, in Linear Algebra in Signal Processing, IMA Vol. Math. Appl. 69, Springer, New York, 1995, pp. 63–81.
[21] G. Heinig and A. Bojanczyk, *Transformation techniques for Toeplitz and Toeplitz-plus-Hankel matrices.* I. *Transformations*, Linear Algebra Appl., 254 (1997), pp. 193–226.
[22] G. Heinig and A. Bojanczyk, *Transformation techniques for Toeplitz and Toeplitz-plus-Hankel matrices.* II. *Algorithms*, Linear Algebra Appl., 278 (1998), pp. 11–36.
[23] G. Heinig and K. Rost, *Algebraic Methods for Toeplitz-Like Matrices and Operators*, Oper. Theory Adv. Appl. 13, Birkhäuser, Basel, Boston, MA, 1984.
[24] M. Kac, W. L. Murdock, and G. Szegő, *On the eigenvalues of certain Hermitian forms*, J. Rational Mech. Anal., 2 (1953), pp. 767–800.

[25] T. Kailath and J. Chun, *Generalized displacement structure for block-Toeplitz, Toeplitz-block, and Toeplitz-derived matrices*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 114–128.
[26] T. Kailath, S. Y. Kung, and M. Morf, *Displacement ranks of matrices and linear equations*, J. Math. Anal. Appl., 68 (1979), pp. 395–407.
[27] T. Kailath and V. Olshevsky, *Diagonal pivoting for partially reconstructible Cauchy-like matrices, with applications to Toeplitz-like linear equations and to boundary rational matrix interpolation problems*, Linear Algebra Appl., 254 (1997), pp. 251–302.
[28] T. Kailath and A. H. Sayed, *Displacement structure: Theory and application*, SIAM Rev., 37 (1995), pp. 297–386.
[29] T. Kailath and A. H. Sayed, eds., *Fast Reliable Algorithms for Matrices with Structure*, SIAM, Philadelphia, PA, 1999.
[30] D. E. Knuth, *The Art of Computer Programming, Vol.* 2: *Seminumerical Algorithms*, 3rd ed., Addison-Wesley, Reading, MA, 1997.
[31] S. MacLane and G. Birkhoff, *Algebra*, 3rd ed., Chelsea, New York, 1988.
[32] *MATLAB ver.* 5.3, The MathWorks, Inc., Natick, MA, 1999.
[33] *MATLAB ver.* 7.1, The MathWorks, Inc., Natick, MA, 2005.
[34] V. Y. Pan, *Structured Matrices and Polynomials: Unified Superfast Algorithms*, Birkhäuser Boston, Boston, MA, 2001.
[35] G. Rodriguez and D. Theis, *Least squares solution of large Toeplitz linear systems*, in Atti del VI Congresso SIMAI, Società Italiana di Matematica Applicata e Industriale, CD-ROM, 2002.
[36] D. R. Sweet, *Fast Toeplitz orthogonalization*, Numer. Math., 43 (1984), pp. 1–21.
[37] A. N. Tikhonov, *Solution of incorrectly formulated problems and the regularization method*, Soviet Math. Dokl., 4 (1963), pp. 1036–1038.
[38] L. N. Trefethen and D. Bau, III, *Numerical Linear Algebra*, SIAM, Philadelphia, 1997.
[39] W. F. Trench, *Numerical solution of the eigenvalue problem for Hermitian Toeplitz matrices*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 135–146.
[40] M. Van Barel, G. Heinig, and P. Kravanja, *A superfast method for solving Toeplitz linear least squares problems*, Linear Algebra Appl., 366 (2003), pp. 441–457.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 3, pp. 749–769

A SYMMETRY PRESERVING SINGULAR VALUE DECOMPOSITION*

MILI I. SHAH† AND DANNY C. SORENSEN†

Abstract. A reduced order representation of a large data set is often realized through a principal component analysis based upon a singular value decomposition (SVD) of the data. The left singular vectors of a truncated SVD provide the reduced basis. In several applications such as facial analysis and protein dynamics, structural symmetry is inherent in the data. Typically, reflective or rotational symmetry is expected to be present in these applications. In protein dynamics, determining this symmetry allows one to provide SVD major modes of motion that best describe the symmetric movements of the protein. In face detection, symmetry in the SVD allows for more efficient compression algorithms. Here we present a method to compute the plane of reflective symmetry or the axis of rotational symmetry of a large set of points. Moreover, we develop a symmetry preserving singular value decomposition (SPSVD) that best approximates the given set while respecting the symmetry. Interesting subproblems arise in the presence of noisy data or in situations where most, but not all, of the structure is symmetric. An important part of the determination of the axis of rotational symmetry or the plane of reflective symmetry is an iterative reweighting scheme. This scheme is rapidly convergent in practice and seems to be very effective in ignoring outliers (points that do not respect the symmetry).

Key words. singular value decomposition, symmetry constraints, large scale, principal components, protein dynamics

AMS subject classifications. 15A18, 65F15

DOI. 10.1137/050646676

1. Introduction.

Determining symmetry within a collection of spatially oriented points is a problem that occurs in many fields including molecular biology and face recognition analysis. In these applications, large amounts of data are generally collected, and it is desirable to approximate this data with a compressed representation. In some applications, the data is known to obey certain symmetry conditions, and it is profitable to preserve such symmetry in the compressed approximation. Taking advantage of symmetry leads to better modeling of physical processes as well as more efficient storage and computational schemes.

For a given set of points $\mathcal{S} = \{\mathbf{x}_i : 1 \le i \le m\}$ in n-dimensional space, we form an $n \times m$ matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m]$. The truncated singular value decomposition (SVD) provides a low rank approximation to $\mathbf{X}$ and therefore also to the data set $\mathcal{S}$. If $\mathbf{USV}^T = \mathbf{X}$ is an SVD of $\mathbf{X}$, then it is well known that the best rank r approximation to $\mathbf{X}$ (in both the 2-norm and the Frobenius norm) is given by $\mathbf{X}_r = \mathbf{U}_r\mathbf{S}_r\mathbf{V}_r^T$, where $\mathbf{U}_r, \mathbf{V}_r$ represent the dominant r columns of $\mathbf{U}, \mathbf{V}$ and $\mathbf{S}_r$ represents the dominant $r \times r$ principal submatrix of $\mathbf{S}$. Here we are concerned with preserving symmetry relations present in the set $\mathcal{S}$ and hence in the matrix $\mathbf{X}$. In particular, we desire the best low rank approximation $\mathbf{X}_r$ that also exhibits the same symmetries as the matrix $\mathbf{X}$. This is accomplished by providing a symmetry preserving singular value decomposition (SPSVD).

*Received by the editors December 3, 2005; accepted for publication (in revised form) by I. C. F. Ipsen April 13, 2006; published electronically September 26, 2006. This work was supported in part by NSF grant CCR-0306503 and by NSF grant ACI-0325081.
http://www.siam.org/journals/simax/28-3/64667.html
†Department of Computational and Applied Mathematics, Rice University, 6100 Main St., MS 134, Houston, TX 77005-1892 (mili@rice.edu, sorensen@rice.edu).

We concentrate on determining two types of symmetry: rotational and reflective. The computational schemes for calculating the best symmetric approximation of a given set involve two steps for each case. For reflective symmetry, the first step is to obtain the normal to an approximate plane of reflective symmetry, where the normal is defined to be the unit vector perpendicular to a hyperplane for which the given set can be split into two mirror image sets. For rotational symmetry, we first determine an approximate axis of rotational symmetry about which the given set can be rotated ($2\pi/k$ degrees in three dimensions) and returned to the same set. Then, in the second step, we find the best approximation to the given set that has the appropriate symmetries with respect to the approximate plane of symmetry or axis of rotation with the aid of the SPSVD.

For practical applications, we must consider noisy data sets. Thus, we need to construct a normal vector or axis of rotation that diminishes the effects of outliers. This is accomplished by creating an iterative reweighting scheme that minimizes deviation from symmetry in a weighted Frobenius norm. With our weighted normal or axis of rotation, we build our SPSVD that preserves the respective symmetries as in the nonweighted scheme.

We also provide a means to compute just the dominant portion (leading r terms) of the SPSVD that is well suited to large scale computation. This computation requires only matrix-vector products involving the point set represented as a matrix. The ARPACK software [8] can be used in this large scale case. The computation is no more expensive than constructing the leading terms of the SVD of the full set of points without the symmetry constraint. Computational examples involving the backbone of the HIV-1 protease molecule are presented here. These examples provide trajectories that result in matrices of dimension 9000 by 10000. The computations were performed on a multiprocessor cluster using the parallel P_ARPACK version of ARPACK.

There has been considerable research in the area of symmetry detection. Atallah [1] constructs an $O(n \log n)$ algorithm that determines the line of reflective symmetry of a perfectly symmetric planar object by reducing the system to a permutation problem. Optimizing a coefficient of symmetry is employed by Marola to determine an axis of symmetry for planar images [9]. Zabrodsky, Peleg, and Avnir [19] employ a continuous symmetry measure and apply it to finding reflective and rotational symmetries in chemistry. Kazhdan extends this idea to three-dimensional (3D) objects by creating a continuous two-dimensional (2D) function that measures the invariance of an object with respect to reflective symmetry about each plane that goes through the object's center of mass [4].

Many papers use the following fundamental properties of symmetry, which can be found in [17, 10, 11], to determine reflective and rotational symmetry. In this literature, the term "principal axes" refers to the eigenvectors of the correlation matrix $\mathbf{X}\mathbf{X}^T$ of the set of points, i.e., the left singular vectors of $\mathbf{X}$. The observations are the following:

- Any plane of symmetry of a body is perpendicular to a principal axis.
- Any axis of rotational symmetry of a body is a principal axis.

Minovic, Ishikawa, and Kato start with this idea and build an octree representation to find symmetries of a 3D object [12]. Sun and Sherrah [16] begin by looking at the extended Gaussian image of an object and then search along the principal axes for the strongest symmetry measure. O'Mara and Owens [14] also search for the principal axis with the largest symmetry measure. However, their symmetry measure is more

refined, since it takes into effect intensity values. Colliot et al. [3] extend O'Mara and Owens' research by starting with the highest symmetry measure principal axis. Then they optimize the axis of symmetry using the Nelder–Mead downhill simplex method. They apply this method to facial recognition and brain scan applications.

The idea of a symmetric approximation to a set of data points has come up in partial differential equations and in face detection. Aubry, Lian, and Titi prove that any truncated approximation to a dynamical system must maintain its respective symmetries. They derive a method of truncation, based on proper orthogonal decomposition, that obeys the symmetries of the original infinite-dimensional system [2]. Smaoui and Armbruster present a way to symmetrize the eigenmodes of the Karhunen–Loeve basis in a computationally efficient matter [15]. Kirby and Sirovich [6, 5] present a symmetric approximation based on taking the average of the even and odd (correctly oriented) symmetric faces. We prove here that taking the average gives the best symmetric approximation (in the Frobenius norm) to the original data set, and we generalize this result to give the best symmetric approximation to a set that possesses k-fold rotational symmetry.

The folding method is employed by Zabrodsky, Peleg, and Avnir [20] to calculate the best symmetric approximation to a set. This method produces an approximation that is equivalent to ours. However, our proof indicates how to calculate an SPSVD that gives the best low rank symmetric approximation to a set efficiently for large scale matrices.

In this paper, we have assumed a correct pairing of symmetric points. In many applications, such as molecular dynamics, this is a valid assumption. However, when this is not true, there are methods to create a pairing of points that has the desired symmetry properties. These methods make certain assumptions about the data set. For example, in [1] Atallah assumes a perfectly symmetric 2D set and employs the idea that reflectively symmetric points must be the same distance from the center of the data. Zabrodsky, Peleg, and Avnir [20] make the assumption that the set of rotationally symmetric points is ordered along a contour.

This paper is organized as follows. Section 2 defines perfect reflective and rotational symmetry. Finding an optimal hyperplane of reflective symmetry for noisy data is developed and analyzed in section 3, while choosing the axes of rotational symmetry for noisy data is discussed in section 4. Finally, section 5 develops an SPSVD that best approximates the given data set and provides an algorithm for directly computing the best low rank symmetry preserving approximation in a way that is suitable for large scale computation. Computational results are presented in section 6.

Throughout the discussion, $\|\cdot\|$ shall denote the 2-norm and $\|\cdot\|_F$ shall represent the Frobenius norm. The term [illegible] will refer to the algebraically smallest eigenvalue of a symmetric matrix. All vectors are column vectors.

2. Perfect symmetry. In this section, we lay out the basic defining properties of reflective and rotational symmetry. We also give analytic specifications of the normal to a plane of reflection and the axis of rotational symmetry when the given data set possesses exact symmetry relations.

2.1. Reflective symmetry. Recall that a hyperplane $\mathcal{H}$ is specified by a constant γ and a vector $\mathbf{w}$ via $\mathcal{H} := \{\mathbf{x} : \gamma + \mathbf{w}^T\mathbf{x} = 0\}$. The vector $\mathbf{w}$ is called the normal to the plane. We say that a set of points $\mathcal{S} \subset \mathbb{R}^n$ is [illegible] $\mathcal{H}$ if for every point $\mathbf{s} \in \mathcal{S}$ there exists a point $\hat{\mathbf{s}} \in \mathcal{S}$ such that $\hat{\mathbf{s}} = \mathbf{s} + \tau\mathbf{w}$ for some scalar τ with $\mathbf{s} + \frac{\tau}{2}\mathbf{w} \in \mathcal{H}$. It is easily shown that the center $\mathbf{c} \equiv \frac{1}{m}\sum_{\mathbf{s}\in\mathcal{S}} \mathbf{s}$ of the point set lies in the plane of symmetry, where m is the number

of elements in $\mathcal{S}$. A simple rigid translation of the point set will allow us to assume that the center is at the origin $\mathbf{c} = 0$ and hence also that $\gamma = 0$. These assumptions will be made throughout this discussion. For simplicity, we shall also assume that no points of $\mathcal{S}$ lie in the plane of symmetry.

The following lemma is an immediate consequence of the fact that for each $\mathbf{s} \in \mathcal{S}$ there is a reflected point $\hat{\mathbf{s}} = \mathbf{s} + \tau\mathbf{w} \in \mathcal{S}$.

LEMMA 2.1. $\mathcal{S}$ $\mathcal{H}$ $\mathbf{w}$

$$\mathcal{S} = (\mathbf{I} - 2\mathbf{w}\mathbf{w}^T)\mathcal{S}.$$

LEMMA 2.2. $\mathcal{S}$ $\mathcal{H}$ $\mathbf{c} \in \mathcal{H}$

If $\mathcal{S}$ is reflectively symmetric about $\mathcal{H}$, we can arrange the points of $\mathcal{S}$ into two sets represented as two $(n \times \frac{m}{2})$-dimensional matrices $\mathbf{X}_0$ and $\mathbf{X}_1$ such that

$$\mathbf{X}_0 = (\mathbf{I} - 2\mathbf{w}\mathbf{w}^T)\mathbf{X}_1.$$

Moreover, there is no loss of generality in assuming that $\mathbf{w}^T\mathbf{X}_0 > 0$ and that $\mathbf{w}^T\mathbf{X}_1 < 0$ (elementwise).

2.2. Rotational symmetry. We say that a set of points $\mathcal{S} \subset \mathbb{R}^n \bigcap \{\mathbf{z}^T\mathbf{q} = 0 : \mathbf{z} \in \mathbb{R}^n\}$ is k $\mathbf{q} \in \mathbb{R}^n$ if there exists an $n \times n$ orthogonal matrix $\mathbf{R}(\mathbf{q})$ such that for every point $\mathbf{s} \in \mathcal{S}$ there are exactly $k-1$ distinct points $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{k-1} \in \mathcal{S}$ with $\mathbf{R}(\mathbf{q})^i\mathbf{s} = \mathbf{s}_i$ for $i = 1, 2, \ldots, k-1$. We call $\mathbf{q}$ the rotational axis of symmetry and $\mathbf{R}(\mathbf{q})$ the rotation matrix. Lemma 2.3 gives an expression for the rotation matrix $\mathbf{R}(\mathbf{q})$.

LEMMA 2.3. $\mathcal{S}$ k $\mathbf{q}$ $i = 1, 2, \ldots, k-1$

$$\mathcal{S} = \mathbf{R}(\mathbf{q})^i\mathcal{S} = (\mathbf{I} - \mathbf{Q}\mathbf{G}\mathbf{Q}^T)^i\mathcal{S},$$

$\mathbf{Q} \in \mathbb{R}^{n\times(n-1)}$ $[\mathbf{q}, \mathbf{Q}]$ $\mathbf{I} - \mathbf{G} \in \mathbb{R}^{(n-1)\times(n-1)}$ $(\mathbf{I} - \mathbf{G})^k = \mathbf{I}$

Note that $(\mathbf{R}(\mathbf{q}))^k = (\mathbf{I} - \mathbf{Q}\mathbf{G}\mathbf{Q}^T)^k = \mathbf{I}$, and for $n = 3$, the matrix $\mathbf{I}_2 - \mathbf{G}$ is a 2×2 plane rotation through an angle of $\theta = 2\pi/k$ degrees.

If $\mathcal{S}$ is k-fold rotationally symmetric about $\mathbf{q}$, we can arrange the points of $\mathcal{S}$ into k sets represented as matrices $\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_{k-1}$ such that

$$\mathbf{X}_i = (\mathbf{I} - \mathbf{Q}\mathbf{G}\mathbf{Q}^T)^i\mathbf{X}_0$$

for $i = 1, 2, \ldots, k-1$. Again, we will assume that the center $\mathbf{c}$ of the data is at the origin. This can always be attained in general by a simple rigid translation of all the points of $\mathcal{S}$.

3. Optimal value of reflective w. Generally, in practice, the given set $\mathcal{S}$ is not exactly symmetric with respect to any particular plane. However, we may think of calculating a $\mathbf{w}$ that does the best possible job of specifying a plane that separates $\mathcal{S}$ into two sets $\mathbf{X}_0$ and $\mathbf{X}_1$ (again represented as matrices) that are "nearly" symmetric with respect to the plane.

It is possible to find an initial separation of $\mathcal{S}$ into $\mathbf{X}_0$ and $\mathbf{X}_1$ that are paired to be nearly symmetric with respect to a plane determined by a calculated $\mathbf{w}$. Methods for this are discussed in [1]. However, for this discussion, we shall assume that a

partitioning of $\mathcal{S}$ into $\mathbf{X}_0$ and $\mathbf{X}_1$ is given such that the columns of the two matrices are correctly paired.

The specification of $\mathbf{w}$ may be expressed as an optimization problem

$$\min_{\|\mathbf{w}\|=1} \{\|\mathbf{X}_0 - \mathbf{W}\mathbf{X}_1\|_F : \mathbf{W} = \mathbf{I} - 2\mathbf{w}\mathbf{w}^T\}. \tag{1}$$

LEMMA 3.1. [illegible] $\mathbf{w}$ [illegible] (1) [illegible]

$$\mathbf{M} = \mathbf{X}_0\mathbf{X}_1^T + \mathbf{X}_1\mathbf{X}_0^T.$$

[illegible]

$$\begin{aligned}
\|\mathbf{X}_0 - \mathbf{W}\mathbf{X}_1\|_F^2 &= \operatorname{tr}\{(\mathbf{X}_0 - \mathbf{X}_1)(\mathbf{X}_0 - \mathbf{X}_1)^T\} + 4\operatorname{tr}\{\mathbf{w}\mathbf{w}^T\mathbf{X}_1(\mathbf{X}_0 - \mathbf{X}_1)^T\} \\
&\quad + 4\operatorname{tr}\{(\mathbf{w}\mathbf{w}^T\mathbf{X}_1)(\mathbf{w}\mathbf{w}^T\mathbf{X}_1)^T\} \\
&= \operatorname{tr}\{(\mathbf{X}_0 - \mathbf{X}_1)(\mathbf{X}_0 - \mathbf{X}_1)^T\} + 4\mathbf{w}^T\mathbf{X}_1(\mathbf{X}_0 - \mathbf{X}_1)^T\mathbf{w} \\
&\quad + 4\mathbf{w}^T(\mathbf{X}_1\mathbf{X}_1^T)\mathbf{w} \\
&= \operatorname{tr}\{(\mathbf{X}_0 - \mathbf{X}_1)(\mathbf{X}_0 - \mathbf{X}_1)^T\} + 4\mathbf{w}^T(\mathbf{X}_1\mathbf{X}_0^T)\mathbf{w} \\
&= \operatorname{tr}\{(\mathbf{X}_0 - \mathbf{X}_1)(\mathbf{X}_0 - \mathbf{X}_1)^T\} + 2\mathbf{w}^T(\mathbf{X}_1\mathbf{X}_0^T + \mathbf{X}_0\mathbf{X}_1^T)\mathbf{w},
\end{aligned}$$

where we have used $\mathbf{w}^T\mathbf{w} = 1$ and that $\operatorname{tr}\{\mathbf{AB}\} = \operatorname{tr}\{\mathbf{BA}\}$.

Clearly, this quantity is minimized when $2\mathbf{w}^T(\mathbf{X}_1\mathbf{X}_0^T + \mathbf{X}_0\mathbf{X}_1^T)\mathbf{w}$ is minimized, and this occurs precisely when $\mathbf{w}$ is the (unit norm) eigenvector corresponding to the smallest eigenvalue of the symmetric matrix

$$\mathbf{M} = \mathbf{X}_1\mathbf{X}_0^T + \mathbf{X}_0\mathbf{X}_1^T. \qquad \square$$

A weighting can be introduced into the minimization problem (1) which gives a way to deemphasize anomalies in the supposed symmetry relation. In this case, we must solve

$$\min_{\|\mathbf{w}\|=1} \{\|(\mathbf{X}_0 - \mathbf{W}\mathbf{X}_1)\mathbf{D}\|_F : \mathbf{W} = \mathbf{I} - 2\mathbf{w}\mathbf{w}^T\}, \tag{2}$$

where $\mathbf{D}$ is a diagonal weighting matrix.

LEMMA 3.2. [illegible] $\mathbf{w}$ [illegible] (2) [illegible]

$$\mathbf{M}_\mathbf{D} = \mathbf{X}_0\mathbf{D}^2\mathbf{X}_1^T + \mathbf{X}_1\mathbf{D}^2\mathbf{X}_0^T. \tag{3}$$

[illegible]. The proof is similar to the proof of Lemma 3.1. $\square$

We have devised an iterative reweighting scheme to construct a $\mathbf{D}$ that diminishes the influence of outliers in the SPSVD. Given a guess $\mathbf{z}$ to the normal vector $\mathbf{w}$, the basic idea is to weight the ith column of $\mathbf{X}_0 - \mathbf{W}\mathbf{X}_1$, i.e., $\mathbf{x}_i^{(0)} - (\mathbf{I} - 2\mathbf{w}\mathbf{w}^T)\mathbf{x}_i^{(1)}$, by the reciprocal of the norm of $\mathbf{x}_i^{(0)} - (\mathbf{I} - 2\mathbf{z}\mathbf{z}^T)\mathbf{x}_i^{(1)}$, where $\mathbf{z}$ is a unit vector. The motivation is to penalize (give the smallest weight to) the pairs $\mathbf{x}_j^0, \mathbf{x}_j^1$ that are farthest from being symmetric with respect to $\mathbf{z}$.

Let us define

$$F(\mathbf{z}, \mathbf{w}) = \sum_{i=1}^{m} \left(\frac{f_i(\mathbf{w})}{f_i(\mathbf{z})} \right)^2 = \|(\mathbf{X}_0 - \mathbf{W}\mathbf{X}_1)\mathbf{D}(\mathbf{z})\|_F^2,$$

where $f_i(\mathbf{z}) = \|\mathbf{x}_i^{(0)} - (\mathbf{I} - 2\mathbf{z}\mathbf{z}^T)\mathbf{x}_i^{(1)}\|$ and $\mathbf{D}(\mathbf{z}) = \operatorname{diag}\left\{f_i(\mathbf{z})^{-1}\right\}$. To find the optimal normal with respect to this weighting, we choose $\mathbf{w}$ as the point that minimizes $\|(\mathbf{X}_0 - \mathbf{W}\mathbf{X}_1)\mathbf{D}(\mathbf{z})\|_F$, as described in Lemma 3.2. Then the approximate $\mathbf{w}$ associated with this weighting solves

$$\min_{\|\mathbf{w}\|=1} F(\mathbf{z}, \mathbf{w}). \tag{4}$$

This suggests an iterative reweighting scheme that will adjust the vector $\mathbf{z}$ to optimally diminish the effect of outliers; begin with an initial guess $\mathbf{z}_0$ and iterate

$$\mathbf{z}_{p+1} = \arg\min_{\|\mathbf{w}\|=1} F(\mathbf{z}_p, \mathbf{w}), \quad k = 0, 1, 2, \ldots, \tag{5}$$

until $\|\mathbf{z}_{p+1} - \mathbf{z}_p\|$ is sufficiently small. Upon convergence, this fixed point iteration will solve the max-min problem

$$\max_{\|\mathbf{z}\|=1} \left\{ \min_{\|\mathbf{v}\|=1} F(\mathbf{z}, \mathbf{v}) \right\}, \tag{6}$$

as the following lemma indicates.

LEMMA 3.3. [illegible] $\mathbf{v} = \mathbf{z}$ [illegible] (4) [illegible] $\mathbf{z}$ [illegible] (6) [illegible] $F(\mathbf{z}, \mathbf{v}) = m$ [illegible]. Given $\mathbf{z}$, $\|\mathbf{z}\| = 1$,

$$\min_{\|\mathbf{v}\|=1} \sum_{i=1}^{m} \left(\frac{f_i(\mathbf{v})}{f_i(\mathbf{z})} \right)^2 \leq \sum_{i=1}^{m} \left(\frac{f_i(\mathbf{z})}{f_i(\mathbf{z})} \right)^2 = m.$$

Hence,

$$\max_{\|\mathbf{z}\|=1} \left\{ \min_{\|\mathbf{v}\|=1} F(\mathbf{z}, \mathbf{v}) \right\} \leq m.$$

If $\mathbf{v} = \mathbf{z}$, then $F(\mathbf{z}, \mathbf{v}) = F(\mathbf{z}, \mathbf{z}) = m$. Therefore, any fixed point of the minimization problem (4) is a solution to the max-min problem (6). □

We have shown in the above lemma that a fixed point of iteration (5) solves the max-min problem (6). Now we will show the existence of a fixed point to the iteration (5) in Theorem 3.4.

THEOREM 3.4. [illegible] $\mathbf{z}_*$ [illegible]

$$\mathbf{z}_* = \arg\min_{\|\mathbf{w}\|=1} F(\mathbf{z}_*, \mathbf{w}).$$

[illegible]. Let $\mathbf{M}_i = \|\mathbf{x}_i^{(0)} - \mathbf{x}_i^{(1)}\|^2\mathbf{I} + 2(\mathbf{x}_i^{(0)}\mathbf{x}_i^{(1)^T} + \mathbf{x}_i^{(1)}\mathbf{x}_i^{(0)^T})$. For a given $\mathbf{z}$, any $\mathbf{w}$ that solves

$$\min_{\|\mathbf{w}\|=1} F(\mathbf{z}, \mathbf{w}) = \min_{\|\mathbf{w}\|=1} \sum_{i=1}^{m} \frac{\mathbf{w}^T\mathbf{M}_i\mathbf{w}}{\mathbf{z}^T\mathbf{M}_i\mathbf{z}}$$

will also solve

$$\min_{\|\mathbf{w}\|=1} \Phi(\mathbf{z})F(\mathbf{z}, \mathbf{w}) = \min_{\|\mathbf{w}\|=1} \sum_{i=1}^{m} \phi_i(\mathbf{z})\mathbf{w}^T\mathbf{M}_i\mathbf{w},$$

where

$$\Phi(\mathbf{z}) = \prod_{i=1}^{m} \mathbf{z}^T \mathbf{M}_i \mathbf{z} \quad \text{and} \quad \phi_i(\mathbf{z}) = \prod_{\substack{j=1 \\ j \neq i}}^{m} \mathbf{z}^T \mathbf{M}_j \mathbf{z}.$$

The function $\Phi(\mathbf{z})$ restricted to the unit n-sphere is a continuous function on a compact set. Therefore, $\min_{\mathbf{z}} \Phi(\mathbf{z}) = \Phi(\mathbf{z}_*)$ is attained at some point $\mathbf{z} = \mathbf{z}_*$ on the unit sphere.

From Lagrange theory, we see that

$$\nabla \Phi(\mathbf{z}_*) = 2 \sum_{i=1}^{m} \phi_i(\mathbf{z}_*) \mathbf{M}_i \mathbf{z}_* = 2 \mathbf{z}_* \lambda,$$

or, if we denote $\mathbf{M}(\mathbf{z}) = \sum_{i=1}^{m} \phi_i(\mathbf{z}) \mathbf{M}_i$,

$$\mathbf{M}(\mathbf{z}_*) \mathbf{z}_* = \mathbf{z}_* \lambda.$$

Now it is straightforward to show that an eigenvector corresponding to the smallest eigenvalue of $\mathbf{M}(\mathbf{z}_*)$ is also an eigenvector corresponding to the smallest eigenvalue of $\mathbf{M}_{\mathbf{D}}$ in (3) with $\mathbf{D} = \mathbf{D}(\mathbf{z}_*)$. Therefore, it is sufficient to show that λ is the smallest eigenvalue of $\mathbf{M}(\mathbf{z}_*)$ to show that $\mathbf{z}_*$ is a fixed point. The following argument will establish this.

Due to the Kurush–Kuhn–Tucker first and second order necessary conditions [13], for all $\mathbf{w}$ such that $\mathbf{w}^T \mathbf{z}_* = 0$, we must have

$$\mathbf{w}^T \nabla \Phi(\mathbf{z}_*) = \mathbf{w}^T \mathbf{M}(\mathbf{z}_*) \mathbf{z}_* = 0$$

and

$$\mathbf{w}^T \left(\nabla^2 \Phi(\mathbf{z}_*) - 2\lambda \mathbf{I} \right) \mathbf{w} \geq 0. \tag{7}$$

Now

$$\nabla^2 \Phi(\mathbf{z}) = 2 \sum_{i=1}^{m} \phi_i(\mathbf{z}) \mathbf{M}_i + 2 \sum_{i=1}^{m} \mathbf{M}_i \mathbf{z} \nabla \phi_i(\mathbf{z})^T$$

and

$$\begin{aligned} \nabla \phi_i(\mathbf{z}) &= \nabla \left(\prod_{\substack{j=1 \\ j \neq i}}^{m} \mathbf{z}^T \mathbf{M}_j \mathbf{z} \right) \\ &= \nabla \left(\frac{\Phi(\mathbf{z})}{\mathbf{z}^T \mathbf{M}_i \mathbf{z}} \right) \\ &= \frac{1}{\mathbf{z}^T \mathbf{M}_i \mathbf{z}} \nabla \Phi(\mathbf{z}) - \frac{2\Phi(\mathbf{z})}{(\mathbf{z}^T \mathbf{M}_i \mathbf{z})^2} \mathbf{M}_i \mathbf{z}. \end{aligned}$$

Therefore,

$$\mathbf{w}^T \nabla \phi_i(\mathbf{z}_*) = -\frac{2\Phi(\mathbf{z}_*)}{({\mathbf{z}_*}^T \mathbf{M}_i \mathbf{z}_*)^2} \mathbf{w}^T \mathbf{M}_i \mathbf{z}_*. \tag{8}$$

Substituting expression (8) into the formula for $\mathbf{w}^T \left(\nabla^2 \Phi(\mathbf{z}_*) - 2\lambda \mathbf{I}\right) \mathbf{w}$ in the second order necessary conditions (7) gives

$$\begin{aligned} 0 &\leq 2\mathbf{w}^T \mathbf{M}(\mathbf{z}_*)\mathbf{w} - 4 \left(\frac{\mathbf{w}^T \mathbf{M}_i \mathbf{z}_*}{\mathbf{z}_*{}^T \mathbf{M}_i \mathbf{z}_*} \right)^2 \Phi(\mathbf{z}_*) - 2\lambda \\ &\leq 2(\mu - \lambda), \end{aligned}$$

where $\mu = \mathbf{w}^T \mathbf{M}(\mathbf{z}_*)\mathbf{w}$. Thus, $\lambda \leq \mu$ for any eigenvalue μ of $\mathbf{M}(\mathbf{z}_*)$. Since λ is the smallest eigenvalue of $\mathbf{M}(\mathbf{z}_*)$, we have established that a constrained minimizer $\mathbf{z}_*$ of $\Phi(\mathbf{z})$ satisfies $\mathbf{z}_* = \arg\min_{\|\mathbf{w}\|=1} F(\mathbf{z}_*, \mathbf{w})$. □

. . . . . We have assumed in Theorem 3.4 that $\Phi(\mathbf{z}) \neq 0$. This is a reasonable assumption, since the only way $\Phi(\mathbf{z}) = 0$ is if $\|\mathbf{x}_j^{(0)}\| = \|\mathbf{x}_j^{(1)}\|$ for some pair $(\mathbf{x}_j^{(0)}, \mathbf{x}_j^{(1)})$. Since we are dealing with noisy sets, it is unlikely that these norms are precisely equal in practice. Nevertheless, we are considering equivalent reformulations that avoid this difficulty altogether.

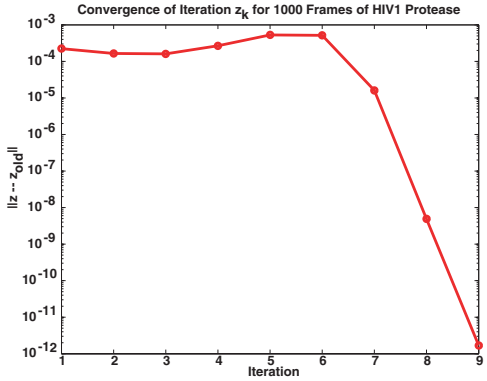


FIG. 1. *Convergence of* 1000 *frames of HIV-*1 *protease using iteration* (5).

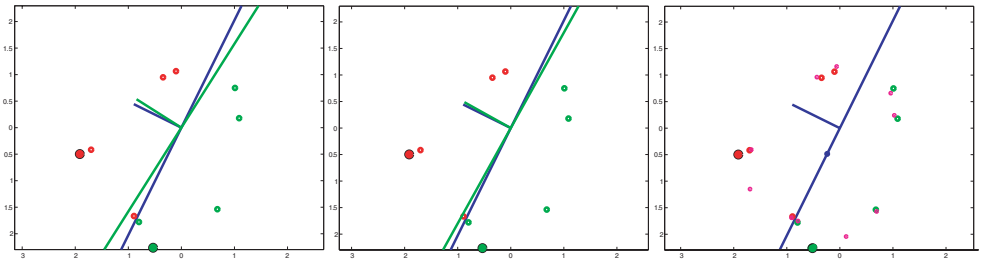


FIG. 2. *Iterations showing that our weighting is a good choice. Notice how as the iterations progress the normal converges to the correct solution, even in the presence of outliers (larger dots). The smaller dots in the last frame show our best symmetric approximation to the original data set.*

Convergence of the iterates $\mathbf{z}_p$ produced by (5) is yet to be proven. However, the convergence history shown in Figures 1 and 2 is typical, and iteration (5) seems to be convergent in practice. Theorem 3.5 does at least establish that the sequence of function values, $\Phi(\mathbf{z}_p)$, is monotonically decreasing and convergent.

THEOREM 3.5. . . . . $\Phi(\mathbf{z}_p)$ . . . $\mathbf{z}_p$ . . . . . (5) . . . . .

. . .

[illegible]. In the proof of Theorem 3.4, we show that a constrained minimizer $\mathbf{z}_*$ of

$$\Phi(\mathbf{z}) = \prod_{i=1}^{m} \mathbf{z}^T \mathbf{M}_i \mathbf{z} = \prod_{i=1}^{m} \|\mathbf{x}_i^{(0)} - (\mathbf{I} - 2\mathbf{z}\mathbf{z}^T)\mathbf{x}_i^{(1)}\|^2$$

is a fixed point to iteration (5). If we can show that $\Phi(\mathbf{z}_p)$, where $\mathbf{z}_p$ satisfies iteration (5), is a monotonically decreasing function, we will have proven that the sequence $\Phi(\mathbf{z}_p)$, with $\mathbf{z}_p$ produced by iteration (5), is convergent. Notice that

$$\frac{\Phi(\mathbf{z}_{p+1})}{\Phi(\mathbf{z}_p)} = \prod_{i=1}^{m} \frac{\|\mathbf{x}_i^{(0)} - (\mathbf{I} - 2\mathbf{z}_{p+1}\mathbf{z}_{p+1}^T)\mathbf{x}_i^{(1)}\|^2}{\|\mathbf{x}_i^{(0)} - (\mathbf{I} - 2\mathbf{z}_p\mathbf{z}_p^T)\mathbf{x}_i^{(1)}\|^2},$$

and $\mathbf{z}_{p+1}$ is chosen such that it minimizes the optimization problem (4); thus

$$\sum_{i=1}^{m} \frac{\|\mathbf{x}_i^{(0)} - (\mathbf{I} - 2\mathbf{z}_{p+1}\mathbf{z}_{p+1}^T)\mathbf{x}_i^{(1)}\|^2}{\|\mathbf{x}_i^{(0)} - (\mathbf{I} - 2\mathbf{z}_p\mathbf{z}_p^T)\mathbf{x}_i^{(1)}\|^2} \leq \sum_{i=1}^{m} \frac{\|\mathbf{x}_i^{(0)} - (\mathbf{I} - 2\mathbf{z}_p\mathbf{z}_p^T)\mathbf{x}_i^{(1)}\|^2}{\|\mathbf{x}_i^{(0)} - (\mathbf{I} - 2\mathbf{z}_p\mathbf{z}_p^T)\mathbf{x}_i^{(1)}\|^2} = m.$$

Since the geometric mean never exceeds the arithmetic mean,

$$\left[\prod_{i=1}^{m} \frac{\|\mathbf{x}_i^{(0)} - (\mathbf{I} - 2\mathbf{z}_{p+1}\mathbf{z}_{p+1}^T)\mathbf{x}_i^{(1)}\|^2}{\|\mathbf{x}_i^{(0)} - (\mathbf{I} - 2\mathbf{z}_p\mathbf{z}_p^T)\mathbf{x}_i^{(1)}\|^2}\right]^{(1/m)} \leq \frac{1}{m}\sum_{i=1}^{m} \frac{\|\mathbf{x}_i^{(0)} - (\mathbf{I} - 2\mathbf{z}_{p+1}\mathbf{z}_{p+1}^T)\mathbf{x}_i^{(1)}\|^2}{\|\mathbf{x}_i^{(0)} - (\mathbf{I} - 2\mathbf{z}_p\mathbf{z}_p^T)\mathbf{x}_i^{(1)}\|^2} \leq 1.$$

Thus,

$$\prod_{i=1}^{m} \frac{\|\mathbf{x}_i^{(0)} - (\mathbf{I} - 2\mathbf{z}_{p+1}\mathbf{z}_{p+1}^T)\mathbf{x}_i^{(1)}\|^2}{\|\mathbf{x}_i^{(0)} - (\mathbf{I} - 2\mathbf{z}_p\mathbf{z}_p^T)\mathbf{x}_i^{(1)}\|^2} \leq 1.$$

Hence, $\Phi(\mathbf{z}_p)$ is a monotonically decreasing sequence that is bounded below and is therefore convergent. □

We have compared the convergence of iteration (5) to a fixed point with the modified compass search method [7] on an equivalent optimization problem:

$$\min_{\|\mathbf{z}\|=1} \|\mathbf{z} - \mathbf{v}\|, \tag{9}$$

where, as before, $\mathbf{v}$ is the eigenvector associated with the smallest eigenvalue of (3) with $\mathbf{D} = \mathrm{diag}(f_i(\mathbf{z})^{-1})$. We have observed that, in general, iteration (5) converges faster and more efficiently when compared to the compass search method. Also, more accurate results are usually obtained with iteration (5).

4. Optimal value of rotational axis q. Recall that for a perfectly rotationally symmetric set,

$$\mathbf{X}_i = (\mathbf{I} - \mathbf{Q}\mathbf{G}\mathbf{Q}^T)^i \mathbf{X}_0, \tag{10}$$

where the columns of $[\mathbf{q}, \mathbf{Q}]$ form an orthogonal set. This specification suggests a means to compute the axis of rotation.

LEMMA 4.1. [illegible] $\mathbf{X}_0$ [illegible] n [illegible] $\mathbf{G}$ [illegible] $\mathbf{q}$ [illegible]

$$\mathbf{q}^T \left[(k-1)\mathbf{X}_0 - \sum_{i=1}^{k-1} \mathbf{X}_i\right] = 0. \tag{11}$$

. First, note that if $\mathbf{q}$ is an axis of rotational symmetry, then $\mathbf{q}^T\mathbf{Q} = 0$ must hold, and thus

$$\mathbf{q}^T\mathbf{X}_i = \mathbf{q}^T(\mathbf{I} - \mathbf{Q}\mathbf{G}\mathbf{Q}^T)^i\mathbf{X}_0 = \mathbf{q}^T\mathbf{X}_0 \quad \text{for} \quad i = 1, 2, \ldots, k,$$

which implies that (11) must hold.

From (10),

$$\begin{aligned}\mathbf{X}_i &= (\mathbf{I} - \mathbf{Q}\mathbf{G}\mathbf{Q}^T)^i\mathbf{X}_0 \\ &= (\mathbf{q}\mathbf{q}^T + \mathbf{Q}(\mathbf{I} - \mathbf{G})\mathbf{Q}^T)^i\mathbf{X}_0 \\ &= (\mathbf{q}\mathbf{q}^T + \mathbf{Q}(\mathbf{I} - \mathbf{G})^i\mathbf{Q}^T)\mathbf{X}_0.\end{aligned}$$

Thus,

$$\begin{aligned}\sum_{i=1}^{k-1}\mathbf{X}_i &= \left((k-1)\mathbf{q}\mathbf{q}^T + \mathbf{Q}\left(\sum_{i=1}^{k-1}(\mathbf{I} - \mathbf{G})^i\right)\mathbf{Q}^T\right)\mathbf{X}_0 \\ &= ((k-1)\mathbf{q}\mathbf{q}^T - \mathbf{Q}\mathbf{Q}^T)\mathbf{X}_0 = k\mathbf{q}\mathbf{q}^T\mathbf{X}_0 - \mathbf{X}_0,\end{aligned}$$

since $(\mathbf{I} - \mathbf{G})^k = \mathbf{I}$ implies that $\sum_{i=1}^{k-1}(\mathbf{I} - \mathbf{G})^i = -\mathbf{I}$ when $\mathbf{G}$ is nonsingular. From this, it follows that

$$(k-1)\mathbf{X}_0 - \sum_{i=1}^{k-1}\mathbf{X}_i = k(\mathbf{I} - \mathbf{q}\mathbf{q}^T)\mathbf{X}_0.$$

Now, suppose $\hat{\mathbf{q}}$ is any unit vector that satisfies (11) (in place of $\mathbf{q}$). Since $\mathbf{X}_0$ is full rank and $\hat{\mathbf{q}}$ satisfies (11),

$$0 = \hat{\mathbf{q}}^T\left[(k-1)\mathbf{X}_0 - \sum_{i=1}^{k-1}\mathbf{X}_i\right] = k\hat{\mathbf{q}}^T(\mathbf{I} - \mathbf{q}\mathbf{q}^T)\mathbf{X}_0$$

implies that $\hat{\mathbf{q}} = \mathbf{q}(\hat{\mathbf{q}}^T\mathbf{q})$. Since both $\mathbf{q}$ and $\hat{\mathbf{q}}$ are unit length, it follows from Cauchy–Schwarz that $\hat{\mathbf{q}} = \pm\mathbf{q}$. □

. In $\mathbb{R}^3$ the only way $\mathbf{G}$ can be singular is if it is identically $\mathbf{0}$, and since we are assuming many points, it is also not unreasonable to assume that $\mathbf{X}_0$ has full rank.

This gives a condition for calculating the axis of rotation, $\mathbf{q}$, when the data is exactly symmetric. However, in general, we are not given a perfectly symmetric data set $\mathcal{S}$. Therefore, we need to be able to specify an approximate rotational axis $\mathbf{q}$ that best fits the data. To this end, we shall assume a partitioning of $\mathcal{S}$ into $\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_{k-1}$ such that the columns of the matrices are correctly paired. Then we can formulate the optimization problem

$$\min_{\|\mathbf{q}\|=1}\left\{\left\|\mathbf{q}^T\left[(k-1)\mathbf{X}_0 - \sum_{i=1}^{k-1}\mathbf{X}_i\right]\right\|_F\right\} \tag{12}$$

to specify our approximate rotational axis of symmetry $\mathbf{q}$. Of course, we can characterize $\mathbf{q}$ as follows.

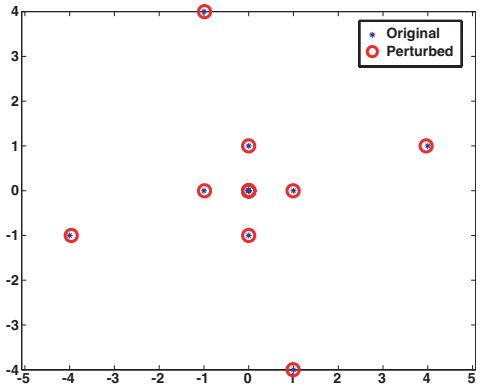


FIG. 3. *Comparison of the projection of the original and perturbed points onto the y-z plane.*

LEMMA 4.2. [illegible] $\mathbf{q}$ [illegible] (12) [illegible] $\mathbf{MM}^T$ [illegible]

$$\mathbf{M} = (k-1)\mathbf{X}_0 - \sum_{i=1}^{k-1} \mathbf{X}_i. \tag{13}$$

Note that this characterization provides a computational mechanism that is robust in the presence of noise. An alternate specification of $\mathbf{q}$ suggested by Minovic et al. is to consider the principal axis of the inertia matrix (correlation matrix) associated with the distinct eigenvalue for an initial guess to the rotational axes of symmetry. The motivation for this is that with exact symmetry the inertia matrix will have a distinct eigenvalue of multiplicity one and another eigenvalue of multiplicity $n-1$. However, in the presence of noise, these criteria may fail. For example, consider the following 4-fold perfectly rotationally symmetric data set with respect to $\mathbf{q} = [1, 0, 0]^T$:

$$\mathbf{X} = \begin{pmatrix} 1 & 4 & 0 & 1 & 4 & 0 & 1 & 4 & 0 & 1 & 4 & 0 \\ 0 & 1 & 4 & 0 & 0 & 1 & 0 & -1 & -4 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 & -4 & 0 & 0 & -1 & 0 & 1 & 4 \end{pmatrix}$$

with eigenvalues 34.667, 36, 36 (or singular values 5.888, 6, 6) after centering. In this case, we can clearly distinguish the distinct eigenvalue and get the corresponding correct axis. However, if we consider the SVD of $\mathbf{X} = \mathbf{USV}^T$, where $\mathbf{S} = \text{diag}\{\sigma_1, \sigma_2, \sigma_3\}$, and perturb the data by

$$\mathbf{X} + \mathbf{E} = \mathbf{USV}^T + \mathbf{US_E V}^T$$

with $\mathbf{S_E} = \text{diag}\{0, -(1+\epsilon)\tau, \tau\}$, where $\tau = (\sigma_2 - \sigma_3)/2 \approx (6 - 5.888)/2 = 0.056$ and $0 \leq \epsilon \ll 1$, then the Minovic condition fails. To see this point, let $\epsilon = 0.001$. Then the residual norm between the original and approximated data is approximately 0.007, which is well within the realm of experimental error in an application. Also, the data points remain symmetric (see Figure 3), and the eigenvalues of the approximated system become 35.330, 35.330, 36 (or singular values 5.944, 5.944, 6). However, the eigenvector associated with the distinct eigenvalue (here 36) corresponds to the vector $[0, 0, 1]^T$. In contrast, our method clearly identifies the correct axis of symmetry.

As with reflective symmetry, we can introduce a weighting scheme that minimizes the influence of outliers in the supposed rotational symmetry relation:

$$\min_{\|\mathbf{q}\|=1} \left\{ \left\| \mathbf{q}^T \left[(k-1)\mathbf{X}_0 - \sum_{i=1}^{k-1} \mathbf{X}_i \right] \mathbf{D} \right\|_F \right\}, \tag{14}$$

where $\mathbf{D}$ is a diagonal weighting matrix. If such a weighting has been specified, then we have the following lemma.

LEMMA 4.3. [illegible] (14) [illegible] $\mathbf{q}$ [illegible] $\mathbf{MD}^2\mathbf{M}^T$ [illegible] $\mathbf{M}$ [illegible] (13)

As in reflective symmetry, we have developed an iterative reweighting scheme to specify the weighting matrix $\mathbf{D}$ of the minimization problem (14) that effectively diminishes the influence of outliers in the final SPSVD approximation. Given a guess $\mathbf{z}$ of unit length, the ith column of $\mathbf{M}$ is weighted by $g_i(\mathbf{z})^{-1}$, where $g_i(\mathbf{z}) = \left\| \mathbf{z}^T \left[(k-1)\mathbf{x}_i^{(0)} - \sum_{j=1}^{k} \mathbf{x}_i^{(j)} \right] \right\|$. If we define

$$G(\mathbf{z}, \mathbf{q}) = \sum_{i=1}^{m} \left(\frac{g_i(\mathbf{q})}{g_i(\mathbf{z})} \right)^2 = \left\| \mathbf{q}^T \left[(k-1)\mathbf{X}_0 - \sum_{i=1}^{k-1} \mathbf{X}_i \right] \mathbf{D}(\mathbf{z}) \right\|_F^2,$$

then the approximate $\mathbf{q}$ associated with this weighting solves

$$\min_{\|\mathbf{q}\|=1} G(\mathbf{z}, \mathbf{q}). \tag{15}$$

The motivation for this is to put greater weight on points that are more symmetric with respect to $\mathbf{z}$ than points that are not. Then $\mathbf{q}$ is constructed to have the optimal normal with respect to the weighting as described in Lemma 4.3. If $\mathbf{q}$ is not acceptable, then $\mathbf{z} \leftarrow \mathbf{q}$, and the process is repeated until an acceptable $\mathbf{q}$ is found. This suggests an iterative reweighting. Given an initial guess $\mathbf{z}_0$ to the axis of rotation, we iterate

$$\mathbf{z}_{p+1} = \arg \min_{\|\mathbf{q}\|=1} G(\mathbf{z}_p, \mathbf{q}) \tag{16}$$

until $\|\mathbf{z}_{p+1} - \mathbf{z}_p\|$ is under a predetermined tolerance. A fixed point of iteration (16) is the solution to the max-min problem

$$\max_{\|\mathbf{z}\|=1} \left\{ \min_{\|\mathbf{q}\|=1} G(\mathbf{z}, \mathbf{q}) \right\}, \tag{17}$$

as the next lemma suggests.

LEMMA 4.4. [illegible] $\mathbf{q} = \mathbf{z}$ [illegible] (16) [illegible] $\mathbf{q}$ [illegible] (17) [illegible] $G(\mathbf{z}, \mathbf{q}) = m$ [illegible]. The proof is essentially the same as the proof of Lemma 3.3. □

Moreover, we have the following theorem.

THEOREM 4.5. [illegible] (16) [illegible]. The proof is essentially the same as the proof of Theorem 3.4. □

We have also compared iteration (16) with the modified compass search method on the equivalent optimization problem

$$\min_{\|\mathbf{z}\|=1} \|\mathbf{z} - \mathbf{q}\|, \tag{18}$$

where $\mathbf{q}$ is the eigenvector associated with the smallest eigenvalue of $\mathbf{M}\mathbf{D}^2\mathbf{M}^T$ with $\mathbf{D} = \mathrm{diag}(g_i(\mathbf{z})^{-1})$. We have observed that iteration (16) is generally more efficient and produces more accurate fixed point solutions when compared to the compass search method.

5. Best symmetric approximation to a set. To find the best reflective or rotational symmetric approximation to a set, we can take advantage of the following theorem. For reflective symmetry $\mathbf{R} = \mathbf{W}$ and $\mathbf{W}^2 = \mathbf{I}$, and in the case of rotational symmetry $\mathbf{R} = \mathbf{R}(\mathbf{q})$ and $\mathbf{R}(\mathbf{q})^k = \mathbf{I}$.

THEOREM 5.1.

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_0 \\ \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_{k-1} \end{pmatrix},$$

$$\mathbf{R}^{k-i}\mathbf{X}_i = \mathbf{X}_0 + \mathbf{E}_i,$$

$\mathbf{R}^k = \mathbf{I}$

$$\min_{\substack{\widehat{\mathbf{X}}_{i+1}=\mathbf{R}\widehat{\mathbf{X}}_i,\\ i=0,1,\ldots,k-2}} \left\| \begin{pmatrix} \mathbf{X}_0 \\ \vdots \\ \mathbf{X}_{k-1} \end{pmatrix} - \begin{pmatrix} \widehat{\mathbf{X}}_0 \\ \vdots \\ \widehat{\mathbf{X}}_{k-1} \end{pmatrix} \right\|_F^2 = \frac{1}{k}\sum_{i=0}^{k-1}\sum_{j=i+1}^{k-1} \|\mathbf{E}_j - \mathbf{R}^{j-i}\mathbf{E}_i\|_F^2$$

$$\mathbf{U}\mathbf{S}\mathbf{V}^T = \begin{pmatrix} \widehat{\mathbf{X}}_0 \\ \vdots \\ \widehat{\mathbf{X}}_{k-1} \end{pmatrix}$$

$$\mathbf{U} = \frac{1}{\sqrt{k}} \begin{pmatrix} \mathbf{U}_0 \\ \vdots \\ \mathbf{U}_{k-1} \end{pmatrix}, \quad \mathbf{S} = \sqrt{k}\mathbf{S}_0, \quad \mathbf{V} = \mathbf{V}_0,$$

$$\mathbf{U}_i = \mathbf{R}^i\mathbf{U}_0 \quad \text{for} \quad i = 0, 1, 2, \ldots, k-1,$$

$$\mathbf{U}_0\mathbf{S}_0\mathbf{V}_0^T = \frac{1}{k}(\mathbf{X}_0 + \mathbf{R}^{k-1}\mathbf{X}_1 + \mathbf{R}^{k-2}\mathbf{X}_2 + \cdots + \mathbf{R}\mathbf{X}_{k-1}).$$

. The proof will consist of a sequence of straightforward lemmas. We begin by assuming that we have perfect symmetry.

LEMMA 5.2. $\mathbf{E}_j = 0$ $j = 0, 1, 2, \ldots, k-1$

$$\begin{pmatrix} \mathbf{X}_0 \\ \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_{k-1} \end{pmatrix} = \begin{pmatrix} \mathbf{U}_0 \\ \mathbf{U}_1 \\ \vdots \\ \mathbf{U}_{k-1} \end{pmatrix} \mathbf{S}\mathbf{V}^T \tag{19}$$

$\mathbf{X}$

$$\mathbf{U}_i = \mathbf{R}^i \mathbf{U}_0,$$

$i = 0, 1, \ldots, k-1$

. From (19), we have

$$\mathbf{U}_i = \mathbf{X}_i \mathbf{V}\mathbf{S}^{-1},$$

where $\mathbf{U}_0^T\mathbf{U}_0 + \mathbf{U}_1^T\mathbf{U}_1 + \cdots + \mathbf{U}_{k-1}^T\mathbf{U}_{k-1} = \mathbf{I}$. Thus,

$$\mathbf{U}_i = \mathbf{X}_i \mathbf{V}\mathbf{S}^{-1} = \mathbf{R}^i \mathbf{X}_0 \mathbf{V}\mathbf{S}^{-1} = \mathbf{R}^i \mathbf{U}_0. \qquad \square$$

Therefore, when $\mathbf{R}$ is known, the SVD of a perfectly symmetric set may be efficiently computed by just taking the SVD of $\mathbf{X}_0$ and putting $\mathbf{U}_i = \mathbf{R}\mathbf{U}_{i-1}$, $1 \leq i \leq k-1$. Combining this fact with the following lemma leads to an algorithm for calculating the best low rank approximation to a matrix that preserves symmetry.

LEMMA 5.3. $\mathbf{X}_0 = \mathbf{U}_0\mathbf{S}_0\mathbf{V}_0^T$ $\mathbf{X}_0$ $\mathbf{U}_0^T\mathbf{U}_0 = \mathbf{V}_0^T\mathbf{V}_0 = \mathbf{I}$

$$\begin{pmatrix} \mathbf{X}_0 \\ \vdots \\ \mathbf{X}_0 \end{pmatrix} = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

$$\mathbf{U} = \frac{1}{\sqrt{k}} \begin{pmatrix} \mathbf{U}_0 \\ \vdots \\ \mathbf{U}_0 \end{pmatrix}, \quad \mathbf{S} = \sqrt{k}\mathbf{S}_0, \quad \mathbf{V} = \mathbf{V}_0.$$

. Clearly, $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, and

$$\begin{aligned} \begin{pmatrix} \mathbf{X}_0 \\ \vdots \\ \mathbf{X}_0 \end{pmatrix} = \begin{pmatrix} \mathbf{U}_0 \\ \vdots \\ \mathbf{U}_0 \end{pmatrix} \mathbf{S}_0\mathbf{V}_0^T &= \frac{1}{\sqrt{k}} \begin{pmatrix} \mathbf{U}_0 \\ \vdots \\ \mathbf{U}_0 \end{pmatrix} \sqrt{k}\mathbf{S}_0\mathbf{V}_0^T \\ &= \mathbf{U}\mathbf{S}\mathbf{V}^T, \end{aligned}$$

which is indeed the SVD. $\square$

We are now ready to give the best low rank approximation that preserves symmetry for a noisy data set.

LEMMA 5.4. $\widehat{\mathbf{Z}} = \frac{1}{k}(\mathbf{Z}_0 + \mathbf{Z}_1 + \cdots + \mathbf{Z}_{k-1})$ $\mathbf{Z} = \widehat{\mathbf{Z}}$

$$\min_{\mathbf{Z}} \left\| \begin{pmatrix} \mathbf{Z}_0 \\ \vdots \\ \mathbf{Z}_{k-1} \end{pmatrix} - \begin{pmatrix} \mathbf{Z} \\ \vdots \\ \mathbf{Z} \end{pmatrix} \right\|_F^2.$$

. Consider

$$\left\| \begin{pmatrix} \mathbf{Z}_0 \\ \vdots \\ \mathbf{Z}_{k-1} \end{pmatrix} - \begin{pmatrix} \mathbf{Z} \\ \vdots \\ \mathbf{Z} \end{pmatrix} \right\|_F^2 = \|\mathbf{Z}_0 - \mathbf{Z}\|_F^2 + \|\mathbf{Z}_1 - \mathbf{Z}\|_F^2 + \cdots + \|\mathbf{Z}_{k-1} - \mathbf{Z}\|_F^2,$$

and note that

$$\|\mathbf{Z}_i - \mathbf{Z}\|_F^2 = \operatorname{tr}(\mathbf{Z}_i^T \mathbf{Z}_i) - 2\operatorname{tr}(\mathbf{Z}_i^T \mathbf{Z}) + \operatorname{tr}(\mathbf{Z}^T \mathbf{Z})$$

for $i = 0, 1, 2, \ldots, k-1$. Therefore,

$$\left\| \begin{pmatrix} \mathbf{Z}_0 \\ \vdots \\ \mathbf{Z}_{k-1} \end{pmatrix} - \begin{pmatrix} \mathbf{Z} \\ \vdots \\ \mathbf{Z} \end{pmatrix} \right\|_F^2 = \operatorname{tr}\left(\sum_{i=0}^{k-1} \mathbf{Z}_i^T \mathbf{Z}_i \right) - 2\operatorname{tr}\left(\sum_{i=0}^{k-1} \mathbf{Z}_i^T \mathbf{Z} \right) + (k)\operatorname{tr}(\mathbf{Z}^T \mathbf{Z}).$$

However,

$$\begin{aligned} -2\operatorname{tr}\left(\sum_{i=0}^{k-1} \mathbf{Z}_i^T \mathbf{Z} \right) + (k)\operatorname{tr}(\mathbf{Z}^T \mathbf{Z}) &= -2\operatorname{tr}\left(\frac{1}{\sqrt{k}} \left(\sum_{i=0}^{k-1} \mathbf{Z}_i \right)^T \sqrt{k}\mathbf{Z} \right) + \operatorname{tr}((\sqrt{k}\mathbf{Z})^T(\sqrt{k}\mathbf{Z})) \\ &= -\operatorname{tr}\left(\frac{1}{\sqrt{k}} \sum_{i=0}^{k-1} \mathbf{Z}_i^T \frac{1}{\sqrt{k}} \sum_{i=0}^{k-1} \mathbf{Z}_i \right) \\ &\quad + \operatorname{tr}\left(\frac{1}{\sqrt{k}} \sum_{i=0}^{k-1} \mathbf{Z}_i^T \frac{1}{\sqrt{k}} \sum_{i=0}^{k-1} \mathbf{Z}_i \right) \\ &\quad - 2\operatorname{tr}\left(\frac{1}{\sqrt{k}} \left(\sum_{i=0}^{k-1} \mathbf{Z}_i \right)^T \sqrt{k}\mathbf{Z} \right) + \operatorname{tr}((\sqrt{k}\mathbf{Z})^T(\sqrt{k}\mathbf{Z})) \\ &= -\frac{1}{k}\operatorname{tr}\left(\sum_{i=0}^{k-1} \mathbf{Z}_i^T \sum_{j=0}^{k-1} \mathbf{Z}_j \right) + \left\| \frac{1}{\sqrt{k}} \sum_{i=0}^{k-1} \mathbf{Z}_i - \sqrt{k}\mathbf{Z} \right\|_F^2. \end{aligned}$$

The fact that $\operatorname{tr} \mathbf{Z}_i^T \mathbf{Z}_j = \operatorname{tr} \mathbf{Z}_j^T \mathbf{Z}_i$ and some tedious bookkeeping will show that

$$\begin{aligned} \operatorname{tr}\left(\sum_{i=0}^{k-1} \mathbf{Z}_i^T \mathbf{Z}_i \right) - \frac{1}{k}\operatorname{tr}\left(\sum_{i=0}^{k-1} \mathbf{Z}_i^T \sum_{j=0}^{k-1} \mathbf{Z}_j \right) &= \frac{k-1}{k}\operatorname{tr}\left(\sum_{i=0}^{k-1} \mathbf{Z}_i^T \mathbf{Z}_i \right) - \frac{2}{k} \sum_{i=0}^{k-1} \sum_{j=i+1}^{k-1} \operatorname{tr}(\mathbf{Z}_i^T \mathbf{Z}_j) \\ &= \frac{1}{k} \sum_{i=0}^{k-1} \sum_{j=i+1}^{k-1} \|\mathbf{Z}_i - \mathbf{Z}_j\|_F^2. \end{aligned}$$

Hence,

$$\begin{aligned} \left\| \begin{pmatrix} \mathbf{Z}_0 \\ \vdots \\ \mathbf{Z}_{k-1} \end{pmatrix} - \begin{pmatrix} \mathbf{Z} \\ \vdots \\ \mathbf{Z} \end{pmatrix} \right\|_F^2 &= \frac{1}{k} \sum_{i=0}^{k-1} \sum_{j=i+1}^{k-1} \|\mathbf{Z}_i - \mathbf{Z}_j\|_F^2 + k \left\| \frac{1}{k} \sum_{i=0}^{k-1} \mathbf{Z}_i - \mathbf{Z} \right\|_F^2 \\ &\geq \frac{1}{k} \sum_{i=0}^{k-1} \sum_{j=i+1}^{k-1} \|\mathbf{Z}_i - \mathbf{Z}_j\|_F^2 \end{aligned}$$

with equality if and only if

$$\mathbf{Z} = \widehat{\mathbf{Z}} = \frac{1}{k}\sum_{i=0}^{k-1} \mathbf{Z}_i. \qquad \square$$

These lemmas establish Theorem 5.1, since solving

$$\min_{\widehat{\mathbf{X}}_{i+1}=\mathbf{R}\widehat{\mathbf{X}}_i} \left\| \begin{pmatrix} \mathbf{X}_0 \\ \vdots \\ \mathbf{X}_{k-1} \end{pmatrix} - \begin{pmatrix} \widehat{\mathbf{X}}_0 \\ \vdots \\ \widehat{\mathbf{X}}_{k-1} \end{pmatrix} \right\|_F^2$$

is equivalent to solving

$$\min_{\widehat{\mathbf{X}}_0} \left\| \begin{pmatrix} \mathbf{X}_0 \\ \mathbf{R}^{k-1}\mathbf{X}_1 \\ \vdots \\ \mathbf{R}\mathbf{X}_{k-1} \end{pmatrix} - \begin{pmatrix} \widehat{\mathbf{X}}_0 \\ \widehat{\mathbf{X}}_0 \\ \vdots \\ \widehat{\mathbf{X}}_0 \end{pmatrix} \right\|_F^2$$

because

$$\begin{pmatrix} \mathbf{I} & & & \\ & \mathbf{R}^{k-1} & & \\ & & \ddots & \\ & & & \mathbf{R} \end{pmatrix}$$

is unitary. Therefore, by Lemma 5.4, $\widehat{\mathbf{X}}_0 = \frac{1}{k}\sum_{i=0}^{k-1} \mathbf{R}^{k-i}\mathbf{X}_i$, and

$$\min_{\widehat{\mathbf{X}}_i=\mathbf{R}^i\widehat{\mathbf{X}}_0} \left\| \begin{pmatrix} \mathbf{X}_0 \\ \vdots \\ \mathbf{X}_{k-1} \end{pmatrix} - \begin{pmatrix} \widehat{\mathbf{X}}_0 \\ \vdots \\ \widehat{\mathbf{X}}_{k-1} \end{pmatrix} \right\|_F^2 = \frac{1}{k}\sum_{i=0}^{k-1}\sum_{j=i+1}^{k-1} \|\mathbf{E}_j - \mathbf{R}^{j-i}\mathbf{E}_i\|_F^2,$$

where $\mathbf{R}^{k-i}\mathbf{X}_i = \mathbf{X}_0 + \mathbf{E}_i$. $\square$

6. Algorithms and computational results. The algorithmic structure for both the reflective and the rotational SPSVD is the same. It consists of two major steps:

1. Determine the normal $\mathbf{w}$ or the axis $\mathbf{q}$ for reflective or rotational symmetry, respectively.
2. Compute the standard SVD

$$\mathbf{U}_0\mathbf{S}_0\mathbf{V}_0^T = \frac{1}{k}(\mathbf{X}_0 + \mathbf{R}^{k-1}\mathbf{X}_1 + \mathbf{R}^{k-2}\mathbf{X}_2 + \cdots + \mathbf{R}\mathbf{X}_{k-1}),$$

where $\mathbf{R}$ is a reflector determined by $\mathbf{w}$ or a rotation about the axis determined by $\mathbf{q}$.

We seek the dominant (largest) singular values, and this can be done in a straightforward manner using the ARPACK software on a serial computer or P_ARPACK on a parallel system. Of course, one might question the use of ARPACK on dense problems. However, the timings shown in Figure 4 clearly verify that it is computationally more efficient to calculate only the leading r terms (singular values) using ARPACK

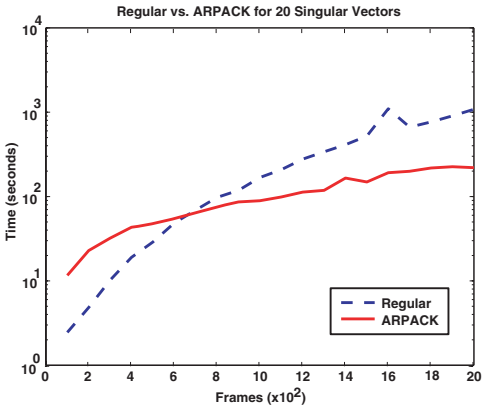


FIG. 4. *Comparison of calculating the largest* 20 *singular vectors of an HIV-*1 *protease trajectory using ARPACK and a dense SVD solver.*

instead of computing all of the singular values and then discarding $n - r$ of them for large scale matrices. One may either specify r or utilize a restarting scheme to adjust r until $\sigma_r \geq \epsilon_1 * \sigma_1 > \sigma_{r+1}$. The important computational point is that only matrix-vector products of the form

$$\mathbf{u} = \frac{1}{k}(\mathbf{X}_0 + \mathbf{R}^{k-1}\mathbf{X}_1 + \mathbf{R}^{k-2}\mathbf{X}_2 + \cdots + \mathbf{R}\mathbf{X}_{k-1})\mathbf{v}$$

are required, and this is slightly less work than is needed to compute the corresponding standard SVD of $\mathbf{X}$ without the symmetry constraint.

6.1. SPSVD in protein dynamics. Given a dynamical system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$, $\mathbf{x}(0) = \mathbf{x}_0$, there are well-known techniques for dimension reduction based upon the Gramian of the trajectory $\{\mathbf{x}(t), t \geq 0\}$. The technique is known as proper orthogonal decomposition in computational fluid dynamics, as Karhunen–Loeve decomposition in face recognition and detection, and as principal component analysis in molecular dynamics. For a system with n-dimensional state vectors, the Gramian

$$\mathcal{P} = \int_0^\infty \mathbf{x}(\tau)\mathbf{x}(\tau)^T d\tau$$

is an $n \times n$ symmetric positive (semi-)definite matrix (assuming it exists). The eigensystem of $\mathcal{P}$

$$\mathcal{P} = \mathbf{U}\mathbf{S}^2\mathbf{U}^T$$

provides an orthogonal basis via the columns of $\mathbf{U}$, and in this basis we have the representation

$$\mathbf{x}(t) = \mathbf{U}\mathbf{S}\mathbf{v}(t)$$

with the components of $\mathbf{v}(t)$ being mutually orthogonal $\mathcal{L}_2(0, \infty)$ functions. If the diagonal elements of the positive semidefinite diagonal matrix $\mathbf{S}$ decay rapidly (assuming they are in decreasing order), then a reduced basis representation of the trajectory may be obtained by discarding the trailing terms and considering the approximation

$\mathbf{x}_r = \mathbf{U}_r\mathbf{S}_r\mathbf{v}_r(t)$, where the subscript r denotes the leading r columns and/or components. This is usually approximated using snapshots consisting of values $\mathbf{x}(t_i)$ of the trajectory at discrete time points and forming the $n \times m$ matrix

$$\mathbf{X} = [\mathbf{x}(t_1), \mathbf{x}(t_2), \ldots, \mathbf{x}(t_m)].$$

The SVD of $\mathbf{X}$ provides

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T \approx \mathbf{U}_r\mathbf{S}_r\mathbf{V}_r^T,$$

where

$$\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_n, \quad \mathbf{S} = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$$

with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$. This is a direct approximation to the continuous derivation if we consider

$$\mathcal{P} \approx \frac{1}{m}\mathbf{X}\mathbf{X}^T = \frac{1}{m}\sum_i \mathbf{x}(t_i)\mathbf{x}(t_i)^T,$$

where the approximation to $\mathcal{P}$ is given by a quadrature rule. Here we are concerned with introducing symmetry constraints into this approximation when appropriate. In molecular dynamics, there is often a known spatial structural symmetry for the state variables, and the purpose of the constrained SVD approximation developed here is to impose such symmetry constraints on the approximate trajectory through the SPSVD.

This method has been implemented using P_ARPACK on a Linux cluster with 6 dual-processor nodes consisting of 1600MHz AMD Athlon processors with 1GB RAM per node and a 1GB/s Ethernet connection. The method was applied to compute the leading 20 symmetric major modes for an HIV-1 protease molecule. The molecule consists of 3120 atoms, and hence the state has 9360 degrees of freedom. The molecular dynamics trajectory consisted of 10000 time steps (snapshots). This resulted in the following:

1. The first 20 [illegible] singular vectors took 244 secs.
   This includes axis of rotation determination.
2. The first 20 standard singular vectors took 118 secs.

This may seem contradictory to the claim that the SPSVD should be as efficient as regular SVD. However, the need to compute the axis of rotation significantly adds to the run time. If more singular vectors are computed, the SPSVD indeed runs faster than regular SVD.

1. The first 50 [illegible] singular vectors took 312 secs.
   This includes axis of rotation determination.
2. The first 50 standard singular vectors took 390 secs.

These computations were done for both reflective and rotational symmetry with essentially the same computational time. The computation of the reflective normal or the axis of rotation was included in both SPSVD approximations. As this normal/axis determination is quite demanding, these computations indicate that obtaining the leading terms of the SVD is comparable for both the symmetry preserving and standard SVD cases. Moreover, both are well suited to the large scale setting when P_ARPACK is used.

It turns out that HIV-1 protease has a 2-fold rotational symmetry, and this aspect is preserved while providing good approximations to the full trajectory, as can be seen in Figure 5. Additional visualizations are available at the web site http://www.caam.rice.edu/∼sorensen/ under "recent talks."

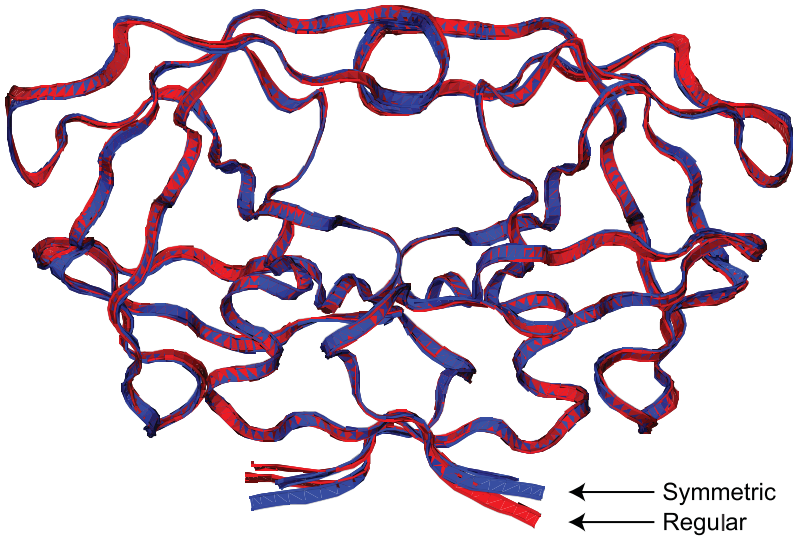


FIG. 5. *Comparison of SVD versus SPSVD. Notice the nice fit for all but the indicated region and its symmetric counterpart.*

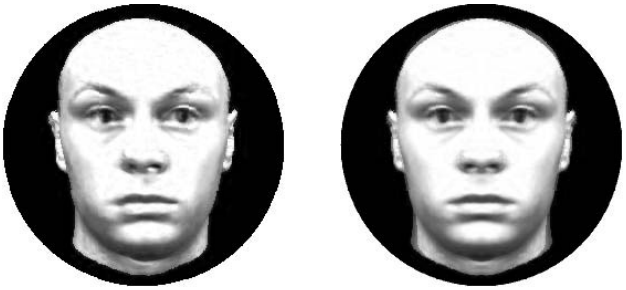
6.2. Face recognition. Generalizations of techniques described here can be used to orient faces once the plane of symmetry has been found. Once the correct orientation is attained, the SPSVD can find the best symmetric approximation to the face.

We notice that a face seems to have reflective symmetry through the vertical midline of the face (through the center of the eyes, middle of the nose, etc.). Therefore, if a face is correctly oriented, we have a reflectively symmetric data set of intensity values. The left half of the face forms $\mathbf{X}_0$, while the right half gives us $\mathbf{X}_1$. Note that the columns of $\mathbf{X}_1$ will have to be in reverse order to maintain correctly paired data points with relation to $\mathbf{X}_0$. Then, using SPSVD, we know that our best symmetric approximation will be formed by taking the average of the intensity levels of the left and right half of the face, i.e., the best symmetric approximation

$$\mathbf{S} = [\mathbf{A}\ \hat{\mathbf{A}}],$$

where $\mathbf{A} = \frac{1}{2}(\mathbf{X}_0 + \mathbf{X}_1)$ and $\hat{\mathbf{A}}$ is the matrix $\mathbf{A}$ with its columns in reverse order. The SPSVD was applied to a series of newly synthesized, laser-scanned (Cyberware TM), 256×256 gray-scaled pixel heads without hair. The face database was provided by the Max-Planck Institute for Biological Cybernetics in Tuebingen, Germany [18] (see http://www.kyb.mpg.de/publications/pdfs/pdf541.pdf). An example of one of the faces and its symmetric counterpart can be seen in Figure 6. The SPSVD gives a good approximation to the original head, while the storage is essentially cut in half. We should also note that the sudden decrease of the singular values in the SPSVD occurs at an index that is approximately half that of the regular SVD (Figure 7). This suggests that a lower rank approximation from the SPSVD could give a better approximation to the original data set when compared to a regular low rank SVD approximation.

7. Conclusion. This paper has described a mathematical formulation of a symmetry preserving singular value decomposition which has led to practical (parallel)



(a) Regular SVD (b) Symmetric SVD

FIG. 6. *Comparison of SVD versus SPSVD on faces.*

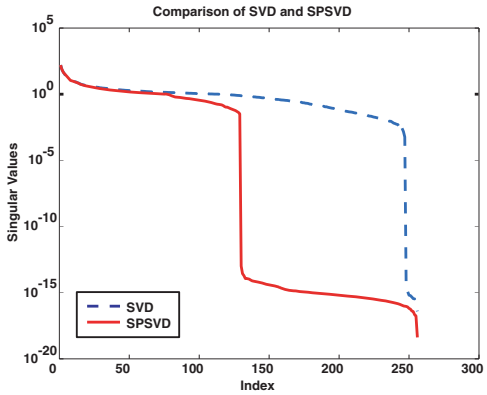


FIG. 7. *Singular values of SVD and SPSVD.*

algorithms suitable for large scale computation. Criteria and methods were given for the calculation of reflective normal and rotational axis of symmetry of objects in $\mathbb{R}^n$ that are able to overcome problems with noisy data and outliers. The resulting technique is able to compute the best low rank symmetry preserving approximation to a given set.

Acknowledgments. The authors would like to thank Prof. Lydia Kavraki for introducing us to the symmetry problem associated with PCA approximation to the HIV-1 protease trajectory. We also thank Dr. Mark Moll for many enlightening discussions concerning this problem and for the parallel computations on the HIV-1 example. Finally, we would like to acknowledge the helpful comments of Prof. Mark Embree concerning earlier versions of this manuscript.

REFERENCES

[1] M. J. ATALLAH, *On symmetry detection*, IEEE Trans. Comput., 34 (1985), pp. 663–666.
[2] N. AUBRY, W.-Y. LIAN, AND E. S. TITI, *Preserving symmetries in the proper orthogonal decomposition*, SIAM J. Sci. Comput., 14 (1993), pp. 483–505.
[3] O. COLLIOT, A. V. TUZIKOV, R. M. CESAR, AND I. BLOCH, *Approximate reflectional symmetries of fuzzy objects with an application in model-based object recognition*, Fuzzy Sets and

Systems, 147 (2004), pp. 141–163.
[4] M. Kazhdan, B. Chazelle, D. Dobkin, T. Funkhouser, and S. Rusinkiewics, *A reflective symmetry descriptor for* 3*D models*, Algorithmica, 38 (2003), pp. 201–225.
[5] M. Kirby and L. Sirovich, *Low-dimensional procedure for the characterization of human faces*, J. Opt. Soc. Amer. A, 4 (1987), pp. 519–524.
[6] M. Kirby and L. Sirovich, *Application of the Karhunen-Loeve procedure for the characterization of human faces*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 12 (1990), pp. 103–108.
[7] T. G. Kolda, R. M. Lewis, and V. Torczon, *Optimization by direct search: New perspectives on some classical and modern methods*, SIAM Rev., 45 (2003), pp. 385–482.
[8] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, SIAM, Philadelphia, 1998.
[9] G. Marola, *On the detection of the axes of symmetry of symmetric and almost symmetric planar images*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 11 (1989), pp. 104–108.
[10] P. Minovic, S. Ishikawa, and K. Kato, *Three dimensional symmetry identification part* I: *Theory*, Memoirs of the Kyushu Institute of Technology, 21 (1992), pp. 1–16.
[11] P. Minovic, S. Ishikawa, and K. Kato, *Three dimensional symmetry identification part* II: *General algorithm and its application to medical images*, Memoirs of the Kyushu Institute of Technology, 21 (1992), pp. 17–26.
[12] P. Minovic, S. Ishikawa, and K. Kato, *Symmetry identification of a* 3*D object represented by octree*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 15 (1993), pp. 507–514.
[13] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Ser. Oper. Res., Springer-Verlag, New York, 1999.
[14] D. O'Mara and R. Owens, *Measuring bilateral symmetry in digital images*, in TENCON Digital Signal Processing Applications, Vol. 1, IEEE, Piscataway, NJ, 1996, pp. 151–156.
[15] N. Smaoui and D. Armbruster, *Symmetry and the Karhunen–Loève analysis*, SIAM J. Sci. Comput., 18 (1997), pp. 1526–1532.
[16] C. Sun and J. Sherrah, 3*D symmetry detection using the extended Gaussian image*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 19 (1997), pp. 164–168.
[17] K. R. Symon, *Mechanics*, Addison–Wesley, Philippines, 1971.
[18] N. F. Troje and H. H. Bülthoff, *Face recognition under varying poses: The role of texture and shape*, Vision Research, 36 (1997), pp. 1761–1771.
[19] H. Zabrodsky, S. Peleg, and D. Avnir, *Continuous symmetry measures*, J. Amer. Chem. Soc., 114 (1992), pp. 7843–7851.
[20] H. Zabrodsky, S. Peleg, and D. Avnir, *Symmetry as a continuous feature*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 19 (1997), pp. 246–247.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 3, pp. 770–787

ON A GENERALIZED EIGENVALUE PROBLEM FOR NONSQUARE PENCILS*

DELIN CHU† AND GENE H. GOLUB‡

Abstract. In this paper a generalized eigenvalue problem for nonsquare pencils of the form $A - \lambda B$ with $A, B \in \mathbf{C}^{m\times n}$ and $m > n$, which was proposed recently by Boutry, Elad, Golub, and Milanfar [*SIAM J. Matrix Anal. Appl.*, 27 (2006), pp. 582–601], is studied. An algebraic characterization for the distance between the pair (A, B) and the pairs (A_0, B_0) with the property that for the pair (A_0, B_0) there exist l distinct eigenpairs of the form $(A_0 - \lambda_k B_0)\underline{v}_k = 0$, $k = 1, \ldots, l$, is given, which implies that this distance can be obtained by solving an optimization problem over the compact set $\{V_l : \ V_l \in \mathbf{C}^{n\times l},\ V_l^H V_l = I\}$. Furthermore, the distance between a controllable descriptor system and uncontrollable ones is also considered, an algebraic characterization is obtained, and hence a well-known result on the distance between a controllable linear time-invariant system to uncontrollable ones is extended to the descriptor systems.

Key words. nonsquare pencil, generalized eigenvalue, descriptor systems

AMS subject classifications. 15A18, 15A22, 65F10

DOI. 10.1137/050628258

1. Introduction.

The generalized eigenvalue problem $(A - \lambda B)\underline{v} = 0$ with $A, B \in \mathbf{C}^{m\times n}$ and $m > n$ for nonsquare pencil $A - \lambda B$ has attracted much attention and has been treated from theoretical and numerical points of view; see [1, 4, 6, 7, 9, 10, 11, 16, 17, 19]. Traditional methods for solving such nonsquare generalized eigenvalue problem are expected to lead to no solutions in most cases. Thus, we may search for the minimal perturbation on the pair (A, B) such that these solutions are indeed possible. This consideration leads to the following problem.

PROBLEM 1. [illegible] $A, B \in \mathbf{C}^{m\times n}$ [illegible] $l \le n$ [illegible] $m > n$ [illegible]

$$\mu_l = \inf \left\{ \| \begin{bmatrix} A_0 - A & B_0 - B \end{bmatrix} \|_F : \begin{pmatrix} A_0, B_0 \in \mathbf{C}^{m\times n},\ A_0\underline{v}_k = \lambda_k B_0 \underline{v}_k \\ \lambda_k \in \mathbf{C},\ \underline{v}_k \in \mathbf{C}^n,\ k = 1, \ldots, l \\ \lambda_k \neq \lambda_j\ \forall 1 \le k \neq j \le l \\ \operatorname{rank} \begin{bmatrix} \underline{v}_1 & \cdots & \underline{v}_l \end{bmatrix} = l \end{pmatrix} \right\}.$$

Very recently, Boutry, Elad, Golub, and Milanfar have treated two cases of Problem 1 in [4]. Starting with the case $n = 1$, they have shown that it leads to a closed form solution. They then treated the case with $n > 1$ and $l = 1$. For this case they proposed an efficient numerical algorithm and demonstrated its behavior. Finally, they pointed out that Problem 1 is complicated and is still open.

In this work we focus on Problem 1. We will show that the infimum μ_l in Problem 1 can be obtained by solving an optimization problem over the compact set $\{V_l : \ V_l \in \mathbf{C}^{m\times l},\ V_l^H V_l = I\}$, and hence Theorem 4 of [4] is extended to the general

*Received by the editors April 1, 2005; accepted for publication (in revised form) by D. Boley February 27, 2006; published electronically October 4, 2006.
http://www.siam.org/journals/simax/28-3/62825.html

†Department of Mathematics, National University of Singapore, 2 Science Drive 2, Singapore 117543 (matchudl@math.nus.edu.sg). The work of this author was supported in part by NUS grant R-146-000-047-112.

‡Department of Computer Science (SCCM), Stanford University, CA 94305-9025 (golub@sccm.stanford.edu). The work of this author was supported in part by an NSF grant.

case $0 < l \le n$. As a by-product, we also derive a formula for the distance between controllable and uncontrollable descriptor systems of the form $E\dot{x} = Ax + Bu$, which generalizes the well-known result in [3, 2, 8] on the distance between controllable and uncontrollable linear systems of the form $\dot{x} = Ax + Bu$.

Throughout this paper, the following notation will be used:

- $\sigma_i(M)$: the ith singular value of $M \in \mathbf{C}^{m\times n}$ with decreasing order, viz. $\sigma_1(M) \ge \sigma_2(M) \ge \cdots \ge \sigma_{\min\{m,n\}}(M)$;
- M^H: the conjugate transpose of M.

2. Main result. In this section we show that the infimum μ_l in Problem 1 can be obtained by solving an optimization problem over the compact set $\{V_l : V_l \in \mathbf{C}^{n\times l},\ V_l^H V_l = I\}$. Our main result is the following theorem.

THEOREM 1.

$$\mu_l = \min\left\{\sqrt{\sum_{i=l+1}^{\min\{m,2l\}} \sigma_i^2(\begin{bmatrix} AV_l & BV_l \end{bmatrix})} : \quad V_l \in \mathbf{C}^{n\times l}, V_l^H V_l = I\right\}. \tag{1}$$

We prove Theorem 1 by the following three arguments.

LEMMA 1. Let $A_0, B_0 \in \mathbf{C}^{m\times n}$. If

$$\begin{cases} A_0\underline{v}_k = \lambda_k B_0\underline{v}_k,\ \lambda_k \in \mathbf{C} \text{ and } \underline{v}_k \in \mathbf{C}^n,\ k = 1,\ldots,l, \\ \operatorname{rank}\begin{bmatrix} \underline{v}_1 & \cdots & \underline{v}_l \end{bmatrix} = l \text{ and } \lambda_k \ne \lambda_j\ (\forall 1 \le k \ne j \le l), \end{cases} \tag{2}$$

then

$$A_0\begin{bmatrix} \underline{v}_1 & \cdots & \underline{v}_l \end{bmatrix} = B_0\begin{bmatrix} \underline{v}_1 & \cdots & \underline{v}_l \end{bmatrix}\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_l \end{bmatrix}.$$

Let the unitary matrix $V_0 \in \mathbf{C}^{n\times n}$ be such that

$$\begin{bmatrix} \underline{v}_1 & \cdots & \underline{v}_l \end{bmatrix} = V_0\begin{bmatrix} \mathcal{K} \\ 0 \end{bmatrix}, \tag{3}$$

where $\mathcal{K} \in \mathbf{C}^{l\times l}$ is nonsingular. Obviously, we have

$$A_0V_0\begin{bmatrix} I_l \\ 0 \end{bmatrix} = B_0V_0\begin{bmatrix} I_l \\ 0 \end{bmatrix}\Lambda_0,$$

i.e.,

$$\begin{aligned} A_0V_0 &= \begin{bmatrix} \overset{l}{A_0^{(1)}} & \overset{n-l}{A_0^{(2)}} \end{bmatrix}, \quad B_0V_0 = \begin{bmatrix} \overset{l}{B_0^{(1)}} & \overset{n-l}{B_0^{(2)}} \end{bmatrix}, \\ A_0^{(1)} &= B_0^{(1)}\Lambda_0, \quad \Lambda_0 = \mathcal{K}\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_l \end{bmatrix}\mathcal{K}^{-1}. \end{aligned} \tag{4}$$

All eigenvalues of Λ_0 are distinct because $\lambda_k \ne \lambda_j$ for all $1 \le k \ne j \le l$. Conversely, if

$$A_0V_0 = \begin{bmatrix} \overset{l}{A_0^{(1)}} & \overset{n-l}{A_0^{(2)}} \end{bmatrix}, \quad B_0V_0 = \begin{bmatrix} \overset{l}{B_0^{(1)}} & \overset{n-l}{B_0^{(2)}} \end{bmatrix}, \quad A_0^{(1)} = B_0^{(1)}\Lambda_0,$$

where V_0 is unitary, all eigenvalues of Λ_0 are distinct, and then we know that (2) holds. Thus,

$$\left\{ (A_0, B_0) : \begin{pmatrix} A_0, B_0 \in \mathbf{C}^{m\times n},\ A_0\underline{v}_k = \lambda_k B_0 \underline{v}_k \\ \lambda_k \in \mathbf{C},\ \underline{v}_k \in \mathbf{C}^n,\ k = 1, \ldots, l \\ \lambda_k \neq \lambda_j\ \forall 1 \leq k \neq j \leq l \\ \operatorname{rank}\left[\ \underline{v}_1 \quad \cdots \quad \underline{v}_l\ \right] = l \end{pmatrix} \right\}$$

$$= \left\{ (A_0, B_0) : \begin{pmatrix} A_0, B_0 \in \mathbf{C}^{m\times n} \\ A_0V_0 = \left[\ A_0^{(1)} \quad A_0^{(2)}\ \right],\ A_0^{(1)} \in \mathbf{C}^{m\times l} \\ B_0V_0 = \left[\ B_0^{(1)} \quad B_0^{(2)}\ \right],\ B_0^{(1)} \in \mathbf{C}^{m\times l} \\ A_0^{(1)} = B_0^{(1)}\Lambda_0 \\ V_0 \text{ unitary} \\ \text{all eigenvalues of } \Lambda_0 \text{ are distinct} \end{pmatrix} \right\}.$$

Hence, we obtain that

$$\mu_l = \inf \left\{ \| \left[\ A_0 - A \quad B_0 - B\ \right] \|_F : \begin{pmatrix} A_0, B_0 \in \mathbf{C}^{m\times n},\ A_0\underline{v}_k = \lambda_k B_0 \underline{v}_k \\ \lambda_k \in \mathbf{C},\ \underline{v}_k \in \mathbf{C}^n,\ k = 1, \ldots, l \\ \lambda_k \neq \lambda_j\ \forall 1 \leq k \neq j \leq l \\ \operatorname{rank}\left[\ \underline{v}_1 \quad \cdots \quad \underline{v}_l\ \right] = l \end{pmatrix} \right\}$$

$$= \inf \left\{ \| \left[\ A_0 - A \quad B_0 - B\ \right] \|_F : \begin{pmatrix} A_0, B_0 \in \mathbf{C}^{m\times n} \\ A_0V_0 = \left[\ A_0^{(1)} \quad A_0^{(2)}\ \right],\ A_0^{(1)} \in \mathbf{C}^{m\times l} \\ B_0V_0 = \left[\ B_0^{(1)} \quad B_0^{(2)}\ \right],\ B_0^{(1)} \in \mathbf{C}^{m\times l} \\ A_0^{(1)} = B_0^{(1)}\Lambda_0 \\ V_0 \in \mathbf{C}^{n\times n} \text{ unitary} \\ \text{all eigenvalues of } \Lambda_0 \text{ are distinct} \end{pmatrix} \right\}$$

$$= \inf \left\{ \| \left[\ A_0V_0 - AV_0 \quad B_0V_0 - BV_0\ \right] \|_F : \right.$$

$$\left. \begin{pmatrix} A_0, B_0 \in \mathbf{C}^{m\times n} \\ A_0V_0 = \left[\ A_0^{(1)} \quad A_0^{(2)}\ \right],\ A_0^{(1)} \in \mathbf{C}^{m\times l} \\ B_0V_0 = \left[\ B_0^{(1)} \quad B_0^{(2)}\ \right],\ B_0^{(1)} \in \mathbf{C}^{m\times l} \\ A_0^{(1)} = B_0^{(1)}\Lambda_0 \\ V_0 \in \mathbf{C}^{n\times n} \text{ unitary} \\ \text{all eigenvalues of } \Lambda_0 \text{ are distinct} \end{pmatrix} \right\}. \tag{5}$$

Next, for any unitary matrix $V_0 \in \mathbf{C}^{n\times n}$, partition

$$\begin{array}{c} \quad l \quad n-l \\ V_0 = \left[\ V_l \quad \mathcal{V}_l\ \right]. \end{array}$$

We have from (5) that

$$\mu_l = \inf \left\{ \| \begin{bmatrix} A_0^{(1)} - AV_l & A_0^{(2)} - A\mathcal{V}_l & B_0^{(1)} - BV_l & B_0^{(2)} - B\mathcal{V}_l \end{bmatrix} \|_F : \right.$$

$$\left. \begin{pmatrix} A_0, B_0 \in \mathbf{C}^{m\times n} \\ A_0 V_0 = \begin{bmatrix} A_0^{(1)} & A_0^{(2)} \end{bmatrix} \\ B_0 V_0 = \begin{bmatrix} B_0^{(1)} & B_0^{(2)} \end{bmatrix} \\ A_0^{(1)} = B_0^{(1)} \Lambda_0 \\ A_0^{(1)}, B_0^{(1)} \in \mathbf{C}^{m\times l} \\ V_0 \text{ unitary} \\ \text{all eigenvalues of } \Lambda_0 \text{ are distinct} \end{pmatrix} \right\}$$

$$= \inf \left\{ \| \begin{bmatrix} A_0^{(1)} - AV_l & B_0^{(1)} - BV_l \end{bmatrix} \|_F : \right.$$

$$\left. \times \begin{pmatrix} A_0^{(1)}, B_0^{(1)} \in \mathbf{C}^{m\times l} \\ V_l \in \mathbf{C}^{n\times l}, V_l^H V_l = I \\ A_0^{(1)} = B_0^{(1)} \Lambda_0 \\ \text{all eigenvalues of } \Lambda_0 \text{ are distinct} \end{pmatrix} \right\}$$

(by taking $A_0^{(2)} = A\mathcal{V}_l,\ B_0^{(2)} = B\mathcal{V}_l$)

$$= \inf \left\{ \| \begin{bmatrix} B_0^{(1)}\Lambda_0 - AV_l & B_0^{(1)} - BV_l \end{bmatrix} \|_F : \right.$$

$$\left. \times \begin{pmatrix} B_0^{(1)} \in \mathbf{C}^{m\times l} \\ V_l \in \mathbf{C}^{n\times l}, V_l^H V_l = I \\ \text{all eigenvalues of } \Lambda_0 \text{ are distinct} \end{pmatrix} \right\}$$

$$= \inf \left\{ \| \begin{bmatrix} AV_l & BV_l \end{bmatrix} - B_0^{(1)} \begin{bmatrix} \Lambda_0 & I \end{bmatrix} \|_F : \right.$$

$$\left. \times \begin{pmatrix} B_0^{(1)} \in \mathbf{C}^{m\times l} \\ V_l \in \mathbf{C}^{n\times l}, V_l^H V_l = I \\ \text{all eigenvalues of } \Lambda_0 \text{ are distinct} \end{pmatrix} \right\}.$$

(6)

Remark 2. In (6), Λ_0 must satisfy that all its eigenvalues are distinct. To remove this constraint from (6), we denote

$$\mathbf{S}_1 := \{\Lambda_0 :\ \Lambda_0 \in \mathbf{C}^{l\times l}\}$$

and

$$\tau_l := \inf \left\{ \| \begin{bmatrix} AV_l & BV_l \end{bmatrix} - B_0^{(1)} \begin{bmatrix} \Lambda_0 & I \end{bmatrix} \|_F : \begin{pmatrix} B_0^{(1)} \in \mathbf{C}^{m\times l} \\ V_l \in \mathbf{C}^{n\times l}, V_l^H V_l = I \\ \Lambda_0 \in \mathbf{S}_1 \end{pmatrix} \right\}.$$

Then we wish to show that $\mu_l = \tau_l$.

Since

$$\{\Lambda_0 : \ \Lambda_0 \in \mathbf{C}^{l\times l}, \text{ all eigenvalues of } \Lambda_0 \text{ are distinct}\} \subset \mathbf{S}_1,$$

we have

$$\mu_l \geq \tau_l. \tag{7}$$

Conversely, for any $\epsilon > 0$, let $B_0^{(1)}(\epsilon) \in \mathbf{C}^{m\times l}$, $V_l(\epsilon) \in \mathbf{C}^{n\times l}$, and $\Lambda_0(\epsilon) \in \mathbf{S}_1$ with $V_l^H(\epsilon)V_l(\epsilon) = I$ be such that

$$\left| \| \begin{bmatrix} AV_l(\epsilon) & BV_l(\epsilon) \end{bmatrix} - B_0^{(1)}(\epsilon) \begin{bmatrix} \Lambda_0(\epsilon) & I \end{bmatrix} \|_F - \tau_l \right| \leq \frac{\epsilon}{2}.$$

Note that there exists a matrix $\Delta\Lambda_0(\epsilon) \in \mathbf{C}^{l\times l}$ satisfying the following:

(i) all eigenvalues of $\Lambda_0(\epsilon) + \Delta\Lambda_0(\epsilon)$ are distinct;

(ii) $\|\Delta\Lambda_0(\epsilon)\|$ is small enough such that

$$\|B_0^{(1)}(\epsilon)\|_F \|\Delta\Lambda_0(\epsilon)\|_F \leq \frac{\epsilon}{2}.$$

A simple calculation gives that

$$\begin{aligned}
&\left| \| \begin{bmatrix} AV_l(\epsilon) & BV_l(\epsilon) \end{bmatrix} - B_0^{(1)}(\epsilon) \begin{bmatrix} \Lambda_0(\epsilon) + \Delta\Lambda_0(\epsilon) & I \end{bmatrix} \|_F - \tau_l \right| \\
&\leq \Big| \| \begin{bmatrix} AV_l(\epsilon) & BV_l(\epsilon) \end{bmatrix} - B_0^{(1)}(\epsilon) \begin{bmatrix} \Lambda_0(\epsilon) + \Delta\Lambda_0(\epsilon) & I \end{bmatrix} \|_F \\
&\quad - \| \begin{bmatrix} AV_l(\epsilon) & BV_l(\epsilon) \end{bmatrix} - B_0^{(1)}(\epsilon) \begin{bmatrix} \Lambda_0(\epsilon) & I \end{bmatrix} \|_F \Big| \\
&\quad + \left| \| \begin{bmatrix} AV_l(\epsilon) & BV_l(\epsilon) \end{bmatrix} - B_0^{(1)}(\epsilon) \begin{bmatrix} \Lambda_0(\epsilon) & I \end{bmatrix} \|_F - \tau_l \right| \\
&\leq \|B_0^{(1)}(\epsilon)\Delta\Lambda_0(\epsilon)\|_F + \frac{\epsilon}{2} \\
&\leq \|B_0^{(1)}(\epsilon)\|_F \ \|\Delta\Lambda_0(\epsilon)\|_F + \frac{\epsilon}{2} \\
&\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,
\end{aligned}$$

which yields

$$\| \begin{bmatrix} AV_l(\epsilon) & BV_l(\epsilon) \end{bmatrix} - B_0^{(1)}(\epsilon) \begin{bmatrix} \Lambda_0(\epsilon) + \Delta\Lambda_0(\epsilon) & I \end{bmatrix} \|_F \leq \tau_l + \epsilon.$$

Thus,

$$\begin{aligned}
\mu_l = \inf \Bigg\{ &\| \begin{bmatrix} AV_l & BV_l \end{bmatrix} - B_0^{(1)} \begin{bmatrix} \Lambda_0 & I \end{bmatrix} \|_F : \\
&\begin{pmatrix} B_0^{(1)} \in \mathbf{C}^{m\times l} \\ V_l \in \mathbf{C}^{n\times l}, V_l^H V_l = I \\ \text{all eigenvalues of } \Lambda_0 \text{ are distinct} \end{pmatrix} \Bigg\} \\
&\leq \| \begin{bmatrix} AV_l(\epsilon) & BV_l(\epsilon) \end{bmatrix} - B_0^{(1)}(\epsilon) \begin{bmatrix} \Lambda_0(\epsilon) + \Delta\Lambda_0(\epsilon) & I \end{bmatrix} \|_F \\
&\leq \tau_l + \epsilon,
\end{aligned}$$

and furthermore, by letting $\epsilon \to 0$ we get that

$$\mu_l \leq \tau_l. \tag{8}$$

Now we have from (7) and (8) that

$$\begin{aligned}\mu_l &= \tau_l \\ &= \inf \left\{ \| \begin{bmatrix} AV_l & BV_l \end{bmatrix} - B_0^{(1)} \begin{bmatrix} \Lambda_0 & I \end{bmatrix} \|_F : \begin{pmatrix} B_0^{(1)} \in \mathbf{C}^{m\times l} \\ V_l \in \mathbf{C}^{n\times l}, V_l^H V_l = I \\ \Lambda_0 \in \mathbf{S}_1 \end{pmatrix} \right\}.\end{aligned} \tag{9}$$

3. For any $\Lambda_0 \in \mathbf{S}_1$, let the QR factorization of $\begin{bmatrix} \Lambda_0 & I \end{bmatrix}$ be

$$\begin{bmatrix} \Lambda_0 & I \end{bmatrix} \begin{bmatrix} P_1 & \hat{P}_1 \\ P_2 & \hat{P}_2 \end{bmatrix} = \begin{bmatrix} 0 & \Delta_0 \end{bmatrix},$$
$$\Delta_0, P_1, P_2 \in \mathbf{C}^{l\times l}, \quad \Delta_0 \in \mathbf{C}^{l\times l} \text{ is nonsingular.}$$

Then it can be verified that

$$P_1 \text{ is nonsingular.}$$

Hence, we get by taking $B_0^{(1)} = \begin{bmatrix} AV_l & BV_l \end{bmatrix} \left[\begin{smallmatrix} \hat{P}_1 \\ \hat{P}_2 \end{smallmatrix}\right] \Delta_0^{-1}$ in (9) that

$$\begin{aligned}\mu_l &= \inf \left\{ \left\| \begin{bmatrix} AV_l & BV_l \end{bmatrix} \begin{bmatrix} P_1 & \hat{P}_1 \\ P_2 & \hat{P}_2 \end{bmatrix} - B_0^{(1)} \begin{bmatrix} 0 & \Delta_0 \end{bmatrix} \right\|_F : \right. \\ &\qquad \left. \times \begin{pmatrix} B_0^{(1)} \in \mathbf{C}^{m\times l} \\ V_l \in \mathbf{C}^{n\times l}, V_l^H V_l = I \\ \Lambda_0 \in \mathbf{S}_1 \end{pmatrix} \right\} \\ &= \inf \left\{ \left\| \begin{bmatrix} AV_l & BV_l \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} \right\|_F : \begin{pmatrix} V_l \in \mathbf{C}^{n\times l}, V_l^H V_l = I \\ \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} \in \mathbf{S}_2 \end{pmatrix} \right\},\end{aligned} \tag{10}$$

where

$$\mathbf{S}_2 := \left\{ \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} : \ P_1, P_2 \in \mathbf{C}^{l\times l}, \ P_1^H P_1 + P_2^H P_2 = I, \ P_1 \text{ is nonsingular} \right\}.$$

Obviously, (10) is an optimization problem over the compact set

$$\{V_l : \ V_l \in \mathbf{C}^{n\times l}, \ V_l^H V_l = I\}$$

and noncompact set $\mathbf{S}_2$. To simplify (10), we have to eliminate the noncompact set $\mathbf{S}_2$. Because the smallest compact set containing $\mathbf{S}_2$ is

$$\mathbf{S}_3 := \left\{ \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} : \ P_1, P_2 \in \mathbf{C}^{l\times l}, \ P_1^H P_1 + P_2^H P_2 = I \right\},$$

now we consider $\mathbf{S}_3$ and show that the set $\mathbf{S}_2$ in (10) can be replaced by $\mathbf{S}_3$. To do so, we define

$$\nu_l := \inf \left\{ \left\| \begin{bmatrix} AV_l & BV_l \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} \right\|_F : \begin{pmatrix} V_l \in \mathbf{C}^{n\times l}, V_l^H V_l = I \\ \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} \in \mathbf{S}_3 \end{pmatrix} \right\}.$$

Since $\mathbf{S}_2 \subset \mathbf{S}_3$, we know

$$\mu_l \geq \nu_l. \tag{11}$$

On the other hand, for any $1 > \epsilon > 0$, let $\left[\begin{smallmatrix} P_1(\epsilon) \\ P_2(\epsilon) \end{smallmatrix}\right] \in \mathbf{S}_3$ and $V_l(\epsilon) \in \mathbf{C}^{n\times l}$ with $V_l^H(\epsilon)V_l(\epsilon) = I$ be such that

$$\left\| \nu_l - \begin{bmatrix} AV_l(\epsilon) & BV_l(\epsilon) \end{bmatrix} \begin{bmatrix} P_1(\epsilon) \\ P_2(\epsilon) \end{bmatrix} \right\|_F \leq \frac{\epsilon}{2}. \tag{12}$$

Denote the cosine-sine decomposition (CSD) [14] of $\left[\begin{smallmatrix} P_1(\epsilon) \\ P_2(\epsilon) \end{smallmatrix}\right]$ by

$$P_1(\epsilon) = U_1 \begin{bmatrix} c_1 & & & & & \\ & \ddots & & & & \\ & & c_k & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix} U_3,$$

$$P_2(\epsilon) = U_2 \begin{bmatrix} s_1 & & & & & \\ & \ddots & & & & \\ & & s_k & & & \\ & & & 1 & & \\ & & & & \ddots & \\ & & & & & 1 \end{bmatrix} U_3,$$

where U_1, U_2, and U_3 are unitary, $0 \leq k \leq l$, and

$$0 < c_i \leq 1, \quad 0 \leq s_i < 1, \quad c_i^2 + s_i^2 = 1, \quad i = 1, \ldots, k.$$

Set

$$\Delta P_1(\epsilon) = U_1 \begin{bmatrix} 0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & \eta & & \\ & & & & \ddots & \\ & & & & & \eta \end{bmatrix} U_3,$$

$$\Delta P_2(\epsilon) = U_2 \begin{bmatrix} 0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & \sqrt{1-\eta^2} - 1 & & \\ & & & & \ddots & \\ & & & & & \sqrt{1-\eta^2} - 1 \end{bmatrix} U_3,$$

where η is small enough that

$$0 < \eta < 1, \quad \eta\sqrt{2l} \parallel \begin{bmatrix} A & B \end{bmatrix} \parallel_F \leq \frac{\epsilon}{2}.$$

We have

$$P_1(\epsilon) + \Delta P_1(\epsilon) \text{ is nonsingular}, \quad \begin{bmatrix} P_1(\epsilon) + \Delta P_1(\epsilon) \\ P_2(\epsilon) + \Delta P_2(\epsilon) \end{bmatrix} \in \mathbf{S}_2,$$

and

$$\left\| \begin{bmatrix} \Delta P_1(\epsilon) \\ \Delta P_2(\epsilon) \end{bmatrix} \right\|_F \leq \eta\sqrt{2l}, \quad \parallel \begin{bmatrix} A & B \end{bmatrix} \parallel_F \left\| \begin{bmatrix} \Delta P_1(\epsilon) \\ \Delta P_2(\epsilon) \end{bmatrix} \right\|_F \leq \frac{\epsilon}{2}.$$

Consequently, we obtain

$$\begin{aligned}
&\left| \nu_l - \left\| \begin{bmatrix} AV_l(\epsilon) & BV_l(\epsilon) \end{bmatrix} \begin{bmatrix} P_1(\epsilon) + \Delta P_1(\epsilon) \\ P_2(\epsilon) + \Delta P_2(\epsilon) \end{bmatrix} \right\|_F \right| \\
\leq &\left| \nu_l - \left\| \begin{bmatrix} AV_l(\epsilon) & BV_l(\epsilon) \end{bmatrix} \begin{bmatrix} P_1(\epsilon) \\ P_2(\epsilon) \end{bmatrix} \right\|_F \right| \\
&+ \left| \left\| \begin{bmatrix} AV_l(\epsilon) & BV_l(\epsilon) \end{bmatrix} \begin{bmatrix} P_1(\epsilon) \\ P_2(\epsilon) \end{bmatrix} \right\|_F \right. \\
&\left. - \left\| \begin{bmatrix} AV_l(\epsilon) & BV_l(\epsilon) \end{bmatrix} \begin{bmatrix} P_1(\epsilon) + \Delta P_1(\epsilon) \\ P_2(\epsilon) + \Delta P_2(\epsilon) \end{bmatrix} \right\|_F \right| \\
\leq &\frac{\epsilon}{2} + \left\| \begin{bmatrix} AV_l(\epsilon) & BV_l(\epsilon) \end{bmatrix} \begin{bmatrix} \Delta P_1(\epsilon) \\ \Delta P_2(\epsilon) \end{bmatrix} \right\|_F \\
\leq &\frac{\epsilon}{2} + \parallel \begin{bmatrix} AV_l(\epsilon) & BV_l(\epsilon) \end{bmatrix} \parallel_F \left\| \begin{bmatrix} \Delta P_1(\epsilon) \\ \Delta P_2(\epsilon) \end{bmatrix} \right\|_F \\
\leq &\frac{\epsilon}{2} + \parallel \begin{bmatrix} A & B \end{bmatrix} \parallel_F \left\| \begin{bmatrix} \Delta P_1(\epsilon) \\ \Delta P_2(\epsilon) \end{bmatrix} \right\|_F \qquad \text{(since } V_l^H(\epsilon)V_l(\epsilon) = I\text{)} \\
\leq &\frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,
\end{aligned}$$

i.e.,

$$\left\| \begin{bmatrix} AV_l(\epsilon) & BV_l(\epsilon) \end{bmatrix} \begin{bmatrix} P_1(\epsilon) + \Delta P_1(\epsilon) \\ P_2(\epsilon) + \Delta P_2(\epsilon) \end{bmatrix} \right\|_F \leq \nu_1 + \epsilon,$$

which together with (10) and the fact $\left[\begin{smallmatrix} P_1(\epsilon)+\Delta P_1(\epsilon) \\ P_2(\epsilon)+\Delta P_2(\epsilon) \end{smallmatrix} \right] \in \mathbf{S}_2$ means that

$$\mu_l \leq \left\| \begin{bmatrix} AV_l(\epsilon) & BV_l(\epsilon) \end{bmatrix} \begin{bmatrix} P_1(\epsilon) + \Delta P_1(\epsilon) \\ P_2(\epsilon) + \Delta P_2(\epsilon) \end{bmatrix} \right\|_F \leq \nu_l + \epsilon.$$

Because $0 < \epsilon < 1$ can be arbitrarily small, we must have

$$\mu_l \leq \nu_l. \tag{13}$$

Therefore, we get from (11) and (13) that

$$\mu_l = \nu_l = \inf \left\{ \left\| \begin{bmatrix} AV_l & BV_l \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} \right\|_F : \begin{pmatrix} V_l \in \mathbf{C}^{n\times l}, V_l^H V_l = I \\ \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} \in \mathbf{S}_3 \end{pmatrix} \right\}$$

$$= \inf \left\{ \sqrt{\sum_{i=l+1}^{\min\{m,2l\}} \sigma_i^2(\begin{bmatrix} AV_l & BV_l \end{bmatrix})} : \quad V_l \in \mathbf{C}^{n\times l}, V_l^H V_l = I \right\}$$

$$= \min \left\{ \sqrt{\sum_{i=l+1}^{\min\{m,2l\}} \sigma_i^2(\begin{bmatrix} AV_l & BV_l \end{bmatrix})} : \quad V_l \in \mathbf{C}^{n\times l}, V_l^H V_l = I \right\}$$

(since the set $\{V_l : V_l \in \mathbf{C}^{n\times l}, V_l^H V_l = I\}$ is compact).

Equivalently, (1) holds. □

COROLLARY 2.

$$\mu_n = \sqrt{\sum_{i=n+1}^{\min\{m,2n\}} \sigma_i^2(\begin{bmatrix} A & B \end{bmatrix})}.$$

[illegible]. By Theorem 1 we have

$$\mu_n = \min \left\{ \sqrt{\sum_{i=n+1}^{\min\{m,2n\}} \sigma_i^2(\begin{bmatrix} AV_n & BV_n \end{bmatrix})} : \quad V_n \in \mathbf{C}^{n\times n}, V_n^H V_n = I \right\}$$

$$= \min \left\{ \sqrt{\sum_{i=n+1}^{\min\{m,2n\}} \sigma_i^2 \left(\begin{bmatrix} A & B \end{bmatrix} \begin{bmatrix} V_n & 0 \\ 0 & V_n \end{bmatrix} \right)} : \quad V_n \in \mathbf{C}^{n\times n}, V_n^H V_n = I \right\}$$

$$= \min \left\{ \sqrt{\sum_{i=n+1}^{\min\{m,2n\}} \sigma_i^2(\begin{bmatrix} A & B \end{bmatrix})} : \quad V_n \in \mathbf{C}^{n\times n}, V_n^H V_n = I \right\}$$

$$\left(\text{since } \begin{bmatrix} V_n & 0 \\ 0 & V_n \end{bmatrix} \text{ is unitary} \right)$$

$$= \sqrt{\sum_{i=n+1}^{\min\{m,2n\}} \sigma_i^2(\begin{bmatrix} A & B \end{bmatrix})}. \qquad \square$$

Corollary 2 indicates clearly that the formula (1) can be simplified significantly when $l = n$. In the following we rederive Theorems 2 and 4 of [4].

COROLLARY 3.

$$\mu_1 = \min\{\sigma_2(\begin{bmatrix} A\underline{v} & B\underline{v} \end{bmatrix}) : \ \underline{v} \in \mathbf{C}^n, \ \|\underline{v}\|_2 = 1\}$$

$$= \inf \left\{ \frac{\|(A - \lambda B)\underline{v}\|_2}{\sqrt{1+|\lambda|^2}} : \ \underline{v} \in \mathbf{C}^n, \ \|\underline{v}\|_2 = 1, \ \lambda \in \mathbf{C} \right\}. \tag{14}$$

[illegible] $n = 1$ $A = \underline{a}$ [illegible] $B = \underline{b}$ [illegible]

$$\mu_1 = \inf \left\{ \frac{\|\underline{a} - \lambda \underline{b}\|_2}{\sqrt{1+|\lambda|^2}} : \ \lambda \in \mathbf{C} \right\}. \tag{15}$$

[illegible]. By Theorem 1 we have

$$\mu_1 = \min\{\sigma_2(\begin{bmatrix} A\underline{v} & B\underline{v} \end{bmatrix}) : \ \underline{v} \in \mathbf{C}^n, \ \|\underline{v}\|_2 = 1\}.$$

Furthermore, because $l = 1$, we have $\mathbf{S}_2 = \{\frac{1}{\sqrt{1+|\lambda|^2}}\left[\begin{smallmatrix} 1 \\ -\lambda \end{smallmatrix}\right] : \lambda \in \mathbf{C}\}$; thus, using (10) we have that

$$\mu_1 = \inf\left\{\| \left[\begin{array}{cc} A\underline{v} & B\underline{v} \end{array}\right] \left[\begin{array}{c} 1 \\ -\lambda \end{array}\right] (1+|\lambda|^2)^{-1/2}\|_F : \ \underline{v} \in \mathbf{C}^n, \|\underline{v}\|_2 = 1, \ \lambda \in \mathbf{C}\right\}$$
$$= \inf\left\{\frac{\|(A-\lambda B)\underline{v}\|_2}{\sqrt{1+|\lambda|^2}} : \ \underline{v} \in \mathbf{C}^n, \ \|\underline{v}\|_2 = 1, \ \lambda \in \mathbf{C}\right\}. \qquad \square$$

It is possible that the infimum μ_l in Problem 1 is not attainable and any pencil achieving this infimum has no l distinct eigenvalues but has eigenvalues at infinity, as shown by the following example.

EXAMPLE 1.[1] Let

$$A = \left[\begin{array}{c} 0 \\ 1 \end{array}\right], \quad B = \left[\begin{array}{c} 0.1 \\ 0 \end{array}\right].$$

Then according to Corollary 2,

$$\mu_1 = 0.1.$$

This infimum cannot be attained by any pencil (A_0, B_0) which has a finite eigenvalue. In other words, all pencils (A_0, B_0) achieving this infimum have no finite eigenvalue but have an eigenvalue at infinity.

The pencil (A_0, B_0) has no eigenvalues at infinity if B_0 is of full column rank. Hence, naturally we may also add the additional constraint $\operatorname{rank}(B_0) = n$ to Problem 1 and consider the following related problem.

PROBLEM 2. Given $A, B \in \mathbf{C}^{m\times n}$ [illegible] $l \le n$ [illegible] $m > n$ [illegible]

$$\hat{\mu}_l = \inf\left\{\| \left[\begin{array}{cc} A_0 - A & B_0 - B \end{array}\right] \|_F : \ \left(\begin{array}{c} A_0, B_0 \in \mathbf{C}^{m\times n}, \ \operatorname{rank}(B_0) = n, \\ A_0\underline{v}_k = \lambda_k B_0 \underline{v}_k \\ \lambda_k \in \mathbf{C}, \ \underline{v}_k \in \mathbf{C}^n, \ k = 1, \ldots, l \\ \lambda_k \ne \lambda_j \ \forall 1 \le k \ne j \le l \\ \operatorname{rank}\left[\begin{array}{ccc} \underline{v}_1 & \cdots & \underline{v}_l \end{array}\right] = l \end{array}\right)\right\}.$$

The following result indicates that the infimum μ_l in Problem 1 and the infimum $\hat{\mu}_l$ in Problem 2 are the same.

THEOREM 4.

$$\hat{\mu}_l = \mu_l.$$

[illegible]. It is trivial that

$$\mu_l \le \hat{\mu}_l, \tag{16}$$

so we need only to show that $\hat{\mu}_l \le \mu_l$.

For any $\epsilon > 0$, let $A_0, B_0 \in \mathbf{C}^{m\times n}$ satisfying

$$A_0\underline{v}_k = \lambda_k B_0 \underline{v}_k, \quad \lambda_k \in \mathbf{C}, \quad \lambda_k \ne \lambda_j \ \forall 1 \le k \ne j \le l,$$

[1] This interesting example is provided by an anonymous referee.

and

$$\underline{v}_k \in \mathbf{C}^n, \quad \text{rank}\left[\begin{array}{ccc} \underline{v}_1 & \cdots & \underline{v}_l \end{array}\right] = l$$

be such that

$$\left|\mu_l - \|\left[\begin{array}{cc} A - A_0 & B - B_0 \end{array}\right]\|_F\right| \leq \frac{\epsilon}{2}.$$

Same as Argument 1, let the QR factorization of $\left[\begin{array}{ccc} \underline{v}_1 & \cdots & \underline{v}_l \end{array}\right]$ be given by (3), then (4) holds. Next, let the QR factorization of $B_0^{(1)}$ be

$$B_0^{(1)} = U_0 \left[\begin{array}{c} B_{11} \\ 0 \end{array}\right],$$

where U_0 is unitary and $B_{11} \in \mathbf{C}^{l\times l}$. Then we have using (4) that $U_0^H A_0 V_0$ and $U_0^H B_0 V_0$ are of the forms

$$U_0^H A_0 V_0 = \left[\begin{array}{cc} B_{11}\Lambda_0 & A_{12} \\ 0 & A_{22} \end{array}\right], \quad U_0^H B_0 V_0 = \left[\begin{array}{cc} B_{11} & B_{12} \\ 0 & B_{22} \end{array}\right]$$

since there always exist $\Delta B_{11} \in \mathbf{C}^{l\times l}$ and $\Delta B_{22} \in \mathbf{C}^{(m-l)\times(n-l)}$ such that

$$\text{rank}(B_{11} + \Delta B_{11}) = l, \quad \text{rank}(B_{22} + \Delta B_{22}) = n - l,$$

and

$$\|\Delta B_{11}\|_F \leq \frac{\epsilon}{4}(1 + \|\Lambda_0\|_F)^{-1}, \quad \|\Delta B_{22}\|_F \leq \frac{\epsilon}{4}.$$

Denote

$$\Delta A_0 = U_0 \left[\begin{array}{cc} \Delta B_{11}\Lambda_0 & 0 \\ 0 & 0 \end{array}\right] V_0^H, \quad \Delta B_0 = U_0 \left[\begin{array}{cc} \Delta B_{11} & 0 \\ 0 & \Delta B_{22} \end{array}\right] V_0^H.$$

Then the pencil $(A_0 + \Delta A_0, B_0 + \Delta B_0)$ satisfies

$$\text{rank}(B_0 + \Delta B_0) = n, \quad (A_0 + \Delta A_0)\underline{v}_k = \lambda_k(B_0 + \Delta B_0)\underline{v}_k, \quad k = 1, \ldots, l, \tag{17}$$

$$\begin{aligned} \|\left[\begin{array}{cc} \Delta A_0 & \Delta B_0 \end{array}\right]\|_F &= \left\|\left[\begin{array}{cccc} \Delta B_{11}\Lambda_0 & 0 & \Delta B_{11} & 0 \\ 0 & 0 & 0 & \Delta B_{22} \end{array}\right]\right\|_F \\ &\leq \|\Delta B_{11}\|_F(1 + \|\Lambda_0\|_F) + \|\Delta B_{22}\|_F \\ &\leq \frac{\epsilon}{2}, \end{aligned} \tag{18}$$

and

$$\begin{aligned} &\left|\mu_l - \|\left[\begin{array}{cc} A - (A_0 + \Delta A_0) & B - (B_0 + \Delta B_0) \end{array}\right]\|_F\right| \\ &\leq \left|\mu_l - \|\left[\begin{array}{cc} A - A_0 & B - B_0 \end{array}\right]\|_F\right| \\ &\quad + \left|\|\left[\begin{array}{cc} A - A_0 & B - B_0 \end{array}\right]\|_F - \|\left[\begin{array}{cc} A - (A_0 + \Delta A_0) & B - (B_0 + \Delta B_0) \end{array}\right]\|_F\right| \\ &\leq \frac{\epsilon}{2} + \|\left[\begin{array}{cc} \Delta A_0 & \Delta B_0 \end{array}\right]\|_F \\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \end{aligned} \tag{19}$$

Because (17) holds and (19) implies

$$\| \begin{bmatrix} A-(A_0+\Delta A_0) & B-(B_0+\Delta B_0) \end{bmatrix} \|_F \le \mu_l + \epsilon,$$

we obtain

$$\hat{\mu}_l \le \| \begin{bmatrix} A-(A_0+\Delta A_0) & B-(B_0+\Delta B_0) \end{bmatrix} \|_F \le \mu_l + \epsilon.$$

Note that ϵ can be arbitrarily small, so, we must have

$$\hat{\mu}_l \le \mu_l. \tag{20}$$

Hence, Theorem 4 follows directly from (16) and (20). □

According to the proofs of Theorems 1 and 4, once the infimum μ_l is computed, for any $1 > \epsilon > 0$, we can always find an $O(\epsilon)$-optimal pencil (A_0, B_0) in the sense that (A_0, B_0) has l distinct eigenvalues, $\operatorname{rank}(B_0) = n$ (so (A_0, B_0) has no eigenvalues at infinity), and

$$\| \begin{bmatrix} A-A_0 & B-B_0 \end{bmatrix} \|_F \le \mu_l + 2\epsilon.$$

Such a pencil can be obtained by the following procedure:

- Let $V_l = \begin{bmatrix} \underline{v}_1 & \cdots & \underline{v}_l \end{bmatrix}$ with $V_l^H V_l = I$ solve (1), i.e.,

$$\mu_l = \sqrt{\sum_{i=l+1}^{\min\{m,2l\}} \sigma_i^2(\begin{bmatrix} AV_l & BV_l \end{bmatrix})}.$$

1. Compute the SVD of $\begin{bmatrix} AV_l & BV_l \end{bmatrix}$ to get matrix $\begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$ such that

$$\begin{bmatrix} P_1 \\ P_2 \end{bmatrix} \in \mathbf{S}_3, \qquad \mu_l = \sqrt{\sum_{i=l+1}^{\min\{m,2l\}} \sigma_i^2(\begin{bmatrix} AV_l & BV_l \end{bmatrix})} = \left\| \begin{bmatrix} AV_l & BV_l \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} \right\|_F.$$

– If P_1 is singular, then similar to the construction of $\begin{bmatrix} \Delta P_1(\epsilon) \\ \Delta P_2(\epsilon) \end{bmatrix}$ in Argument 3 we construct $\begin{bmatrix} \Delta P_1 \\ \Delta P_2 \end{bmatrix}$ such that

$$P_1 + \Delta P_1 \text{ is nonsingular}, \quad \| \begin{bmatrix} A & B \end{bmatrix} \|_F \left\| \begin{bmatrix} \Delta P_1 \\ \Delta P_2 \end{bmatrix} \right\|_F \le \frac{\epsilon}{2}.$$

Set

$$P_1 := P_1 + \Delta P_1, \quad P_2 := P_2 + \Delta P_2.$$

We have

$$\left| \mu_l - \left\| \begin{bmatrix} AV_l & BV_l \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} \right\|_F \right| \le \frac{\epsilon}{2},$$

$$\left\| \begin{bmatrix} AV_l & BV_l \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} \right\|_F \le \mu_1 + \frac{\epsilon}{2}.$$

, , 2. Let $\left[\begin{smallmatrix} P_1 & \hat{P}_1 \\ P_2 & \hat{P}_2 \end{smallmatrix}\right]$ be unitary. Since P_1 is nonsingular, we know by using the CSD [14] of $\left[\begin{smallmatrix} P_1 & \hat{P}_1 \\ P_2 & \hat{P}_2 \end{smallmatrix}\right]$ that $\hat{P}_2$ is also nonsingular. Then define

$$\Delta_0 := \hat{P}_2^{-1}, \quad \Lambda_0 = \Delta_0 \hat{P}_1, \quad B_0^{(1)} = \begin{bmatrix} AV_l & BV_l \end{bmatrix} \begin{bmatrix} \hat{P}_1 \\ \hat{P}_2 \end{bmatrix} \Delta_0^{-1}.$$

We have

$$\left\| \begin{bmatrix} AV_l & BV_l \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} \right\|_F = \| \begin{bmatrix} AV_l & BV_l \end{bmatrix} - B_0^{(1)} \begin{bmatrix} \Lambda_0 & I \end{bmatrix} \|_F.$$

– If all eigenvalues of Λ_0 are not distinct, similar to Argument 2, we compute $\Delta\Lambda_0$ such that all eigenvalues of $\Lambda_0 + \Delta\Lambda_0$ are distinct and $\|B_0^{(1)}\|_F \|\Delta\Lambda_0\|_F \le \frac{\epsilon}{2}$. Set

$$\Lambda_0 := \Lambda_0 + \Delta\Lambda_0.$$

Now, all eigenvalues of Λ_0, denoted by $\lambda_1, \ldots, \lambda_l$, are distinct and

$$\| \begin{bmatrix} AV_l & BV_l \end{bmatrix} - B_0^{(1)} \begin{bmatrix} \Lambda_0 & I \end{bmatrix} \|_F \le \mu_l + \epsilon.$$

, , 3. Let

$$V_0 := \begin{bmatrix} V_l & \mathcal{V}_l \end{bmatrix}$$

be unitary. Define

$$A_0 = \begin{bmatrix} B_0^{(1)}\Lambda_0 & A\mathcal{V}_l \end{bmatrix} V_0^H, \quad B_0 = \begin{bmatrix} B_0^{(1)} & B\mathcal{V}_l \end{bmatrix} V_0^H$$

if $\operatorname{rank}(B_0) < n$. Then similar to the proof of Theorem 4, we construct ΔA_0 and ΔB_0 such that (17) and (18) hold. Set

$$A_0 := A_0 + \Delta A_0, \quad B_0 := B_0 + \Delta B_0.$$

For the pencil (A_0, B_0) above, we have

$$\operatorname{rank}(B_0) = n, \quad A_0 \underline{v}_k = \lambda_k B_0 \underline{v}_k, \quad k = 1, \ldots, l,$$

and

$$\| \begin{bmatrix} A - A_0 & B - B_0 \end{bmatrix} \|_F = \| \begin{bmatrix} AV_l & BV_l \end{bmatrix} - B_0^{(1)} \begin{bmatrix} \Lambda_0 & I \end{bmatrix} \|_F \le \mu_l + 2\epsilon.$$

, , 2. Let's consider Example 1 again. For any $1 > \epsilon > 0$, by using the procedure above, we obtain an $O(\epsilon)$-optimal pencil

$$(A_0, B_0) = \left(\begin{bmatrix} -\frac{\epsilon}{40}\sqrt{1 - \frac{\epsilon^2}{16}} \\ 1 - \frac{\epsilon^2}{16} \end{bmatrix}, \begin{bmatrix} \frac{\epsilon^2}{160} \\ -\frac{\epsilon}{4}\sqrt{1 - \frac{\epsilon^2}{16}} \end{bmatrix} \right).$$

It can be verified that (A_0, B_0) has a finite eigenvalue and

$$\| \begin{bmatrix} A - A_0 & B - B_0 \end{bmatrix}_F \le 0.1 + 2\epsilon = \mu_1 + 2\epsilon.$$

3. Distance between controllable and uncontrollable descriptor systems. In this section we apply the idea in proving Theorem 1 to the distance problem

between controllable and uncontrollable descriptor systems, which is defined as follows.

PROBLEM 3.

$$E\dot{x} = Ax + Bu, \tag{21}$$

$E, A \in \mathbf{C}^{n\times n}$ $B \in \mathbf{C}^{n\times p}$ E (21) $(E;\ A,\ B)$

$$\operatorname{rank}\begin{bmatrix} \alpha A - \beta E & B \end{bmatrix} = n \quad \forall (\alpha,\beta) \in \mathbf{C}^2\backslash(0,0).$$

$$\mu_{(E;\ A,\ B)} = \inf\left\{ \| \begin{bmatrix} E_0 - E & A_0 - A & B_0 - B \end{bmatrix} \|_F : \begin{pmatrix} E_0, A_0 \in \mathbf{C}^{n\times n}, B_0 \in \mathbf{C}^{n\times p} \\ (E_0;\ A_0,\ B_0) \text{ is uncontrollable} \end{pmatrix} \right\}.$$

$\mu_{(E;\ A,\ B)}$ above is called the distance between controllable system (21) and the set of uncontrollable descriptor systems.

The distance between controllable and uncontrollable linear systems has been an interesting problem in the past two decades; see [2, 3, 5, 8, 12, 13, 15, 18]. When $E = I$ is supposed not to be perturbed, it was shown in [8] that

$$\begin{aligned} \mu_{(A,B)} &:= \inf\left\{ \| \begin{bmatrix} A - A_0 & B - B_0 \end{bmatrix} \|_F : \begin{pmatrix} A_0 \in \mathbf{C}^{n\times n},\ B_0 \in \mathbf{C}^{n\times p} \\ (I; A_0, B_0) \text{ is uncontrollable} \end{pmatrix} \right\} \\ &= \inf\{\sigma_n(\begin{bmatrix} \lambda I - A & B \end{bmatrix}) :\ \lambda \in \mathbf{C}\}. \end{aligned} \tag{22}$$

When E is singular, some upper and lower bounds for $\mu_{(E;\ A,\ B)}$ were given in [20, 21]. In this section we study $\mu_{(E;\ A,\ B)}$. Our purpose is to extend (22) to $\mu_{(E;\ A,\ B)}$.

LEMMA 5. $E_0, A_0 \in \mathbf{C}^{n\times n}$ $B_0 \in \mathbf{C}^{n\times p}$

(i) $(E_0; A_0, B_0)$

(ii) $(\alpha,\beta) \in \mathbf{C}^2\backslash(0,0)$ $\underline{v} \in \mathbf{C}^n$ $\underline{v} \neq 0$

$$\underline{v}^H \begin{bmatrix} \alpha A_0 - \beta E_0 & B_0 \end{bmatrix} = 0.$$

(iii)

$$V = \begin{bmatrix} \underline{v}^H \\ \mathcal{V}^H \end{bmatrix} \begin{matrix} \}1 \\ \}n-1 \end{matrix}$$

$(\alpha,\beta) \in \mathbf{C}^2\backslash(0,0)$

$$VE_0 = \begin{bmatrix} \underline{e}_0^H \\ \mathcal{E}_0^H \end{bmatrix} \begin{matrix} \}1 \\ \}n-1 \end{matrix}, \quad VA_0 = \begin{bmatrix} \underline{a}_0^H \\ \mathcal{A}_0^H \end{bmatrix} \begin{matrix} \}1 \\ \}n-1 \end{matrix}, \tag{23}$$

$$VB_0 = \begin{bmatrix} 0 \\ \mathcal{B}_0^H \end{bmatrix} \begin{matrix} \}1 \\ \}n-1 \end{matrix}, \quad \alpha\underline{a}_0^H = \beta\underline{e}_0^H.$$

. The proof is trivial and thus is omitted. □

We are ready now to present our main result in this section.

THEOREM 6. [illegible] (21) [illegible] $E, A \in \mathbf{C}^{n\times n}$ $B \in \mathbf{C}^{n\times p}$ [illegible]

$$\mu_{(E;\ A,\ B)} = \min\left\{\sigma_n(\begin{bmatrix} E & B \end{bmatrix}),\ \inf\left\{\sigma_n\left(\begin{bmatrix} \frac{A-\lambda E}{\sqrt{1+|\lambda|^2}} & B \end{bmatrix}\right) :\ \lambda \in \mathbf{C}\right\}\ \right\}. \tag{24}$$

[illegible] According to Lemma 5, the following relation holds:

$$\{(E_0, A_0, B_0) :\ E_0, A_0 \in \mathbf{C}^{n\times n},\ B_0 \in \mathbf{R}^{n\times p},\ (E_0; A_0, B_0) \text{ is uncontrollable}\} = \mathbf{S}_4 \cup \mathbf{S}_5 \tag{25}$$

with

$$\mathbf{S}_4 = \Bigg\{(E_0, A_0, B_0) :\ E_0, A_0 \in \mathbf{C}^{n\times n},\ B_0 \in \mathbf{R}^{n\times p},$$
$$VE_0 = \begin{bmatrix} 0 \\ \mathcal{E}_0^H \end{bmatrix} \begin{matrix} \}1 \\ \}n-1 \end{matrix},\ VA_0 = \begin{bmatrix} \underline{a}_0^H \\ \mathcal{A}_0^H \end{bmatrix} \begin{matrix} \}1 \\ \}n-1 \end{matrix},$$
$$VB_0 = \begin{bmatrix} 0 \\ \mathcal{B}_0^H \end{bmatrix} \begin{matrix} \}1 \\ \}n-1 \end{matrix},\ V = \begin{bmatrix} \underline{v}^H \\ \mathcal{V}^H \end{bmatrix} \begin{matrix} \}1 \\ \}n-1 \end{matrix} \text{ is unitary}\Bigg\},$$
$$\mathbf{S}_5 = \Bigg\{(E_0, A_0, B_0) :\ \ E_0, A_0 \in \mathbf{C}^{n\times n},\ B_0 \in \mathbf{R}^{n\times p},$$
$$VE_0 = \begin{bmatrix} \underline{e}_0^H \\ \mathcal{E}_0^H \end{bmatrix} \begin{matrix} \}1 \\ \}n-1 \end{matrix},\ VA_0 = \begin{bmatrix} \underline{a}_0^H \\ \mathcal{A}_0^H \end{bmatrix} \begin{matrix} \}1 \\ \}n-1 \end{matrix},$$
$$VB_0 = \begin{bmatrix} 0 \\ \mathcal{B}_0^H \end{bmatrix} \begin{matrix} \}1 \\ \}n-1 \end{matrix},\ V = \begin{bmatrix} \underline{v}^H \\ \mathcal{V}^H \end{bmatrix} \begin{matrix} \}1 \\ \}n-1 \end{matrix} \text{ is unitary},\ \underline{a}_0^H = \lambda \underline{e}_0^H,\ \lambda \in \mathbf{C}\Bigg\}.$$

As a direct result of (25) we have

$$\begin{aligned} &\mu_{(E;\ A,\ B)} \\ &= \inf\{\| \begin{bmatrix} E_0 - E & A_0 - A & B_0 - B \end{bmatrix} \|_F :\ (E_0, A_0, B_0) \in \mathbf{S}_4 \cup \mathbf{S}_5\} \\ &= \min\{\inf\{\| \begin{bmatrix} E_0 - E & A_0 - A & B_0 - B \end{bmatrix} \|_F :\ (E_0, A_0, B_0) \in \mathbf{S}_4\}, \\ &\qquad \inf\{\| \begin{bmatrix} E_0 - E & A_0 - A & B_0 - B \end{bmatrix} \|_F :\ (E_0, A_0, B_0) \in \mathbf{S}_5\}\ \}. \end{aligned} \tag{26}$$

Now we consider

$$\inf\{\| \begin{bmatrix} E_0 - E & A_0 - A & B_0 - B \end{bmatrix} \|_F :\ (E_0, A_0, B_0) \in \mathbf{S}_4\}$$

and

$$\inf\{\| \begin{bmatrix} E_0 - E & A_0 - A & B_0 - B \end{bmatrix} \|_F :\ (E_0, A_0, B_0) \in \mathbf{S}_5\}$$

separately.

First we have

$$\begin{aligned} &\inf\{\| \begin{bmatrix} E_0 - E & A_0 - A & B_0 - B \end{bmatrix} \|_F :\ (E_0, A_0, B_0) \in \mathbf{S}_4\} \\ &= \inf\Bigg\{\| \begin{bmatrix} VE_0 - VE & VA_0 - VA & VB_0 - VB \end{bmatrix} \|_F : \\ &\qquad \begin{pmatrix} (E_0, A_0, B_0) \in \mathbf{S}_4 \\ \text{with } V = \begin{bmatrix} \underline{v}^H \\ \mathcal{V}^H \end{bmatrix} \begin{matrix} \}1 \\ \}n-1 \end{matrix} \text{ unitary} \end{pmatrix}\Bigg\} \end{aligned}$$

$$= \inf \left\{ \left\| \begin{bmatrix} -\underline{v}^H E & \underline{a}_0^H - \underline{v}^H A & -\underline{v}^H B \\ \mathcal{E}_0^H - \mathcal{V}^H E & \mathcal{A}_0^H - \mathcal{V}^H A & \mathcal{B}_0^H - \mathcal{V}^H B \end{bmatrix} \right\|_F : \right.$$

$$\left. \left(\begin{array}{c} \mathcal{E}_0, \mathcal{A}_0 \in \mathbf{C}^{n\times(n-1)} \\ \mathcal{B}_0 \in \mathbf{C}^{p\times(n-1)},\ \underline{a}_0 \in \mathbf{C}^n \\ V = \begin{bmatrix} \underline{v}^H \\ \mathcal{V}^H \end{bmatrix} \begin{matrix} \}1 \\ \}n-1 \end{matrix} \text{ unitary} \end{array} \right) \right\}$$

$$= \inf\{\| \begin{bmatrix} -\underline{v}^H E & -\underline{v}^H B \end{bmatrix} \|_F :\ \underline{v} \in \mathbf{C}^n,\ \|\underline{v}\|_2 = 1\}$$

$$\left(\text{by taking } \begin{cases} \mathcal{E}_0^H = \mathcal{V}^H E \\ \mathcal{A}_0^H = \mathcal{V}^H A \\ \mathcal{B}_0^H = \mathcal{V}^H B \\ \underline{a}_0^H = \underline{v}^H A \end{cases} \right)$$

$$= \inf\{\|\underline{v}^H \begin{bmatrix} E & B \end{bmatrix} \|_F :\ \underline{v} \in \mathbf{C}^n,\ \|\underline{v}\|_2 = 1\}$$

$$= \sigma_n(\begin{bmatrix} E & B \end{bmatrix}). \tag{27}$$

Next, we know

$$\inf\{\| \begin{bmatrix} E_0 - E & A_0 - A & B_0 - B \end{bmatrix} \|_F^2 :\ (E_0, A_0, B_0) \in \mathbf{S}_5\}$$

$$= \inf \left\{ \| \begin{bmatrix} VE_0 - VE & VA_0 - VA & VB_0 - VB \end{bmatrix} \|_F^2 : \right.$$

$$\left. \left(\begin{array}{c} (E_0, A_0, B_0) \in \mathbf{S}_5 \\ \text{with } V = \begin{bmatrix} \underline{v}^H \\ \mathcal{V}^H \end{bmatrix} \begin{matrix} \}1 \\ \}n-1 \end{matrix} \text{ unitary} \end{array} \right) \right\}$$

$$= \inf \left\{ \left\| \begin{bmatrix} \underline{e}_0^H - \underline{v}^H E & \underline{a}_0^H - \underline{v}^H A & -\underline{v}^H B \\ \mathcal{E}_0^H - \mathcal{V}^H E & \mathcal{A}_0^H - \mathcal{V}^H A & \mathcal{B}_0^H - \mathcal{V}^H B \end{bmatrix} \right\|_F^2 : \right.$$

$$\left. \left(\begin{array}{c} \mathcal{E}_0, \mathcal{A}_0 \in \mathbf{C}^{n\times(n-1)},\ \mathcal{B}_0 \in \mathbf{C}^{p\times(n-1)} \\ \underline{e}_0, \underline{a}_0 \in \mathbf{C}^n,\ \underline{a}_0^H = \lambda \underline{e}_0^H,\ \lambda \in \mathbf{C} \\ V = \begin{bmatrix} \underline{v}^H \\ \mathcal{V}^H \end{bmatrix} \begin{matrix} \}1 \\ \}n-1 \end{matrix} \text{ unitary} \end{array} \right) \right\}$$

$$= \inf \left\{ \left\| \begin{bmatrix} \underline{e}_0^H - \underline{v}^H E & \lambda \underline{e}_0^H - \underline{v}^H A & -\underline{v}^H B \\ \mathcal{E}_0^H - \mathcal{V}^H E & \mathcal{A}_0^H - \mathcal{V}^H A & \mathcal{B}_0^H - \mathcal{V}^H B \end{bmatrix} \right\|_F^2 : \right.$$

$$\left. \left(\begin{array}{c} \mathcal{E}_0, \mathcal{A}_0 \in \mathbf{C}^{n\times(n-1)},\ \mathcal{B}_0 \in \mathbf{C}^{p\times(n-1)}, \\ \underline{e}_0 \in \mathbf{C}^n, \lambda \in \mathbf{C} \\ V = \begin{bmatrix} \underline{v}^H \\ \mathcal{V}^H \end{bmatrix} \begin{matrix} \}1 \\ \}n-1 \end{matrix} \text{ unitary} \end{array} \right) \right\}$$

$$= \inf \left\{ \| \begin{bmatrix} \underline{e}_0^H - \underline{v}^H E & \lambda \underline{e}_0^H - \underline{v}^H A & \underline{v}^H B \end{bmatrix} \|_F^2 :\ \left(\begin{array}{c} \underline{e}_0 \in \mathbf{C}^n, \lambda \in \mathbf{C} \\ \underline{v} \in \mathbf{C}^n,\ \|\underline{v}\|_2 = 1 \end{array} \right) \right\} \tag{28}$$

$$\left(\text{by taking } \begin{cases} \mathcal{E}_0^H = \mathcal{V}^H E \\ \mathcal{A}_0^H = \mathcal{V}^H A \\ \mathcal{B}_0^H = \mathcal{V}^H B \end{cases} \right).$$

Note that

$$\mathcal{Q}(\lambda) := \frac{1}{\sqrt{1+|\lambda|^2}} \begin{bmatrix} I & \lambda I \\ \bar{\lambda} I & -I \end{bmatrix} \in \mathbf{C}^{2n\times 2n}$$

is unitary for any $\lambda \in \mathbf{C}$. Here $\bar{\lambda}$ is the complex conjugate of λ. Thus,

$$\begin{aligned}
&\inf\{\| \begin{bmatrix} E_0 - E & A_0 - A & B_0 - B \end{bmatrix} \|_F^2 : \ (E_0, A_0, B_0) \in \mathbf{S}_5\} \\
&= \inf \left\{ \left\| \begin{bmatrix} \underline{e}_0^H - \underline{v}^H E & \lambda \underline{e}_0^H - \underline{v}^H A & \underline{v}^H B \end{bmatrix} \begin{bmatrix} \mathcal{Q}(\lambda) & 0 \\ 0 & I \end{bmatrix} \right\|_F^2 : \right. \\
&\qquad \left. \begin{pmatrix} \underline{e}_0 \in \mathbf{C}^n, \ \lambda \in \mathbf{C} \\ \underline{v} \in \mathbf{C}^n, \ \|\underline{v}\|_2 = 1 \end{pmatrix} \right\} \\
&= \inf \left\{ \left\| \begin{bmatrix} \frac{(1+|\lambda|^2)\underline{e}_0^H - (\underline{v}^H E + \bar{\lambda}\underline{v}^H A)}{\sqrt{1+|\lambda|^2}} & \frac{\underline{v}^H A - \lambda \underline{v}^H E}{\sqrt{1+|\lambda|^2}} & \underline{v}^H B \end{bmatrix} \right\|_F^2 : \right. \\
&\qquad \left. \begin{pmatrix} \underline{e}_0 \in \mathbf{C}^n, \ \lambda \in \mathbf{C} \\ \underline{v} \in \mathbf{C}^n, \ \|\underline{v}\|_2 = 1 \end{pmatrix} \right\} \\
&= \inf \left\{ \left\| \begin{bmatrix} \frac{\underline{v}^H A - \lambda \underline{v}^H E}{\sqrt{1+|\lambda|^2}} & \underline{v}^H B \end{bmatrix} \right\|_F^2 : \ \begin{pmatrix} \lambda \in \mathbf{C} \\ \underline{v} \in \mathbf{C}^n, \ \|\underline{v}\|_2 = 1 \end{pmatrix} \right\} \\
&\quad \left(\text{by taking } e_0^H = \frac{\underline{v}^H E + \bar{\lambda}\underline{v}^H A}{1+|\lambda|^2} \right) \\
&= \inf \left\{ \left\| \underline{v}^H \begin{bmatrix} \frac{A - \lambda E}{\sqrt{1+|\lambda|^2}} & B \end{bmatrix} \right\|_F^2 : \ \begin{pmatrix} \lambda \in \mathbf{C} \\ \underline{v} \in \mathbf{C}^n, \ \|\underline{v}\|_2 = 1 \end{pmatrix} \right\} \\
&= \inf \left\{ \sigma_n^2 \left(\begin{bmatrix} \frac{A - \lambda E}{\sqrt{1+|\lambda|^2}} & B \end{bmatrix} \right) : \ \lambda \in \mathbf{C} \right\},
\end{aligned}$$

or equivalently

$$\begin{aligned}
&\inf\{\| \begin{bmatrix} E_0 - E & A_0 - A & B_0 - B \end{bmatrix} \|_F : \ (E_0, A_0, B_0) \in \mathbf{S}_5\} \\
&= \inf \left\{ \sigma_n \left(\begin{bmatrix} \frac{A - \lambda E}{\sqrt{1+|\lambda|^2}} & B \end{bmatrix} \right) : \ \lambda \in \mathbf{C} \right\}.
\end{aligned} \tag{29}$$

Therefore, (24) follows directly from (26), (27), and (29). ∎

4. Conclusions. In this paper we have considered the nonsquare generalized eigenvalue problem and obtained algebraic characterizations for Problems 1 and 3 in Theorems 1 and 6, respectively. The numerical algorithms for solving (1) are still under investigation. Since

$$\inf \left\{ \sigma_n \left(\begin{bmatrix} \frac{A - \lambda E}{\sqrt{1+|\lambda|^2}} & B \end{bmatrix} \right) : \lambda \in \mathbf{C} \right\}$$

is quite similar to $\inf\{\sigma_n(\begin{bmatrix} A - \lambda I & B \end{bmatrix}) : \lambda \in \mathbf{C}\}$, some existing algorithms in [5, 12, 15, 18] for computing $\mu_{(A,B)} = \inf\{\sigma_n(\begin{bmatrix} A - \lambda I & B \end{bmatrix}) : \ \lambda \in \mathbf{C}\}$ can be

extended for computing $\mu_{(E;A,B)}$ based on the formula (24). However, because the factor $\sqrt{1+|\lambda|^2}$ is in the denominator of

$$\inf\left\{\sigma_n\left(\begin{bmatrix} \frac{A-\lambda E}{\sqrt{1+|\lambda|^2}} & B \end{bmatrix}\right) : \lambda \in \mathbf{C}\right\},$$

such extensions are not trivial, and some numerical investigations should be done. We leave this topic to interested readers.

REFERENCES

[1] D. Boley, *Estimating the sensitivity of the algebraic structure of pencils with simple eigenvalue estimates*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 632–643.

[2] D. Boley, *A perturbation result for linear control problems*, SIAM J. Algebraic Discrete Methods, 6 (1985), pp. 66–72.

[3] D. Boley and L. Wu-Sheng, *Measuring how far a controllable system is from an uncontrollable one*, IEEE Trans. Automat. Control, 31 (1986), pp. 249–251.

[4] G. Boutry, M. Elad, G. H. Golub, and P. Milanfar, *The generalized eigenvalue problem for nonsquare pencils using a minimal perturbation approach*, SIAM J. Matrix Anal. Appl., 27 (2005), pp. 582–601.

[5] R. Byers, *Detecting nearly uncontrollable pairs*, in Signal Processing, Scattering and Operator Theory, and Numerical Methods (Amsterdam, 1989), Progr. Systems Control Theory 5, Birkhäuser Boston, Boston, MA, 1990, pp. 447–457.

[6] J. W. Demmel and B. Kågström, *Computing stable eigendecompositions of matrix pencils*, Linear Algebra Appl., 88/89 (1987), pp. 139–186.

[7] 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.

[8] R. Eising, *Between controllable and uncontrollable*, Systems Control Lett., 4 (1985), pp. 263–264.

[9] M. Elad, P. Milanfar, and G. H. Golub, *Shape from moments-an estimation theory perspective*, IEEE Trans. Signal Process., 52 (2004), pp. 1814–1829.

[10] E. Elmroth, P. Johansson, and B. Kågström, *Computation and presentation of graphs displaying closure hierarchies of Jordan and Kronecker structures*, Numer. Linear Algebra Appl., 8 (2001), pp. 381–399.

[11] E. Elmroth, P. Johansson, and B. Kågström, *Bounds for the distance between nearly Jordan and Kronecker structure in a closure hierarchy*, Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov (POMI), 268 (2000), pp. 24–48.

[12] L. Elsner and C. He, *An algorithm for computing the distance to uncontrollability*, Systems Control Lett., 17 (1991), pp. 453–464.

[13] P. Gahinet and A. J. Laub, *Algebraic Riccati equations and the distance to the nearest uncontrollable pair*, SIAM J. Control Optim., 30 (1992), pp. 765–786.

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

[15] C. He, *Estimating the distance to uncontrollability: A fast method and a slow one*, Systems Control Lett., 26 (1995), pp. 275–281.

[16] G. W. Stewart, *Perturbation theory for rectangular matrix pencils*, Linear Algebra Appl., 208/209 (1994), pp. 297–301.

[17] L. N. Trefethen, *Computation of pseudospectra*, Acta Numerica, Cambridge University Press, Cambridge, UK, 1999, pp. 247–295.

[18] M. Wicks and R. A. DeCarlo, *Computing the distance to an uncontrollable system*, IEEE Trans. Automat. Control, 36 (1991), pp. 39–49.

[19] T. G. Wright and L. N. Trefethen, *Pseudospectra of rectangular matrices*, IMA J. Numer. Anal., 22 (2002), pp. 501–519.

[20] Y. Zou and C. Yang, *Formulae for the distance between controllable and uncontrollable linear systems*, Systems Control Lett., 21 (1993), pp. 173–180.

[21] Y. Zou and C. Yang, *The distance between controllable and uncontrollable singular systems*, Acta Automat. Sinica, 17 (1991), pp. 220–224 (in Chinese).

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 3, pp. 788–804

A SCHUR–NEWTON METHOD FOR THE MATRIX pTH ROOT AND ITS INVERSE*

CHUN-HUA GUO† AND NICHOLAS J. HIGHAM‡

Abstract. Newton's method for the inverse matrix pth root, $A^{-1/p}$, has the attraction that it involves only matrix multiplication. We show that if the starting matrix is $c^{-1}I$ for $c \in \mathbb{R}^+$ then the iteration converges quadratically to $A^{-1/p}$ if the eigenvalues of A lie in a wedge-shaped convex set containing the disc $\{z : |z - c^p| < c^p\}$. We derive an optimal choice of c for the case where A has real, positive eigenvalues. An application is described to roots of transition matrices from Markov models, in which for certain problems the convergence condition is satisfied with $c = 1$. Although the basic Newton iteration is numerically unstable, a coupled version is stable and a simple modification of it provides a new coupled iteration for the matrix pth root. For general matrices we develop a hybrid algorithm that computes a Schur decomposition, takes square roots of the upper (quasi-) triangular factor, and applies the coupled Newton iteration to a matrix for which fast convergence is guaranteed. The new algorithm can be used to compute either $A^{1/p}$ or $A^{-1/p}$, and for large p that are not highly composite it is more efficient than the method of Smith based entirely on the Schur decomposition.

Key words. matrix pth root, principal pth root, matrix logarithm, inverse, Newton's method, preprocessing, Schur decomposition, numerical stability, convergence, Markov model, transition matrix

AMS subject classifications. 65F30, 15A18, 15A51

DOI. 10.1137/050643374

1. Introduction.

Newton methods for computing the principal matrix square root have been studied for almost fifty years and are now well understood. Since Laasonen proved convergence but observed numerical instability [25], several Newton variants have been derived and proved numerically stable, for example by Higham [13], [15], Iannazzo [18], and Meini [27]. For matrix pth roots, with p an integer greater than 2, Newton methods were until recently little used, for two reasons: their convergence in the presence of complex eigenvalues was not well understood and the iterations were found to have poor numerical stability. The subtlety of the question of convergence is clear from the scalar case, since the starting values for which Newton's method for $z^p - 1 = 0$ converges to some pth root of unity form fractal Julia sets in the complex plane for $p > 2$ [28], [30], [33]. Nevertheless, Iannazzo [19] has recently proved a new convergence result for the scalar Newton iteration and has thereby shown how to build a practical algorithm for the matrix pth root.

Throughout this work we assume that $A \in \mathbb{C}^{n\times n}$ has no eigenvalues on $\mathbb{R}^-$, the closed negative real axis. The particular pth root of interest is the principal pth root (and its inverse), denoted by $A^{1/p}$ ($A^{-1/p}$), which is the unique matrix X such that $X^p = A$ ($X^{-p} = A$) and the eigenvalues of X lie in the segment $\{ z : -\pi/p <$

*Received by the editors October 24, 2005; accepted for publication (in revised form) by A. Frommer March 6, 2006; published electronically October 4, 2006. This work was supported in part by a Royal Society-Wolfson Research Merit Award to the second author.
http://www.siam.org/journals/simax/28-3/64337.html
†Department of Mathematics and Statistics, University of Regina, Regina, SK S4S 0A2, Canada (chguo@math.uregina.ca, http://www.math.uregina.ca/~chguo/). The work of this author was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada.
‡School of Mathematics, The University of Manchester, Sackville Street, Manchester, M60 1QD, UK (higham@ma.man.ac.uk, http://www.ma.man.ac.uk/~higham/).

$\arg(z) < \pi/p\}$. We are interested in methods both for computing $A^{1/p}$ and for computing $A^{-1/p}$.

We briefly summarize Iannazzo's contribution, which concerns Newton's method for $X^p - A = 0$, and then turn to the inverse Newton iteration. Newton's method takes the form

$$X_{k+1} = \frac{1}{p}\big[(p-1)X_k + X_k^{1-p}A\big], \qquad X_0A = AX_0. \tag{1.1}$$

Iannazzo [19] shows that $X_k \to A^{1/p}$ quadratically if $X_0 = I$ and each eigenvalue of A belongs to the set

$$S = \{\, z \in \mathbb{C} : \operatorname{Re} z > 0 \text{ and } |z| \le 1 \,\} \cup \mathbb{R}^+, \tag{1.2}$$

where $\mathbb{R}^+$ denotes the open positive real axis. Based on this result, he obtains the following algorithm for computing the principal pth root.

ALGORITHM 1.1 (matrix pth root via Newton iteration [19]). [illegible] $A \in \mathbb{C}^{n\times n}$ [illegible] $\mathbb{R}^-$ [illegible] $X = A^{1/p}$ [illegible]

1 $B = A^{1/2}$
2 $C = B/\|B\|$ (any norm)
3 Use the iteration (1.3) to compute $X = C^{2/p}$ (p even) or $X = \big(C^{1/p}\big)^2$ (p odd).
4 $X \leftarrow \|B\|^{2/p} X$

The iteration used in the algorithm is a rewritten version of (1.1):

$$\begin{aligned} X_{k+1} &= X_k\left(\frac{(p-1)I + M_k}{p}\right), & X_0 &= I, \\ M_{k+1} &= \left(\frac{(p-1)I + M_k}{p}\right)^{-p} M_k, & M_0 &= A, \end{aligned} \tag{1.3}$$

where $M_k \equiv X_k^{-p}A$. Iannazzo shows that, unlike (1.1), this coupled form is numerically stable.

Newton's method can also be applied to $X^{-p} - A = 0$, for which it takes the form

$$X_{k+1} = \frac{1}{p}\big[(p+1)X_k - X_k^{p+1}A\big], \qquad X_0A = AX_0. \tag{1.4}$$

The iteration has been studied by several authors. R. A. Smith [34] uses infinite product expansions to show that X_k converges to an inverse pth root of A if the initial matrix X_0 satisfies $\rho(I - X_0^pA) < 1$, where ρ denotes the spectral radius. Lakić [26] reaches the same conclusion, under the assumption that A is diagonalizable, for a family of iterations that includes (1.4). Bini,[1] Higham, and Meini take $X_0 = I$ and prove convergence of the residuals $I - X_k^pA$ to zero when $\rho(I - A) < 1$ (see Lemma 2.1 below) as well as convergence of X_k to $A^{-1/p}$ if A has real, positive eigenvalues and $\rho(A) < p + 1$ [4]. They also show that (1.4) has poor numerical stability properties. In none of these papers is it proved to which inverse pth root the iteration converges when $\rho(I - X_0^pA) < 1$. The purpose of our work is to determine a larger region of convergence to $A^{-1/p}$ for (1.4) and to build a numerically stable algorithm applicable to arbitrary A having no eigenvalues on $\mathbb{R}^-$.

[1] The authors of [4] were unaware of the papers of Lakić [26] and R. A. Smith [34], and Lakić appears to have been unaware of Smith's paper.

In section 2 we present convergence analysis to show that if the spectrum of A is contained in a certain wedge-shaped convex region depending on a parameter $c \in \mathbb{R}^+$ then quadratic convergence of the inverse Newton method with $X_0 = c^{-1}I$ to $A^{-1/p}$ is guaranteed—with no restrictions on the Jordan structure of A. In section 3 we consider the practicalities of choosing c and implementing the inverse Newton iteration. We derive an optimal choice of c for the case where A has real, positive eigenvalues, and we prove a finite termination property for a matrix with just one distinct eigenvalue. A stable coupled version of (1.4) is noted, and by a simple modification a new iteration is obtained for $A^{1/p}$. For general A we propose a hybrid algorithm for computing $A^{-1/p}$ or $A^{1/p}$ that precedes application of the Newton iteration with a preprocessing step, in which a Schur reduction to triangular form is followed by the computation of a sequence of square roots. An interesting and relatively unexplored application of pth roots is to Markov models; in section 4 we discuss this application and show that convergence of the inverse Newton iteration is ensured with $c = 1$ in certain cases. Numerical experiments are presented in section 5, wherein we derive a particular scaling of the residual that is appropriate for testing numerical stability. Section 6 presents our conclusions.

Finally, we mention some other reasons for our interest in computing the inverse matrix pth root. The pth root arises in the computation of the matrix logarithm by the inverse scaling and squaring method. This method uses the relation $\log(A) = p \log A^{1/p}$, where p is typically a power of 2, and approximates $\log A^{1/p}$ using a Padé approximant [6], [22, App. A]. Since $\log(A) = -p \log A^{-1/p}$, the inverse pth root can equally well be employed. The inverse pth root also appears in the matrix sector function, defined by $\mathrm{sect}_p(A) = A(A^p)^{-1/p}$ (of which the matrix sign function is the special case with $p = 2$) [23], [31], and in the expression $A(A^*A)^{-1/2}$ for the unitary polar factor of a matrix [12], [29]. For scalars $a \in \mathbb{R}$ the inverse Newton iteration is employed in floating point hardware to compute the square root $a^{1/2}$ via $a^{-1/2} \times a$, since the whole computation can be done using only multiplications [7], [21]. The inverse Newton iteration is also used to compute $a^{1/p}$ in arbitrarily high precision in the MPFUN and ARPREC packages [1], [2], [3]. Our work will be useful for computing [illegible] pth roots in high precision—a capability currently lacking in MATLAB's Symbolic Math Toolbox (Release 14, Service Pack 3).

2. Convergence to the inverse principal pth root. We begin by recalling a result of Bini, Higham, and Meini [4, Prop. 6.1].

LEMMA 2.1. [illegible] $R_k = I - X_k^p A$ [illegible] (1.4) [illegible]

$$R_{k+1} = \sum_{i=2}^{p+1} a_i R_k^i, \tag{2.1}$$

[illegible] a_i [illegible] $\sum_{i=2}^{p+1} a_i = 1$ [illegible] $0 < \|R_0\| < 1$ [illegible] $\|R_k\|$ [illegible] 0 [illegible] $k \to \infty$ [illegible] $\|R_{k+1}\| < \|R_k\|^2$

In the scalar case, Lemma 2.1 implies the convergence of (1.4) to an inverse pth root when $\|R_0\| < 1$, and we will use this fact below; the limit is not necessarily the inverse principal pth root, however. R. A. Smith [34] shows likewise that $\|R_0\| < 1$ implies convergence to an inverse pth root for matrices. Note that the convergence of X_k in the matrix case does not follow immediately from the convergence of R_k in Lemma 2.1. Indeed, when $\|R_0\| < 1$, the sequence of pth powers, $\{X_k^p\}$, is bounded

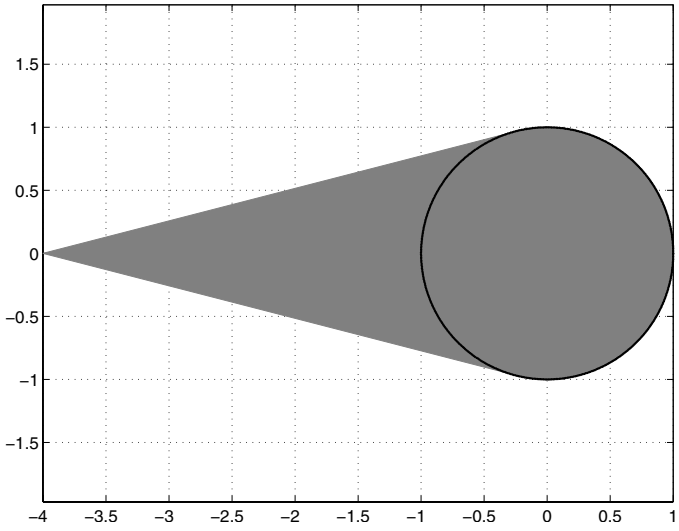


FIG. 2.1. *The region E for $p = 4$. The solid line marks the disk of radius 1, center 0, whose interior is D.*

since $X_k^p = (I - R_k)A^{-1}$, but the boundedness of $\{X_k\}$ itself does not follow when $n > 1$.

Our aim in this section is to show that for an appropriate range of X_0 the Newton iterates X_k converge to $A^{-1/p}$. We begin with the scalar case. Thus, for a given $\lambda \in \mathbb{C} \setminus \mathbb{R}^-$ we wish to determine for which $x_0 \in \mathbb{C}$ the iteration

$$x_{k+1} = \frac{1}{p}\left[(p+1)x_k - x_k^{p+1}\lambda\right] \tag{2.2}$$

yields $\lambda^{-1/p}$, the principal inverse pth root of λ, which we know lies in the segment

$$\{\, z : -\pi/p < \arg(z) < \pi/p \,\}. \tag{2.3}$$

We denote by $D = \{\, z : |z| < 1 \,\}$ the open unit disc and by $\overline{D}$ its closure. Let

$$E = \operatorname{conv}\{\overline{D}, -p\} \setminus \{-p, 1\},$$

where conv denotes the convex hull. Figure 2.1 depicts E for $p = 4$. The next result is a restatement of [34, Thm. 4].

LEMMA 2.2. [illegible] (2.2) [illegible] $1 - x_0^p\lambda \in E$ [illegible] $1 - x_1^p\lambda \in D$

The following result generalizes the scalar version of [4, Prop. 6.2] from $x_0 = 1$ to $x_0 > 0$ and the proof is essentially the same.

LEMMA 2.3. [illegible] $\lambda \in \mathbb{R}^+$ [illegible] $x_0 \in \mathbb{R}^+$ [illegible] $1 - x_0^p\lambda \in (-p, 1)$ [illegible] $\{x_k\}$ [illegible] (2.2) [illegible] $\lambda^{-1/p}$

We will also need the following complex mean value theorem from [9]. We denote by $\operatorname{Re}(z)$ and $\operatorname{Im}(z)$ the real and imaginary parts of $z \in \mathbb{C}$ and define the line

$$L(a, b) = \{\, a + t(b - a) : t \in (0, 1) \,\}.$$

LEMMA 2.4. [illegible] Ω [illegible] $\mathbb{C}$ [illegible] $f : \Omega \mapsto \mathbb{C}$ [illegible] a b [illegible] Ω [illegible] u v [illegible] $L(a, b)$ [illegible]

$$\mathrm{Re}\left(\frac{f(b)-f(a)}{b-a}\right)=\mathrm{Re}(f'(u)),\quad \mathrm{Im}\left(\frac{f(b)-f(a)}{b-a}\right)=\mathrm{Im}(f'(v)).$$

The next result improves Lemma 2.3 by extending the region of allowed $1-x_0^p\lambda$ from the interval $(-p,1)$ to the convex set E in the complex plane.

LEMMA 2.5. [illegible] $\lambda\in\mathbb{C}\setminus\mathbb{R}^-$ [illegible] $x_0\in\mathbb{R}^+$ [illegible] $1-x_0^p\lambda\in E$ [illegible] x_k [illegible] (2.2) [illegible] $\lambda^{-1/p}$ [illegible]. By Lemma 2.2 we have $1-x_1^p\lambda\in D$. It then follows from the scalar version of Lemma 2.1 that x_k converges quadratically to $x(\lambda)$, an inverse pth root of λ (see the discussion after Lemma 2.1). We need to show that $x(\lambda)=\lambda^{-1/p}$. There is nothing to prove for $p=1$, so we assume $p\ge 2$.

For any $\lambda\in\mathbb{R}^+$ with $1-x_0^p\lambda\in(-p,1)$ we know from Lemma 2.3 that $x(\lambda)=\lambda^{-1/p}$. Intuition suggests that $x(\lambda)$ is a continuous function of λ. Since the principal segment (2.3) is disjoint from the other $p-1$ segments it then follows that for each λ with $1-x_0^p\lambda\in E$, $x(\lambda)$ must be the inverse of the principal pth root. We now provide a rigorous proof of $x(\lambda)=\lambda^{-1/p}$. (Once this is proved, the continuity of $x(\lambda)$ as a function of λ follows.)

We write $x_0=1/c$. Then $1-x_0^p\lambda\in(-p,1)$ becomes $\lambda\in(0,(p+1)c^p)$, and $1-x_0^p\lambda\in E$ is the same as $\lambda\in E_c$, where

$$E_c=\mathrm{conv}\big\{\,\{\,z:|z-c^p|\le c^p\,\},(p+1)c^p\,\big\}\setminus\{\,0,(p+1)c^p\,\}.$$

We rewrite E_c in polar form:

$$E_c=\{\,(r,\theta):0<r<(p+1)c^p,\ -\theta_r\le\theta\le\theta_r\,\},$$

where the exact expression for $\theta_r\equiv\theta(r)$ is unimportant. We fix $\delta\in(0,1)$ and define the compact set

$$E_{c,\delta}=\{\,(r,\theta):\delta c^p\le r\le(p+1-\delta)c^p,\ -\theta_r\le\theta\le\theta_r\,\}.$$

We will prove that $x(\lambda)$ is in the segment (2.3) for each $\lambda\in E_{c,\delta}$. This will yield $x(\lambda)=\lambda^{-1/p}$ for $\lambda\in E_c$, since δ can be arbitrarily small. More precisely, for each fixed $r\in[\delta c^p,(p+1-\delta)c^p]$, we will show that $x(\lambda)$ is in the same segment for each λ on the arc given in polar form by

$$\Gamma_r=\{\,(r,\theta):-\theta_r\le\theta\le\theta_r\,\}.$$

This will complete the proof, since we already know that $x(\lambda)$ is in the segment (2.3) when $\theta=0$. Thus we only need to show that there exists $\epsilon>0$ such that for all $a,b\in\Gamma_r$ with $|a-b|<\epsilon$, $x(a)$ and $x(b)$ are in the same segment. To do so, we suppose that for all $\epsilon>0$ there exist $a,b\in\Gamma_r$ with $|a-b|<\epsilon$ such that $x(a)$ is in segment i and $x(b)$ is in segment $j\ne i$, and we will obtain a contradiction.

Let a and b be any such pair for a suitably small ϵ to be chosen below. Let $\widetilde{x}(b)$ be the inverse pth root of b in segment i. Then $|x(b)-\widetilde{x}(b)|$ is at least the distance between two neighboring inverse pth roots of b, i.e.,

$$|x(b)-\widetilde{x}(b)|\ge 2r^{-1/p}\sin\frac{\pi}{p}=:4\eta.$$

Also, we have, by Lemma 2.4,

$$|x(a) - \widetilde{x}(b)| \le \sqrt{2} \sup_{\xi \in L(a,b)} \left| -\frac{1}{p} \xi^{-1/p-1} \right| |a-b| \le \frac{\sqrt{2}}{p} \left(\frac{r}{2}\right)^{-1/p-1} |a-b|$$

when $|a-b| \le \sqrt{3}r$. Therefore

$$|x(a) - \widetilde{x}(b)| \le \eta$$

when $|a-b| \le \min\{\sqrt{3}r, \frac{p}{\sqrt{2}}(\frac{r}{2})^{1/p+1}\eta\} =: \epsilon_1$.

For every $\lambda \in E_{c,\delta} \subset E_c$, we have $1 - x_0^p \lambda \in E$. Thus $1 - x_1^p \lambda \in D$ by Lemma 2.2. Since $E_{c,\delta}$ is compact, the set $\{\, 1 - x_1^p \lambda : \lambda \in E_{c,\delta} \,\}$ is a compact subset of D. Therefore there is constant $\delta_1 \in (0,1)$, independent of λ, such that $|1 - x_1^p \lambda| \le 1 - \delta_1$.

Now, for the iteration (2.2) with $\lambda \in \Gamma_r$, Lemma 2.1 implies

$$|1 - x_k^p \lambda| \le |1 - x_1^p \lambda|^{2^{k-1}} \le (1-\delta_1)^{2^{k-1}}$$

for $k \ge 1$. So

$$|(x_k - r_1)(x_k - r_2) \cdots (x_k - r_p)| = |x_k^p - \lambda^{-1}| \le \frac{1}{r}(1-\delta_1)^{2^{k-1}},$$

where $r_1, r_2, \ldots, r_p$ are the pth roots of λ^{-1}. Let

$$|x_k - r_s| = \min_{1 \le j \le p} |x_k - x_j|.$$

Then

$$|x_k - r_s| \le r^{-1/p}(1-\delta_1)^{2^{k-1}/p} =: \eta_1.$$

The iteration (2.2) is given by $x_{k+1} = g(x_k)$, where

$$g(x) = \frac{1}{p}\left[(p+1)x - x^{p+1}\lambda\right].$$

Note that for all x with $|x - r_s| \le \eta_1$,

$$|x - r_j| \le |r_s| + |r_j| + \eta_1 = 2r^{-1/p} + \eta_1, \qquad j \ne s,$$

and

$$\begin{aligned} |g'(x)| &= \frac{p+1}{p}|1 - x^p\lambda| = \frac{p+1}{p} r |(x-r_1)(x-r_2)\cdots(x-r_p)| \\ &\le \frac{p+1}{p} r\, \eta_1 (2r^{-1/p} + \eta_1)^{p-1}. \end{aligned}$$

We now take a sufficiently large k, independent of λ, such that $\eta_1 \le \eta$ and $\frac{p+1}{p} r \eta_1 (2r^{-1/p} + \eta_1)^{p-1} \le \frac{1}{2}$. Then, by Lemma 2.4,

$$|x_{k+1} - r_s| = |g(x_k) - g(r_s)| \le \frac{\sqrt{2}}{2}|x_k - r_s|$$

and hence $|x_{k+m} - r_s| \le (\frac{\sqrt{2}}{2})^m |x_k - r_s|$ for all $m \ge 0$. Thus $x_i \to r_s$ as $i \to \infty$ and $|x_k - r_s| \le \eta_1 \le \eta$. It follows that $r_s = x(\lambda)$ and $|x_k(\lambda) - x(\lambda)| \le \eta$, where we write $x_k(\lambda)$ for x_k to indicate its dependence on λ. In particular, we have

$$|x_k(a) - x(a)| \le \eta, \quad |x_k(b) - x(b)| \le \eta.$$

Now

$$\begin{aligned}|x_k(a) - x_k(b)| &= |(x_k(a) - x(a)) + (x(a) - \widetilde{x}(b)) + (\widetilde{x}(b) - x(b)) + (x(b) - x_k(b))| \\ &\geq |\widetilde{x}(b) - x(b)| - |x_k(a) - x(a)| - |x(a) - \widetilde{x}(b)| - |x(b) - x_k(b)| \\ &\geq 4\eta - \eta - \eta - \eta = \eta.\end{aligned}$$

On the other hand, for the chosen k, $x_k(\lambda)$ is a continuous function of λ on the compact set Γ_r and is therefore uniformly continuous on Γ_r. Thus there exists $\epsilon \in (0, \epsilon_1)$ such that for all $a, b \in \Gamma_r$ with $|a - b| < \epsilon$, $|x_k(a) - x_k(b)| < \eta$. This is a contradiction since we have just shown that for any $\epsilon \in (0, \epsilon_1)$, $|x_k(a) - x_k(b)| \geq \eta$ for some $a, b \in \Gamma_r$ with $|a - b| < \epsilon$. Our earlier assumption is therefore false, and the proof is complete. ∎

We are now ready to prove the convergence of (1.4) in the matrix case. The iterations (1.4) and (2.2) have the form $X_{k+1} = g(X_k, A)$ and $x_{k+1} = g(x_k, \lambda)$, respectively, where $g(x, t)$ is a polynomial in two variables. We will need the following special case of Theorem 4.16 in [11].

LEMMA 2.6. $g(x,t)$ $x_{k+1} = g(x_k, \lambda)$ $f(\lambda)$ λ x_0 $X_{k+1} = g(X_k, J(\lambda))$ $X_0 = x_0 I$ $J(\lambda)$ X_* $\operatorname{diag}(X_*) = \operatorname{diag}(f(J(\lambda)))$

We now apply Lemmas 2.5 and 2.6 with $x_0 = 1/c$ and $f(\lambda) = \lambda^{-1/p}$, where $c > 0$ is a constant.

THEOREM 2.7. $A \in \mathbb{C}^{n\times n}$ $\mathbb{R}^-$ $p \geq 1$ X_k (1.4) $X_0 = \frac{1}{c} I$ $c \in \mathbb{R}^+$ $A^{-1/p}$ A

$$E(c,p) = \operatorname{conv}\{\, \{\, z : |z - c^p| \leq c^p \,\}, (p+1)c^p \,\} \setminus \{\, 0, (p+1)c^p \,\}.$$

Since X_0 is a multiple of I the X_k are all rational functions of A. The Jordan canonical form of A therefore enables us to reduce the proof to the case of Jordan blocks $J(\lambda)$, where $\lambda \in E(c,p)$. Using Lemmas 2.5 and 2.6 we deduce that X_k has a limit X_* that satisfies $X_*^{-p} = A$ and has the same eigenvalues as $A^{-1/p}$. Since $A^{-1/p}$ is the only inverse pth root having these eigenvalues, $X_* = A^{-1/p}$. Now

$$\begin{aligned}X_{k+1} - A^{-1/p} &= \frac{1}{p}\left[(p+1)X_k(A^{-1/p})^p - p(A^{-1/p})^{p+1} - X_k^{p+1}\right] A \\ &= \frac{1}{p}\left[-(X_k - A^{-1/p})^2 \sum_{i=1}^{p} i X_k^{p-i}(A^{-1/p})^{i-1}\right] A,\end{aligned}$$

and hence we have

$$\|X_{k+1} - A^{-1/p}\| \leq \|X_k - A^{-1/p}\|^2 \cdot p^{-1}\|A\| \sum_{i=1}^{p} i \|X_k^{p-i}\| \, \|A^{(1-i)/p}\|,$$

which implies that the convergence is quadratic. ∎

Recall that the convergence results summarized in section 1 require $\rho(I - X_0^p A) < 1$ and do not specify to which root the iteration converges. When $X_0 = c^{-1} I$ this condition is $\max_i |\lambda_i - c^p| < c^p$, where $\Lambda(A) = \{\lambda_1, \ldots, \lambda_n\}$ is the spectrum of A. Theorem 2.7 guarantees convergence to the inverse principal pth root for $\Lambda(A)$ lying in the much larger region $E(c,p)$. The actual convergence region, determined experimentally, is shown together with $E(c,p)$ in Figure 2.2 for $c = 1$ and several values of p.

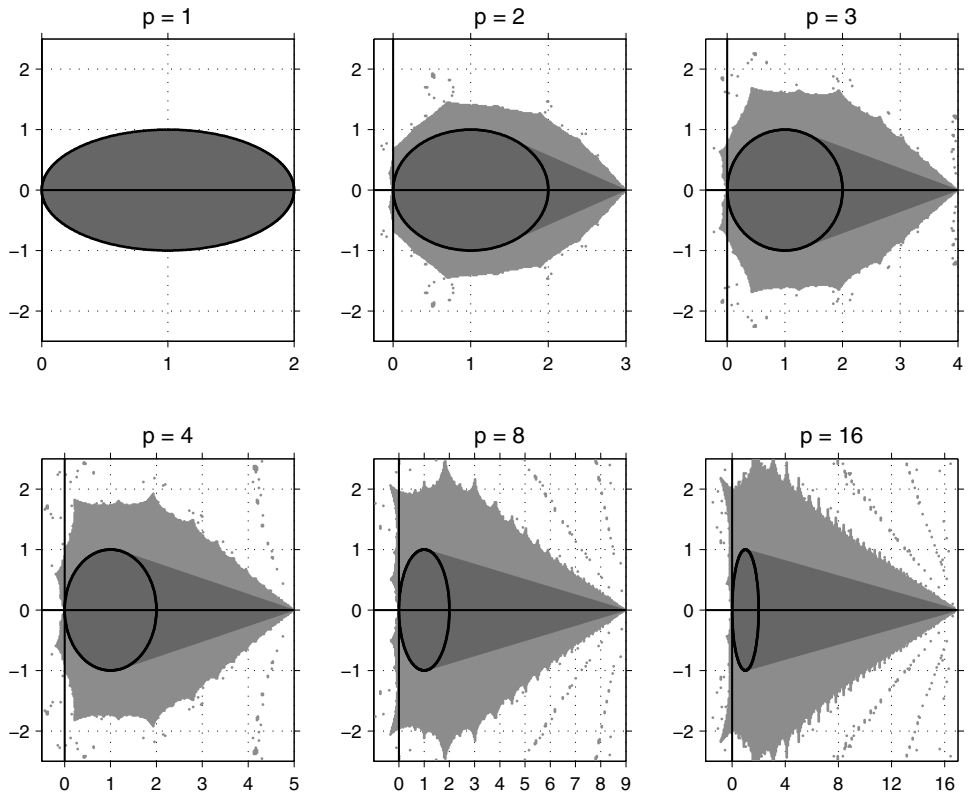


FIG. 2.2. *Regions of $\lambda \in \mathbb{C}$ for which the inverse Newton iteration* (2.2) *with $x_0 = 1$ converges to $\lambda^{-1/p}$. The dark shaded region is $E(1,p)$. The union of that region with the lighter shaded points is the experimentally determined region of convergence. The solid line marks the disk of radius* 1*, center* 1*. Note the differing x-axis limits.*

3. Practical algorithms. Armed with the convergence result in Theorem 2.7, we now build two practical algorithms applicable to arbitrary $A \in \mathbb{C}^{n\times n}$ having no eigenvalues on $\mathbb{R}^-$. Both preprocess A by computing square roots before applying the Newton iteration, one by computing a Schur decomposition and thereby working with (quasi-) triangular matrices.

We take $X_0 = c^{-1}I$, where the parameter $c \in \mathbb{R}^+$ is at our disposal. Thus, to recap, the iteration is

$$X_{k+1} = \frac{1}{p}\left[(p+1)X_k - X_k^{p+1}A\right], \qquad X_0 = \frac{1}{c}I. \tag{3.1}$$

Note that scaling X_0 through c is equivalent to fixing $X_0 = I$ and scaling A: if $X_k(X_0, A)$ denotes the dependence of X_k on X_0 and A then

$$X_k(c^{-1}I, A) = c^{-1}X_k(I, c^{-p}A).$$

We begin, in the next section, by considering numerical stability.

3.1. Coupled iterations. The Newton iteration (3.1) is usually numerically unstable. Indeed, the iteration can be guaranteed to be stable only if the eigenvalues

of A satisfy [4]

$$\frac{1}{p}\left|p-\sum_{r=1}^{p}\left(\frac{\lambda_i}{\lambda_j}\right)^{r/p}\right|\le 1,\quad i,j=1\colon n.$$

This is a very restrictive condition on A. However, by introducing the matrix $M_k = X_k^p A$, the iteration can be rewritten in the coupled form

$$\begin{aligned} X_{k+1} &= X_k\left(\frac{(p+1)I-M_k}{p}\right), & X_0 &= \frac{1}{c}I,\\ M_{k+1} &= \left(\frac{(p+1)I-M_k}{p}\right)^p M_k, & M_0 &= \frac{1}{c^p}A. \end{aligned} \tag{3.2}$$

When $X_k \to A^{-1/p}$ we have $M_k \to I$. This coupled iteration was suggested, and its unconditional stability noted, by Iannazzo [19]. In fact, (3.2) is a special case of a family of iterations of Lakić [26], and stability of the whole family is proved in [26].

Since the X_k in (3.2) are the same as those in the original iteration, their residuals R_k satisfy Lemma 2.1. Since $M_k = I - R_k$ and $M_k \to I$, the R_k are errors for the M_k.

Note that by setting $Y_k = X_k^{-1}$ we obtain from (3.2) a new coupled iteration for computing $A^{1/p}$:

$$\begin{aligned} Y_{k+1} &= \left(\frac{(p+1)I-M_k}{p}\right)^{-1} Y_k, & Y_0 &= cI,\\ M_{k+1} &= \left(\frac{(p+1)I-M_k}{p}\right)^p M_k, & M_0 &= \frac{1}{c^p}A. \end{aligned} \tag{3.3}$$

If $A^{1/p}$ is wanted without computing any inverses then $A^{1/p}$ can be computed from (3.2) and the formula $A^{1/p} = A(A^{-1/p})^{p-1}$ used (cf. (1.3)).

3.2. Algorithm not requiring eigenvalues. We now outline an algorithm that works directly on A and does not compute any spectral information. We begin by taking the square root twice by any iterative method [15]. This preprocessing step brings the spectrum into the sector $\arg z \in (-\pi/4, \pi/4)$. The nearest point to the origin that is both within this sector and on the boundary of $E(c,p)$ is at a distance $c^p\sqrt{2}$. Hence the inverse Newton iteration in the form (3.2) can be applied to $B = A^{1/4}$ with $c \ge (\rho(B)/\sqrt{2})^{1/p}$. If $\rho(B)$ is not known and cannot be estimated then we can replace it by the upper bound $\|B\|$, for some norm. This corresponds with the scaling used by Iannazzo in Algorithm 1.1 for $A^{1/p}$. A disadvantage of using the norm is that for nonnormal matrices $\rho(B) \ll \|B\|$ is possible, and this can lead to much slower convergence, as illustrated by the following example.

We use the inverse Newton iteration to compute $B^{-1/2}$, where $B = \left[\begin{smallmatrix}\epsilon & 1\\ 0 & \epsilon\end{smallmatrix}\right]$ and $\epsilon \ll 1$. If we use $c = (\|B\|_1/\sqrt{2})^{1/2}$, the convergence will be very slow, since for the eigenvalue ϵ, $r_0(\epsilon) = 1 - x_0^2\epsilon \approx 1 - \sqrt{2}\epsilon$. If we use $c = (\rho(B)/\sqrt{2})^{1/2}$, then we have $r_0(\epsilon) = 1 - \sqrt{2}$ and the convergence will be fast (modulo the nonnormality). The best choice of c for this example, however, is $c = \epsilon^{1/2}$. For this c we have immediate convergence to the inverse square root: $X_1 = B^{-1/2}$. This finite convergence behavior is a special case of that described in the next result.

LEMMA 3.1. [illegible] $A \in \mathbb{C}^{n\times n}$ [illegible] λ [illegible] n [illegible] q [illegible] (3.1) [illegible] $c = \lambda^{1/p}$ [illegible] $X_m = A^{-1/p}$ [illegible] m [illegible] $2^m \ge q$

Let A have the Jordan form $A = ZJZ^{-1}$. Then $R_0 = I - X_0^p A = Z(I - \frac{1}{\lambda}J)Z^{-1}$. Thus $R_0^q = 0$. By Lemma 2.1, $R_m = (R_0)^{2^m} h(R_0)$, where $h(R_0)$ is a polynomial in R_0. Thus $R_m = 0$ if $2^m \geq q$. □

As for the complexity of iteration (3.2), the benchmark with which to compare is the Schur method for the pth root of M. I. Smith [33]. It computes a Schur decomposition and obtains the pth root of the triangular factor by a recurrence, with a total cost of $(28 + (p-1)/3)n^3$ flops. The cost of one iteration of (3.2) is about $2n^3(2 + \theta \log_2 p)$ flops, where $\theta \in [1, 2]$, assuming that the pth power in (3.2) is evaluated by binary powering [10, Alg. 11.2.2]. Since at least four iterations will typically be required, unless p is large ($p \geq 200$, say) it is difficult for (3.2) to be competitive in its operation count with the Schur method. However, the Newton iterations are rich in matrix multiplication and matrix inversion, and on a modern machine with a hierarchical memory these operations are much more efficient relative to a Schur decomposition than their flop counts suggest. For special matrices A, such as the strictly diagonally dominant stochastic matrices arising in the Markov model application in section 4, we can apply (3.2) and (3.3) with $c = 1$ without any preprocessing, which makes this approach more efficient.

3.3. Schur–Newton algorithm. We now develop a more sophisticated algorithm that begins by computing a Schur decomposition $A = QRQ^*$ (Q unitary, R upper triangular). The Newton iteration is applied to a triangular matrix obtained from R, thereby greatly reducing the cost of each iteration. We begin by considering the choice of c, exploiting the fact that the spectrum of A is now available.

We consider first the case where the eigenvalues λ_i of A are all real and positive: $0 < \lambda_n \leq \cdots \leq \lambda_1$. Consider the residual $r_k(\lambda) = 1 - x_k^p \lambda$, and note that

$$r_{k+1}(\lambda) = 1 - \frac{1}{p^p}(1 - r_k(\lambda))(p + r_k(\lambda))^p. \tag{3.4}$$

Recall from Lemmas 2.1 and 2.2 that if $r_0 \in E$, or equivalently $\lambda \in E(c, p)$, then $|r_1| < 1$ and $|r_{i+1}| \leq |r_i|^2$ for $i \geq 1$. For c large enough, the spectrum of A lies in $E(c, p)$ and convergence is guaranteed. However, if c is too large, then $r_0(\lambda_n) = 1 - (\frac{1}{c})^p \lambda_n$ is extremely close to 1; $r_1(\lambda_n)$ is then also close to 1, by (3.4), and the convergence for the eigenvalue λ_n is very slow. On the other hand, if c is so small that $(\frac{1}{c})^p \lambda_1$ is close to (but still less than) $p + 1$, then $r_0(\lambda_1) = 1 - (\frac{1}{c})^p \lambda_1$ is close to $-p$, and, by (3.4), $r_1(\lambda_1)$ is very close to 1. Ideally we would like to choose c to minimize $\max_i |r_1(\lambda_i)|$.

LEMMA 3.2. [illegible] A [illegible] $0 < \lambda_n \leq \cdots \leq \lambda_1$ [illegible] $r_k(\lambda) = 1 - x_k^p \lambda$ [illegible] $c \in \mathbb{R}^+$ [illegible]

$$-p < r_0(\lambda_1) \leq r_0(\lambda_2) \leq \cdots \leq r_0(\lambda_n) < 1, \tag{3.5}$$

[illegible] $0 \leq r_j(\lambda_i) < 1$ [illegible] $j \geq 1$ [illegible] $i = 1\colon n$ [illegible]

$$\widehat{r}_j := \max_{1 \leq i \leq n} r_j(\lambda_i) = \max\bigl(r_j(\lambda_1), r_j(\lambda_n)\bigr).$$

[illegible] $j \geq 1$ $\widehat{r}_j$ [illegible]

$$c = \left(\frac{\alpha^{1/p}\lambda_1 - \lambda_n}{(\alpha^{1/p} - 1)(p+1)}\right)^{1/p}, \qquad \alpha = \frac{\lambda_1}{\lambda_n}, \tag{3.6}$$

[illegible] $\lambda_1 > \lambda_n$ [illegible] $\lambda_1 = \lambda_n$ [illegible] $\widehat{r}_j = 0$ [illegible] $j \geq 0$ [illegible] $c = \lambda_n^{1/p}$

TABLE 1
Values of $f(\alpha,p)$ for some particular α and p.

α	2	5	10	50	100
$p=2$	0.0852	0.3674	0.5883	0.8877	0.9403
$p=5$	0.0690	0.3109	0.5190	0.8452	0.9125
$p=10$	0.0635	0.2902	0.4915	0.8247	0.8979
$p=1000$	0.0580	0.2688	0.4618	0.7999	0.8795

. For each eigenvalue λ, we have, by (3.4), $r_{k+1}(\lambda)=f(r_k(\lambda))$ with $f(x)=1-\frac{1}{p^p}(1-x)(p+x)^p$. Since $f'(x)=\frac{p+1}{p^p}x(p+x)^{p-1}$, $f(x)$ is decreasing on $(-p,0]$ and increasing on $[0,1)$, and since $f(-p)=f(1)=1$ and $f(0)=0$ it follows that $0\le f(x)<1$ on $(-p,1)$. The first part of the result follows immediately. Since $f(x)$ is increasing on $[0,1)$, $\widehat{r}_j$ is minimized for all $j\ge 1$ if and only if $\widehat{r}_1$ is minimized. If $\lambda_1>\lambda_n$ it is easily seen that $\widehat{r}_1$ is minimized when $r_1(\lambda_1)=r_1(\lambda_n)$, i.e.,

$$\lambda_1\left(p+1-\lambda_1/c^p\right)^p=\lambda_n\left(p+1-\lambda_n/c^p\right)^p,$$

from which we find that c is given by (3.6). It is straightforward to verify that for this c, (3.5) holds. The formula (3.6) is not valid when $\lambda_1=\lambda_n$. However, we have

$$\lim_{\lambda_1\to\lambda_n} c=\lim_{\alpha\to 1}\left(\frac{\alpha^{1+1/p}-1}{\alpha^{1/p}-1}\frac{\lambda_n}{p+1}\right)^{1/p}=\lambda_n^{1/p}.$$

Note that when $\lambda_1=\lambda_n$, $r_0(\lambda_1)=r_0(\lambda_n)=0$ for $c=\lambda_n^{1/p}$. Therefore $\widehat{r}_j=0$ for all $j\ge 0$. □

When $\lambda_1>\lambda_n$, a little computation shows that the minimum value of $\widehat{r}_1$, achieved for c in (3.6), is

$$f(\alpha,p)=1-\alpha\frac{(p+1)^{p+1}}{p^p}\frac{(\alpha-1)^p(\alpha^{1/p}-1)}{(\alpha^{1+1/p}-1)^{p+1}}.$$

Numerical experiments suggest that $f(\alpha,p)$ is increasing in α for fixed p, and decreasing in p for fixed α. Moreover, it is easy to show that $\lim_{\alpha\to 1^+} f(\alpha,p)=0$. Some particular values of $f(\alpha,p)$ are given in Table 1. From the table, we can see that the values of $f(\alpha,p)$ are not sensitive to p but are sensitive to α. It is advisable to preprocess the matrix A to achieve $\alpha\le 2$, since $f(\alpha,p)$ is then safely less than 1 and rapid convergence can be expected.

We develop the idea of preprocessing in the context of general A with possibly nonreal eigenvalues. Suppose the eigenvalues are ordered $|\lambda_n|\le\cdots\le|\lambda_1|$. A convenient way to reduce $\chi(A):=|\lambda_1|/|\lambda_n|$ is to take k_1 square roots of the triangular matrix R in the Schur form, which can be done using the method of Björck and Hammarling [5], or that of Higham [14] if R is real and quasi-triangular. Since $\chi(A)=\chi(R)\le\kappa_2(R)$, in IEEE double precision arithmetic we can reasonably assume that $\chi(R)\le 10^{16}$, and then $k_1\le 6$ square roots are enough to achieve $\chi(R^{1/2^{k_1}})\le 2$. Write $p=2^{k_0}q$ where q is odd. If $q=1$, $R^{1/p}$ can be computed simply by k_0 square roots. If $q\ge 3$, we will take a total of $\max(k_0,k_1)$ square roots, compute the qth root by the Newton iteration, and finish with k_1-k_0 squarings if $k_1>k_0$. Taking $k_1>k_0$ is justified by the operation counts if it saves just one iteration of the Newton process, because for triangular matrices the cost of a square root and a squaring is at most half of the cost of one Newton iteration. When R has nonreal eigenvalues we will increase k_1, if

necessary, so that the matrix $B = R^{1/2^{k_1}}$ to which we apply the Newton iteration has spectrum in the sector $\arg z \in (-\pi/8, \pi/8)$; in general we therefore require $k_1 \geq 3$. Then we take $c = (\frac{\mu_1+\mu_n}{2})^{1/q}$, where $\mu_i = |\lambda_i|^{1/2^{k_1}}$. For any eigenvalue μ of B we have $\frac{2}{3} \leq (\frac{1}{c})^q|\mu| \leq \frac{4}{3}$, since $\mu_1/\mu_n \leq 2$, and thus $|1 - (\frac{1}{c})^q\mu| \leq |1 - \frac{4}{3}e^{i\frac{\pi}{8}}| \approx 0.56$. So the convergence of (3.2) is expected to be fast.

We now present our algorithm for computing the (inverse) principal pth root of a general A. We state the algorithm for real matrices, but an analogous algorithm is obtained for complex matrices by using the complex Schur decomposition.

ALGORITHM 3.3. [illegible] $A \in \mathbb{R}^{n\times n}$ [illegible] $\mathbb{R}^-$ [illegible] $X = A^{1/p}$ [illegible] $X = A^{-1/p}$ [illegible] $p = 2^{k_0}q$ [illegible] $k_0 \geq 0$ [illegible] q

```
1  Compute a real Schur decomposition A = QRQ^T.
2  if q = 1
3     k_1 = k_0
4  else
5     Choose k_1 ≥ k_0 such that |λ_1/λ_n|^{1/2^{k_1}} ≤ 2,
      where the eigenvalues of A are ordered |λ_n| ≤ ··· ≤ |λ_1|.
6  end
7  If the λ_i are not all real and q ≠ 1, increase k_1 as necessary so that
   arg(λ_i^{1/2^{k_1}}) ∈ (−π/8, π/8) for all i.
8  Compute B = R^{1/2^{k_1}} by k_1 invocations of the method of Higham [14] for the
   square root of a quasi-triangular matrix. If q = 1, goto line 21.
9  Let μ_1 = |λ_1|^{1/2^{k_1}}, μ_n = |λ_n|^{1/2^{k_1}}.
10 if the λ_i are all real
11    if μ_1 ≠ μ_n
12       determine c by (3.6) with λ_1, λ_n, p in (3.6) replaced by μ_1, μ_n, q
13    else
14       c = μ_n^{1/q}
15    end
16 else
17    c = ((μ_1+μ_n)/2)^{1/q}
18 end
19 Compute { X = B^{-1/q} by (3.2),   if A^{-1/p} required,
           { X = B^{1/q}  by (3.3),   if A^{1/p}  required.
20 X ← X^{2^{k_1−k_0}} (repeated squaring).
21 X ← QXQ^T
```

The cost of the algorithm is about

$$\left(28 + \frac{2}{3}(k_1 + k_2) - \left(\frac{1}{3} + \frac{k_2}{2}\right)k_0 + \frac{k_2}{2}\log_2 p\right)n^3 \text{ flops},$$

where we assume that k_2 iterations of (3.2) or (3.3) are needed (the cost per iteration is the same for both for triangular matrices, except on the first iteration, where (3.2) requires $n^3/3$ fewer flops because X_1 does not require a matrix multiplication). When $k_0 = 0$, $k_1 = 3$, and $k_2 = 4$, for example, the flop count becomes $(32\frac{2}{3} + 2\log_2 p)n^3$, while the count is always $(28 + \frac{p-1}{3})n^3$ for Smith's method. Note, however, that the computational work can be reduced for Smith's method if p is not prime by applying the method over the prime factors of p (this is not beneficial for Algorithm 3.3).

Our algorithm is slightly more expensive than Smith's method if p is small or highly composite, but it is much less expensive than Smith's method if p is large and has a small number of prime factors.

Algorithm 3.3 can be modified to compute $A^{1/p}$ in a different way: by computing $X = B^{-1/q}$ in line 19 and replacing line 21 with $X \leftarrow QX^{-1}Q^T$, which is implemented as a multiple right-hand-side triangular solve followed by a matrix multiplication. The modified line 21 costs the same as the original, so the cost of the algorithm is unchanged. We will call this variant Algorithm 3.3a.

A key feature of Algorithm 3.3 is that it applies the Newton iteration to a (quasi-) triangular matrix—one that has been "preconditioned" so that few iterations will be required. This can be expected to improve the numerical properties of the iteration, not least because for triangular matrices inversion and the solution of linear systems tend to be more accurate than the conventional error bounds suggest [16, Chap. 8].

4. An application to Markov models. Let $P(t)$ be a transition matrix for a time-homogeneous continuous-time Markov process. Thus $P(t)$ is a stochastic matrix: an $n \times n$ real matrix with nonnegative entries and row-sums 1. A generator Q of the Markov process is an $n \times n$ real matrix with nonnegative off-diagonal entries and zero row-sums such that $P(t) = e^{Qt}$. Clearly, Q must satisfy $e^Q = P \equiv P(1)$. If P has distinct, real positive eigenvalues then the only real logarithm, and hence the only candidate generator, is the principal logarithm, $\log P$. In general, a generator may or may not exist, and if it exists it need not be the principal logarithm of P [32].

Suppose a given transition matrix $P \equiv P(1)$ has a generator $Q = \log P$. Then Q can be used to construct $P(t)$ at other times, through $P(t) = \exp(Qt)$. For example, if P is the transition matrix for the time period of one year then the transition matrix for a month is $P(1/12) = e^{\frac{1}{12}\log P}$. However, it is more direct and efficient to compute $P(1/12)$ as $P^{1/12}$, thus avoiding the computation of a generator. Indeed, the standard inverse scaling and squaring method for the principal logarithm of a matrix requires the computation of a matrix root, as noted in section 1. Similarly, the transition matrix for a week can be computed directly as $P^{1/52}$.

This use of matrix roots is suggested by Waugh and Abel [35], mentioned by Israel, Rosenthal, and Wei [20], and investigated in detail by Kreinin and Sidelnikova [24]. The latter authors, who are motivated by credit risk models, address the problems that the principal root and principal logarithm of P may have the wrong sign patterns; for example, the root may have negative elements, in which case it is not a transition matrix. They show how to optimally adjust these matrices to achieve the required properties, a process they term regularization. Their preferred method for obtaining transition matrices for short times is to regularize the appropriate matrix root.

Transition matrices arising in the credit risk literature are typically strictly diagonally dominant [20], and such matrices are known to have at most one generator [8]. For any strictly diagonally dominant stochastic matrix P, Gershgorin's theorem shows that every eigenvalue lies in one of the disks $|z - a_{ii}| \le 1 - a_{ii}$, and we have $a_{ii} > 0.5$, so the spectrum lies in $E(1, p)$ and the convergence of (3.2) and (3.3) (with $A = P$) is guaranteed with $c = 1$. Note, however, that faster convergence is possible by choosing $c < 1$ when P has eigenvalues close to 0. For $c = 1$, it is easy to see that $X_k e = e$ and $M_k e = e$ for each $k \ge 0$. Thus all approximations to $P^{1/p}$ obtained from (3.2) and (3.3) have unit row sums, though they are not necessarily nonnegative matrices.

To illustrate, consider the strictly diagonally dominant stochastic matrix [35]

$$P = \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.2 & 0.7 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}.$$

Suppose we wish to compute $P(1/12)$ and $P(1/52)$. After (for example) four iterations of (3.3) with $c = 1$ we obtain (to four decimal places)

$$p = \frac{1}{12}: \qquad X = \begin{bmatrix} 0.9518 & 0.0384 & 0.0098 \\ 0.0253 & 0.9649 & 0.0098 \\ 0.0106 & 0.0089 & 0.9805 \end{bmatrix}, \qquad \|X^{12} - P\|_F = 4.7 \times 10^{-7}$$

and

$$p = \frac{1}{52}: \qquad X = \begin{bmatrix} 0.9886 & 0.0092 & 0.0023 \\ 0.0060 & 0.9917 & 0.0023 \\ 0.0025 & 0.0021 & 0.9954 \end{bmatrix}, \qquad \|X^{52} - P\|_F = 2.5 \times 10^{-7},$$

and both matrices are stochastic to the working precision of about 10^{-16}. Note that such a computation, requiring just matrix multiplication and the solution of multiple right-hand side linear systems, is easily carried out in a spreadsheet, which is a computing environment used by some finance practitioners.

In summary, Markov models provide an application of matrix roots that is little known to numerical analysts, and the Newton iterations (3.2) and (3.3) for computing these roots are well suited to the application.

5. Numerical experiments. We present some numerical experiments to compare the behavior of Algorithm 1.1, Algorithm 3.3, and the Schur method of Smith [33]. First, we need to develop appropriate residual-based measures of numerical stability for pth roots and inverse pth roots.

Let $\widetilde{X} = X + E$ be an approximation to a pth root X of $A \in \mathbb{C}^{n\times n}$. Then $\widetilde{X}^p = A + \sum_{i=0}^{p-1} X^i E X^{p-1-i} + O(\|E\|^2)$. An obvious residual bound is $\|A - \widetilde{X}^p\| \le p\|X\|^{p-1}\|E\| + O(\|E\|^2)$. While this bound is satisfactory for $p = 2$ [14], for $p \ge 3$ it can be very weak, since $\|X^i\| \le \|X\|^i$ can be an arbitrarily weak bound. Therefore we use the vec operator, which stacks the columns of a matrix into one long column, and the Kronecker product [17, Chap. 4] to write

$$\operatorname{vec}(A - \widetilde{X}^p) = -\left(\sum_{i=0}^{p-1} \bigl(X^{p-1-i}\bigr)^T \otimes X^i\right) \operatorname{vec}(E) + O(\|E\|^2).$$

For the 2-norm, it follows that

$$\|A - \widetilde{X}^p\|_F \le \|E\|_F \left\| \sum_{i=0}^{p-1} \bigl(X^{p-1-i}\bigr)^T \otimes X^i \right\|_2 + O(\|E\|_F^2)$$

is a sharp bound, to first order in E. If we suppose that $\|E\|_F \le \epsilon \|X\|_F$, then

$$\frac{\|A - \widetilde{X}^p\|_F}{\|X\|_F \bigl\|\sum_{i=0}^{p-1} \bigl(X^{p-1-i}\bigr)^T \otimes X^i\bigr\|_2} \le \epsilon + O(\epsilon^2).$$

We conclude that if $\widetilde{X}$ is a correctly rounded approximation to a pth root $\widetilde{X}$ of A in floating point arithmetic with unit roundoff u, then we expect the [illegible]

$$\rho_A(\widetilde{X}) := \frac{\|A - \widetilde{X}^p\|}{\|\widetilde{X}\| \left\| \sum_{i=0}^{p-1} \left(\widetilde{X}^{p-1-i}\right)^T \otimes \widetilde{X}^i \right\|}$$

to be of order u, where for practical purposes any norm can be taken. Therefore $\rho_A(\widetilde{X})$ is the appropriate residual to compute and compare with u. In [4] and [19] the scaled residual $\|A - \widetilde{X}^p\|/\|A\|$ was computed; this makes the interpretation of the numerical results therein difficult when the denominator of $\rho_A(\widetilde{X})$ is not of the same order as $\|A\|$.

For an approximate inverse pth root $\widetilde{X} \approx A^{-1/p}$ the situation is more complicated, as there is no natural residual. Criteria can be based on $A\widetilde{X}^p - I$, $\widetilde{X}^p A - I$, or indeed $\widetilde{X}^i A \widetilde{X}^{p-i} - I$ for any $i = 0\colon p$, as well as $\widetilde{X}^{-p} - A$ and $\widetilde{X}^p - A^{-1}$. Since they reduce to the pth root case discussed above, we will use the latter two residuals, which lead to the relative residuals $\rho_A(\widetilde{X}^{-1})$ and $\rho_{A^{-1}}(\widetilde{X})$. We compute the inverses in high precision to ensure that errors in the inversion do not significantly influence the computed residuals.

Iterations (3.2) and (3.3) can be terminated when $\|M_k - I\|$ is less than a suitable tolerance (nu in our experiments). This test has negligible cost and has proved to be reliable when used within Algorithm 3.3. In Algorithm 1.1 square roots were computed using the Schur method [14].

Our computational experience on a wide variety of matrices is easily summarized. The Schur method invariably produces a computed $\widehat{X} \approx A^{1/p}$ with $\rho_A(\widehat{X}) \approx u$, and $\rho_{A^{-1}}(\widehat{X}^{-1})$ is usually of order u but occasionally much larger. When computing $A^{-1/p}$, Algorithm 3.3 usually produces an $\widehat{X}$ with $\rho_A(\widehat{X}^{-1})$ order u, but occasionally this residual is a couple of orders of magnitude larger. When computing $A^{1/p}$, Algorithms 3.3 and 3.3a invariably yield $\rho_A(\widehat{X}) \approx u$.

We describe MATLAB tests with two particular matrices and $p = 5$. The first matrix is `gallery('frank',8)^5`, where the Frank matrix is upper Hessenberg and has real eigenvalues, the smaller of which are ill conditioned. The second matrix is a random nonnormal 8×8 matrix constructed as $A = QTQ^T$, where Q is a random orthogonal matrix and T, is in real Schur form with eigenvalues $\alpha_j \pm i\beta_j$, $\alpha_j = -j^2/10$, $\beta_j = -j$, $j = 1\colon n/2$ and $(2j, 2j+1)$ elements -450. The infinity norm is used in evaluating ρ. The results are summarized in Tables 2 and 3. The values for k_0, k_1, and the number of iterations are the same for Algorithms 3.3 and 3.3a. For the Frank matrix, $\rho_A(\widehat{X}^{-1}) \gg u$ but for the pth root approximation obtained using Algorithms 3.3 and 3.3a the residual is of order u. The five iterations required by the iterative phase of Algorithm 3.3 are typical. Both matrices reveal two weaknesses of Algorithm 1.1: it can require many iterations, making it significantly more expensive than the Schur method, and it can suffer from instability, as indicated by the relative residuals.

6. Conclusions. Our initial aim in this work was to strengthen existing convergence results for Newton's method for the inverse pth root. The analysis has led us to develop a hybrid algorithm—employing a Schur decomposition, matrix square roots, and two coupled versions of the Newton iteration—that computes either $A^{1/p}$ or $A^{-1/p}$. The new algorithm performs stably in practice and it is more efficient than the Schur method of Smith for large p that are not highly composite. Although the Newton iterations for $A^{1/p}$ and $A^{-1/p}$ have until recently rarely been used for

TABLE 2
Results for Frank matrix. $p = 5$, $\|A\|_2 = 4.3 \times 10^6$, $\|A^{1/p}\|_2 = 2.4 \times 10^1$, $\|A^{-1/p}\|_2 = 1.0 \times 10^4$.

Schur $\widehat{X} \approx A^{1/p}$	Inverse Newton $\widehat{X} \approx A^{-1/p}$, $\widehat{Y} \approx A^{1/p}$ (Alg. 3.3) $\widehat{Z} \approx A^{1/p}$ (Alg. 3.3a)	Newton (Alg. 1.1) $\widehat{X} \approx A^{1/p}$
$\rho_A(\widehat{X}) = 1.5\text{e-}16$ $\rho_{A^{-1}}(\widehat{X}^{-1}) = 1.8\text{e-}7$	$\rho_A(\widehat{X}^{-1}) = 2.5\text{e-}13$ $\rho_{A^{-1}}(\widehat{X}) = 1.8\text{e-}7$ $\rho_A(\widehat{Y}) = 8.2\text{e-}15$ $\rho_A(\widehat{Z}) = 9.8\text{e-}16$	$\rho_A(\widehat{X}) = 1.8\text{e-}14$ $\rho_{A^{-1}}(\widehat{X}^{-1}) = 1.8\text{e-}7$
	$k_0 = 0$, $k_1 = 6$; 5 iterations	19 iterations

TABLE 3
Results for random nonnormal matrix. $p = 5$, $\|A\|_2 = 4.5 \times 10^2$, $\|A^{1/p}\|_2 = 9.2 \times 10^5$, $\|A^{-1/p}\|_2 = 1.0 \times 10^6$.

Schur $\widehat{X} \approx A^{1/p}$	Inverse Newton $\widehat{X} \approx A^{-1/p}$, $\widehat{Y} \approx A^{1/p}$ (Alg. 3.3) $\widehat{Z} \approx A^{1/p}$ (Alg. 3.3a)	Newton (Alg. 1.1) $\widehat{X} \approx A^{1/p}$
$\rho_A(\widehat{X}) = 3.6\text{e-}18$ $\rho_{A^{-1}}(\widehat{X}^{-1}) = 4.1\text{e-}18$	$\rho_A(\widehat{X}^{-1}) = 5.0\text{e-}18$ $\rho_{A^{-1}}(\widehat{X}) = 9.7\text{e-}19$ $\rho_A(\widehat{Y}) = 1.5\text{e-}18$ $\rho_A(\widehat{Z}) = 5.4\text{e-}18$	$\rho_A(\widehat{X}) = 3.1\text{e-}12$ $\rho_{A^{-1}}(\widehat{X}^{-1}) = 1.6\text{e-}11$
	$k_0 = 0$, $k_1 = 3$; 5 iterations	21 iterations

$p > 2$, our work and that of Iannazzo [19] shows that these iterations are valuable practical tools and that general-purpose algorithms can be built around them based on understanding of their convergence properties.

Acknowledgments. This work was carried out while the first author visited MIMS in the School of Mathematics at the University of Manchester; he thanks the School for its hospitality. Both authors thank the referees for their helpful comments.

REFERENCES

[1] D. H. BAILEY, *MPFUN: A Portable High Performance Multiprecision Package*, Technical Report RNR-90-022, NASA Ames Research Center, Moffett Field, CA, 1990.
[2] D. H. BAILEY, *A Fortran 90-based multiprecision system*, ACM Trans. Math. Software, 21 (1995), pp. 379–387.
[3] D. H. BAILEY, Y. HIDA, X. S. LI, AND B. THOMPSON, *ARPREC: An Arbitrary Precision Computation Package*, Technical Report LBNL-53651, Lawrence Berkeley National Laboratory, Berkeley, CA, 2002.
[4] D. A. BINI, N. J. HIGHAM, AND B. MEINI, *Algorithms for the matrix pth root*, Numer. Algorithms, 39 (2005), pp. 349–378.
[5] Å. BJÖRCK AND S. HAMMARLING, *A Schur method for the square root of a matrix*, Linear Algebra Appl., 52/53 (1983), pp. 127–140.
[6] S. H. CHENG, N. J. HIGHAM, C. S. KENNEY, AND A. J. LAUB, *Approximating the logarithm of a matrix to specified accuracy*, SIAM J. Matrix Anal. Appl., 22 (2001), pp. 1112–1125.
[7] M. CORNEA-HASEGAN AND B. NORIN, *IA-64 floating-point operations and the IEEE standard for binary floating-point arithmetic*, Intel Technology Journal, 3 (1999), ftp://download.intel.com/technology/itj/q41999/pdf/ia64fpbf.pdf.
[8] J. R. CUTHBERT, *On uniqueness of the logarithm for Markov semi-groups*, J. London Math. Soc., 4 (1972), pp. 623–630.
[9] J.-C. EVARD AND F. JAFARI, *A complex Rolle's theorem*, Amer. Math. Monthly, 99 (1992), pp. 858–861.

[10] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[11] N. J. Higham, *Functions of a Matrix: Theory and Computation*, book in preparation.
[12] N. J. Higham, *Computing the polar decomposition—with applications*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1160–1174.
[13] N. J. Higham, *Newton's method for the matrix square root*, Math. Comp., 46 (1986), pp. 537–549.
[14] N. J. Higham, *Computing real square roots of a real matrix*, Linear Algebra Appl., 88/89 (1987), pp. 405–430.
[15] N. J. Higham, *Stable iterations for the matrix square root*, Numer. Algorithms, 15 (1997), pp. 227–242.
[16] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., SIAM, Philadelphia, PA, 2002.
[17] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.
[18] B. Iannazzo, *A note on computing the matrix square root*, Calcolo, 40 (2003), pp. 273–283.
[19] B. Iannazzo, *On the Newton method for the matrix pth root*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 503–523.
[20] R. B. Israel, J. S. Rosenthal, and J. Z. Wei, *Finding generators for Markov chains via empirical transition matrices, with applications to credit ratings*, Math. Finance, 11 (2001), pp. 245–265.
[21] A. H. Karp and P. Markstein, *High-precision division and square root*, ACM Trans. Math. Software, 23 (1997), pp. 561–589.
[22] C. S. Kenney and A. J. Laub, *Condition estimates for matrix functions*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 191–209.
[23] Ç. K. Koç and B. Bakkaloğlu, *Halley's method for the matrix sector function*, IEEE Trans. Automat. Control, 40 (1995), pp. 944–949.
[24] A. Kreinin and M. Sidelnikova, *Regularization algorithms for transition matrices*, Algo Research Quarterly, 4 (2001), pp. 23–40.
[25] P. Laasonen, *On the iterative solution of the matrix equation* $AX^2 - I = 0$, Math. Tables Aids Comput., 12 (1958), pp. 109–116.
[26] S. Lakić, *On the computation of the matrix k-th root*, Z. Angew. Math. Mech., 78 (1998), pp. 167–172.
[27] B. Meini, *The matrix square root from a new functional perspective: Theoretical results and computational issues*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 362–376.
[28] H.-O. Peitgen, H. Jürgens, and D. Saupe, *Fractals for the Classroom. Part Two: Complex Systems and Mandelbrot Set*, Springer-Verlag, New York, 1992.
[29] B. Philippe, *An algorithm to improve nearly orthonormal sets of vectors on a vector processor*, SIAM J. Alg. Discrete Methods, 8 (1987), pp. 396–403.
[30] M. Schroeder, *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*, W. H. Freeman, New York, 1991.
[31] L. S. Shieh, Y. T. Tsay, and C. T. Wang, *Matrix sector functions and their applications to system theory*, IEE Proc., 131 (1984), pp. 171–181.
[32] B. Singer and S. Spilerman, *The representation of social processes by Markov models*, Amer. J. Sociology, 82 (1976), pp. 1–54.
[33] M. I. Smith, *A Schur algorithm for computing matrix pth roots*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 971–989.
[34] R. A. Smith, *Infinite product expansions for matrix n-th roots*, J. Austral. Math. Soc., 8 (1968), pp. 242–249.
[35] F. V. Waugh and M. E. Abel, *On fractional powers of a matrix*, J. Amer. Statist. Assoc., 62 (1967), pp. 1018–1021.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 3, pp. 805–821

REDUCING THE TOTAL BANDWIDTH OF A SPARSE UNSYMMETRIC MATRIX*

J. K. REID† AND J. A. SCOTT†

Abstract. For a sparse symmetric matrix, there has been much attention given to algorithms for reducing the bandwidth. As far as we can see, little has been done for the unsymmetric matrix A, which has distinct lower and upper bandwidths l and u. When Gaussian elimination with row interchanges is applied, the lower bandwidth is unaltered, while the upper bandwidth becomes $l+u$. With column interchanges, the upper bandwidth is unaltered, while the lower bandwidth becomes $l+u$. We therefore seek to reduce $\min(l,u)+l+u$, which we call the *total bandwidth.* We compare applying the reverse Cuthill–McKee algorithm to $A+A^T$, to the row graph of A, and to the bipartite graph of A. We also propose an unsymmetric variant of the reverse Cuthill–McKee algorithm. In addition, we have adapted the node-centroid and hill-climbing ideas of Lim, Rodrigues, and Xiao to the unsymmetric case. We have found that using these to refine a Cuthill–McKee-based ordering can give significant further bandwidth reductions. Numerical results for a range of practical problems are presented and comparisons made with the recent lexicographical method of Baumann, Fleischmann, and Mutzbauer.

Key words. matrix bandwidth, sparse unsymmetric matrices, Gaussian elimination, Cuthill–McKee algorithm

AMS subject classification. 65F50

DOI. 10.1137/050629938

1. Introduction.

If Gaussian elimination is applied without interchanges to an unsymmetric matrix $A=\{a_{ij}\}$ of order n, each fill-in takes place between the first entry of a row and the diagonal or between the first entry of a column and the diagonal. It is therefore sufficient to store all the entries in the lower triangle from the first entry in each row to the diagonal and all the entries in the upper triangle from the first entry in each column to the diagonal. This simple structure allows straightforward code using static data structures to be written. We will call the sum of the lengths of the rows the [illegible] and the sum of the lengths of the columns the [illegible].

We will also use the term [illegible] for $l=\max_{a_{ij}\neq 0}(i-j)$ and the term [illegible] for $u=\max_{a_{ij}\neq 0}(j-i)$. For a symmetric matrix, these are the same and are called the [illegible]. A particularly simple data structure is available by taking account of only the bandwidths l and u. If row interchanges are used for stability reasons during the factorization, it may be readily verified that the lower bandwidth remains l but the upper bandwidth may increase to $l+u$. With column interchanges (or row interchanges applied while factorizing A^T), the upper bandwidth is unaltered, while the lower bandwidth becomes $l+u$. We may therefore always have one triangular factor of bandwidth $\min(l,u)$ and the other of bandwidth $l+u$. Thus we seek to reduce $\min(l,u)+l+u$, which we call the [illegible].

Many algorithms for reducing the bandwidth of a sparse symmetric matrix A have been proposed in the literature, most of which make extensive use of the adjacency graph $\mathcal{G}$ of the matrix. This is an undirected graph that has a node for each row (or

*Received by the editors April 25, 2005; accepted for publication (in revised form) by V. Simoncini March 22, 2006; published electronically October 4, 2006.
http://www.siam.org/journals/simax/28-3/62993.html
†Computational Science and Engineering Department, Atlas Centre, Rutherford Appleton Laboratory, Oxon OX11 0QX, England (j.k.reid@rl.ac.uk, j.a.scott@rl.ac.uk). The work of the second author was supported by the EPSRC grant GR/S42170.

column) of the matrix, and node i is a neighbor of node j if a_{ij} (and by symmetry a_{ji}) is an entry (nonzero) of A. An important and well-known example of an algorithm that uses $\mathcal{G}$ is that of Cuthill and McKee [2]. The main aim of this paper is to consider how variants of the Cuthill–McKee algorithm can be used to order an unsymmetric matrix for small total bandwidth.

In some circumstances, reordering the matrix and then using a band solver will be the method of choice for solving large sparse linear systems. However, in many situations it is more appropriate to use other sparse direct methods. In this study, we concentrate solely on the reduction of the bandwidth of unsymmetric matrices and do not address the question of when a band solver is the best choice.

The rest of this paper is organized as follows. We begin (in section 2) by commenting on the importance of reordering a matrix to block form prior to applying a bandwidth reduction algorithm. In section 3, we briefly describe the Cuthill–McKee algorithm and the variant that reverses the order (RCM). Then, in section 4, we discuss three undirected graphs that can be associated with an unsymmetric matrix A and that can be reordered using RCM. We then propose in section 5 an unsymmetric variant of RCM. In section 6, we look at modifying the hill-climbing algorithm of Lim, Rodrigues, and Xiao [10] to improve a given ordering, and in section 7, we propose a variant of the node-centroid algorithm of [10] for the unsymmetric case. In section 8, we discuss the recently published algorithm of Baumann, Fleischmann, and Mutzbauer [1] for reducing the bandwidth of an unsymmetric matrix. In section 9, we use our proposed algorithms to reorder a set of matrices that arise from a range of practical problems; we report the total bandwidths before and after reordering, and we summarize our findings in section 10.

2. The block triangular form. In the symmetric case, it may be possible to preorder the matrix A to block diagonal form

$$\begin{bmatrix} A_{11} & & & & & \\ & A_{22} & & & & \\ & & A_{33} & & & \\ & & & A_{44} & & \\ & & & & \ddots & \end{bmatrix}. \tag{2.1}$$

In this case, each block may be permuted to band form, and the overall matrix is a band matrix; the profile is the sum of the profiles of the blocks, and the bandwidth is the greatest bandwidth of a block.

The unsymmetric case is not so straightforward because we need also to exploit the block triangular form

$$\begin{bmatrix} A_{11} & & & & & \\ A_{21} & A_{22} & & & & \\ A_{31} & A_{32} & A_{33} & & & \\ A_{41} & A_{42} & A_{43} & A_{44} & & \\ \ldots & \ldots & \ldots & \ldots & \ldots \end{bmatrix}, \tag{2.2}$$

where the blocks A_{ll}, $l = 1, 2, \ldots, N$, are all square. A matrix that can be permuted to this form with $N > 1$ diagonal blocks is said to be [illegible]; if no block triangular form other than the trivial one with a single block ($N = 1$) can be found, the matrix is [illegible]. The advantage of the block triangular form (2.2) is that the corresponding

set of equations $Ax = b$ may be solved by the block forward substitution

$$A_{ii}x_i = b_i - \sum_{j=1}^{i-1} A_{ij}x_j, \quad i = 1, 2, \ldots, N. \tag{2.3}$$

There is no fill in the off-diagonal blocks, which are involved only in matrix-by-vector multiplications. It therefore suffices to permute each diagonal block A_{ii} to band form. We will take the upper and lower profiles to be the sums of the upper and lower profiles of the diagonal blocks, and the upper and lower bandwidths to be the greatest of the upper and lower bandwidths of the diagonal blocks.

3. The Cuthill–McKee algorithm. The Cuthill–McKee algorithm is a well-known and successful algorithm for reducing the bandwidth of a symmetric matrix of order n. It does this for a given starting node s by relabeling the nodes of the adjacency graph $\mathcal{G}$ in order of increasing distance from s. The algorithm is outlined in Figure 1. Here the degree of a node i is defined as the number of its neighbors. If $\mathcal{G}$ has more than one component, the procedure is repeated from a starting node in each component.

```
ALGORITHM CUTHILL–MCKEE.
Label s as node 1; l_1 = {s}; i = 1
do k = 2, 3, ... until i = n
    l_k = {}
    do for each v ∈ l_{k-1} in label order
        do for each neighbor u of v that has not been labeled,
                in order of increasing degree
            add u to l_k; i = i + 1; label u as node i
        end do
    end do
end do
```

FIG. 1. *Cuthill–McKee ordering algorithm.*

Ordering the nodes in this way groups them into "level sets," that is, nodes at the same distance from the starting node. Since nodes in level set l_k can have neighbors only in level sets l_{k-1}, l_k, and l_{k+1}, the reordered matrix is block tridiagonal with blocks corresponding to the level sets. It is therefore desirable that the level sets be small, which is likely if there are many of them. The size of the largest level set is called the [illegible] of the level structure. The width and number of level sets ([illegible] of the level structure) are dependent on the choice of the starting node s. Algorithms for finding a good starting node are usually based on finding a pseudodiameter (pair of nodes that are a maximum distance apart or nearly so). Much effort has gone into efficiently finding a pseudodiameter; see, for example, [7] and [12] and the references therein. The modified Gibbs Poole Stockmeyer (MGPS) algorithm of Reid and Scott [12] is outlined in Figure 2. For efficiency, the test on $w(r)$ may be performed during the formation of the level set structure so that the structure can be discarded as soon as a large value of $w(r)$ is found. In the inner loop, the choice of 5 nodes was made on the basis of numerical experimentation.

George [6] found that the profile may be reduced if the Cuthill–McKee ordering is reversed (the bandwidth is unchanged). The reverse Cuthill–McKee (RCM) algorithm and variants of it remain in common use. For example, an implementation is available

ALGORITHM MGPS.
Construct $\mathcal{G}$ and choose a starting node s of smallest degree
Form the level structure rooted at s, of height $h(s)$ and width $w(s)$
outer: do
  $w_{least} = \infty$
  inner: do for up to 5 nodes r of the final level set that are not
      neighbors, in order of increasing degree
    Form the level set structure rooted at r,
    of height $h(r)$ and width $w(r)$
    if $w(r) \geq w_{least}$ **cycle inner**
    if $h(r) > h(s)$ **then**
      $s = r$; **cycle outer**
    end if
    $e = r$; $w_{least} = w(r)$;
  end do inner
  exit outer
end do outer
Pseudodiameter is defined by the pair (s, e).

FIG. 2. *Modified Gibbs Poole Stockmeyer algorithm.*

within MATLAB as the function `symrcm`, and RCM is included as an option within the package `MC60` from the mathematical software library HSL [8]. We note that both these implementations apply RCM directly to the supplied matrix A, without attempting to first reorder the matrix to block diagonal form (2.1).

4. Undirected graphs for unsymmetric matrices. In this section, we consider three adjacency graphs that can be associated with an unsymmetric matrix A. In each case, we employ the RCM algorithm to reduce the semibandwidth of the graph, and this permutation is then used to reorder A.

4.1. Using $A + A^T$. For a matrix whose structure is nearly symmetric, an effective strategy is to find a symmetric permutation that reduces the bandwidth of the structure of the symmetric matrix $A+A^T$. The MATLAB function `symrcm` applies RCM to the adjacency graph of $A + A^T$. If the symmetric permutation is applied to A, the lower and upper bandwidths are no greater than the semibandwidth of the permuted $A + A^T$. Of course, the same algorithm may be applied to a matrix that is far from symmetric, and the same results apply, but the effectiveness is uncertain. It is likely to be helpful to permute A to make it more symmetric. We will judge this by its [illegible], that is, the number of off-diagonal entries a_{ij} for which a_{ji} is also an entry, divided by the total number of off-diagonal entries. Permuting a large number of off-diagonal entries onto the diagonal reduces the number of unmatched off-diagonal entries, which in turn generally increases the symmetry index (see, for example, [5], [9]).

Note that most algorithms for preordering the matrix to block triangular form (2.2) begin with a permutation that places entries on the diagonal. Thus, in this case, permuting entries onto the diagonal is not available as a strategy for improving the symmetry index.

4.2. Bipartite graph. The bipartite graph of A, which we will denote by $\mathcal{G}_{bipart}$, has a node for each row and a node for each column, and row node i and column node

j are neighbors if a_{ij} is an entry. It is straightforward to see that $\mathcal{G}_{bipart}$ is actually the adjacency graph of the $2n \times 2n$ symmetric matrix

$$\hat{A} = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}. \tag{4.1}$$

Starting the Cuthill–McKee algorithm with any node of $\mathcal{G}_{bipart}$, the level sets are alternately sets of rows and sets of columns. If we start from a row node and perform the corresponding symmetric permutation on the matrix (4.1), we find the matrix

$$\begin{bmatrix} 0 & A_{11} & & & & & & \\ A_{11}^T & 0 & A_{21}^T & & & & & \\ & A_{21} & 0 & A_{22} & & & & \\ & & A_{22}^T & 0 & A_{32}^T & & & \\ & & & A_{32} & 0 & A_{33} & & \\ & & & & A_{33}^T & 0 & A_{34}^T & \\ & & & & & \cdots & \cdots & \cdots \end{bmatrix}, \tag{4.2}$$

where A_{lm} is the submatrix of A corresponding to the rows of row level set l and columns of column level set m.

If we permute the rows of A by the row level sets and the orderings within them, and permute the columns by the column level sets and the orderings within them, we find the block bidiagonal form

$$\begin{bmatrix} A_{11} & & & & \\ A_{21} & A_{22} & & & \\ & A_{32} & A_{33} & & \\ & & A_{43} & A_{44} & \\ & & & \cdots & \cdots \end{bmatrix}, \tag{4.3}$$

which is also the submatrix of (4.2) consisting of block rows 1, 3, ... and block columns 2, 4, ... We illustrate with a small reordered example in Figure 3. Here there are four row level sets, of sizes 1, 3, 2, 2, and three column level sets, of sizes 3, 2, 3.

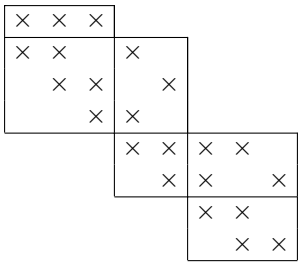


FIG. 3. *A matrix with rows and columns reordered using the permutations obtained by applying Cuthill–McKee to its bipartite graph.*

This example has entries on the whole of the diagonal, which will not necessarily be the case. However, if the matrix is structurally nonsingular, the diagonal will always intersect each of the blocks. This is because the leading k columns must be of full structural rank and similarly for the leading k rows.

The semibandwidth of the reordered matrix (4.2) is at most one less than the largest sum of the sizes of two adjacent level sets, that is, one less than the largest

sum of the sizes of a row level set and an adjacent column level set. The corresponding results for the reordered unsymmetric matrix (4.3) are that the lower bandwidth is at most one less than the sum of the sizes of two adjacent column level sets and the upper bandwidth is at most one less than the sum of the sizes of two adjacent row level sets. Note, however, that all these bounds are pessimistic; they do not take into account the ordering of the nodes within each level set (and RCM does well in this respect) and, in the case (4.3), of the position of the matrix diagonal within the blocks.

4.3. Row graph. Another alternative is to consider the [illegible] [11] of A, which is defined to be the adjacency graph of the symmetric matrix AA^T, where matrix multiplication is performed without taking cancellations into account (so that, if a coefficient of AA^T is zero as a result of numerical cancellation, it is still considered to be an entry). The nodes of the row graph correspond to the rows of A, and nodes i and j ($i \neq j$) are neighbors if and only if there is at least one column k of A for which a_{ik} and a_{jk} are both entries. The row graph has been used by Scott [13], [14] to order the rows of unsymmetric matrices prior to solving the linear system using a frontal solver. We can obtain an ordering for the rows of A by applying the RCM algorithm to AA^T. This will ensure that rows with entries in common are nearby; that is, the first and last entry of each column will not be too far apart. If the columns are now ordered according to their last entry, the lower bandwidth will be small and the upper bandwidth will not be large.

A potential disadvantage of computing and working with the pattern of AA^T is that it can be costly in terms of time and memory requirements. This is because AA^T may contain many more entries than A. It fails completely if A has a full column (AA^T is full), but such a matrix cannot be permuted to have small lower and upper bandwidths.

5. Unsymmetric RCM. Any reordering within a Cuthill–McKee level set of section 4.2 will alter the positions of the leading entries of the columns of a submatrix A_{ii} or the rows of a submatrix A_{ji}, $j = i+1$. It will make exactly the same change to the profile of the matrix (4.2) as it does to the sum of the upper and lower profiles of the matrix (4.3). If it reduces the bandwidth of the matrix (4.2), it will reduce either the upper or lower bandwidth of the matrix (4.3); however, the converse is not true: It might reduce the upper or lower bandwidth of the matrix (4.3) without reducing the bandwidth of the matrix (4.2). It follows that it may be advantageous for bandwidth reduction to develop a special-purpose code for the unsymmetric case, rather than giving the matrix (4.1) to a general-purpose code for reducing the bandwidth of a symmetric matrix. We have developed a prototype unsymmetric bandwidth reduction code of this kind. Our algorithm is based on the MGPS algorithm outlined in Figure 2. As in the bipartite approach discussed in section 4.2, we use the adjacency graph $\mathcal{G}_{bipart}$ so that the level sets alternate between a set of rows and a set of columns, but our unsymmetric algorithm bases its decisions on the total bandwidth of the unsymmetric matrix A rather than on the bandwidth of the matrix (4.2). Our algorithm is given in Figure 4.

The profile and bandwidth are likely to be reduced if the rows of each level set are ordered according to their leading entries. This happens automatically with the Cuthill–McKee algorithm and was done for the example in Figure 3.

Reversing the Cuthill–McKee ordering may reduce the profile. It is reduced if the column index of the trailing entry in a row is lower than the column index of the trailing entry in an earlier row. There is an example in block A_{22} of Figure 3.

ALGORITHM UNSYMMETRIC RCM.
Construct $\mathcal{G}_{bipart}$ and choose a starting node s of smallest degree
Apply Cuthill–McKee from s, finding height $h(s)$ and total bandwidth $t(s)$
$t_{least} = \infty$; $h_{most} = 1$
outer: do
  inner: do for up to 5 nodes r of the final level set that are not neighbors,
      in order of increasing degree
    Apply Cuthill–McKee from r, finding height $h(r)$
    and total bandwidth $t(r)$
    if $t(r) > t_{least}$ **cycle inner**
    if $t(r) < t_{least}$ or $h(r) > h(s)$ **then**
      $e = s$; $s = r$; $t_{least} = t(r)$; $h_{most} = h(s)$; **cycle outer**
    end if
  end do inner
  exit outer
end do outer
Order using the Cuthill–McKee ordering from e
Reverse the order

FIG. 4. *Unsymmetric RCM ordering algorithm.*

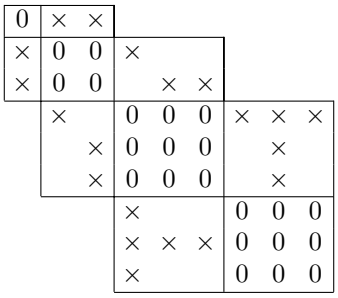


FIG. 5. *A symmetric matrix ordered by Cuthill–McKee.*

6. Hill climbing to improve a given ordering. In this and the next section, we consider algorithms that are not based on level sets in a graph and are therefore completely different.

Lim, Rodrigues, and Xiao [10] propose a hill-climbing algorithm for reducing the semibandwidth of a symmetric matrix. An entry a_{ij} in a matrix A with semibandwidth b is called critical if $|i-j| = b$. For each critical entry a_{ij} in the lower-triangular part, an interchange of i with $k < i$ or j with $k > j$ is sought that will reduce the number of critical entries. For example, a_{94} is critical in Figure 5, and the semibandwidth is reduced from 5 to 4 by interchanging column 4 with column 5 and row 4 with row 5. As a column is moved backwards, its first entry is moved away from the diagonal, while its last entry is moved nearer. If the distance of the first entry from the diagonal is d, we can therefore limit the choice of k to the range $j < k < j + b - d$ since we want $d + k - j$ to be smaller than the bandwidth b. Similarly, if the distance of the last entry in row i from the diagonal is l, we limit the choice of k to the range $i - b + l < k < i$. Each interchange while the semibandwidth is b reduces the number of critical entries by one. If the number of critical entries becomes zero, we recommence the algorithm for semibandwidth $b - 1$ and continue until none of the critical entries

for the current semibandwidth can be interchanged to reduce their number. The algorithm is summarized as Figure 6. Note that this hill-climbing algorithm cannot increase the semibandwidth.

```
ALGORITHM HC (SYMMETRIC).
outer: do
    Form the set V_c of critical nodes
    do until V_c is empty
        if there are nodes u ∈ V_c and v ∉ V_c such that
          swapping u and v leaves both noncritical then
          swap u and v and remove u from V_c
        else
          exit outer
        end if
    end do
end do outer
```

FIG. 6. *Hill climbing algorithm for symmetric matrices.*

We have adapted this idea to reduce the lower and upper bandwidths of an unsymmetric matrix. If the lower bandwidth is l and the upper bandwidth is u, we call an entry a_{ij} in the lower triangle for which $i - j = l$ a [illegible] entry, and an entry a_{ij} in the upper triangle for which $j - i = u$ a [illegible] entry. We have found it convenient to alternate between making row interchanges while the column permutation is fixed and making column interchanges while the row permutation is fixed. While making row interchanges to reduce the number of critical upper entries, we seek to exchange a row i containing a critical upper entry with another row so that the number of critical upper entries is reduced by one while the lower bandwidth is not increased. If the distance between the leading entry in the row and the diagonal is d, we limit our search to rows in the range $i-l+d \leq k < i$. For example, we do not exchange rows 3 and 4 in Figure 3, since this would increase the lower bandwidth.

Similarly, while making row interchanges to reduce the number of critical lower entries, we seek to exchange a row i containing a critical lower entry with another row so that the number of critical lower entries is reduced by one while the upper bandwidth is not increased. The row hill-climbing algorithm is outlined in Figure 7. Column hill climbing is analogous, using column interchanges to first reduce the upper bandwidth as much as possible and then to reduce the lower bandwidth as much as possible.

One complete iteration of our hill-climbing algorithm for unsymmetric matrices consists of row hill climbing followed by column hill climbing. We continue until a complete iteration fails to reduce one of the bandwidths or the total number of critical entries. This is illustrated in Figure 8.

7. Node centroid ordering. The hill-climbing algorithm of the previous section is essentially a local search and is very dependent on the initial order that it is given. To generate other initial orderings, Lim, Rodrigues, and Xiao [10] propose an algorithm that they call "node-centroid." For the graph of a symmetric matrix, they define $N_\lambda(i)$ to be the set of neighbors j of node i for which the distance $|i - j|$ is at least λb, where b is the semibandwidth and $\lambda \leq 1$ is a parameter for which they recommend a value of 0.95. They refer to such neighbors as λ [illegible]. $w(i)$ is then defined as the average node index over $i \cup N_\lambda(i)$, and the nodes are ordered by increas-

```
Algorithm HC (row).
rows: do
    Form the set V_u of rows that contain a critical upper entry
    do until V_u is empty
        if there are rows u ∈ V_u and v ∉ V_u such that swapping leaves both
                noncritical and does not increase the lower bandwidth then
            swap u and v and remove u from V_u
        else
            exit rows
        end if
    end do
end do rows
cols: do
    Form the set V_l of columns that contain a critical lower entry
    do until V_l is empty
        if there are columns u ∈ V_u and v ∉ V_u such that swapping leaves
            both noncritical and does not increase the upper bandwidth then
            swap u and v and remove u from V_l
        else
            exit cols
        end if
    end do
end do cols
```

Fig. 7. *Row hill climbing algorithm for unsymmetric matrices.*

```
Algorithm HC (unsymmetric).
do while lower bandwidth, upper bandwidth, or
        number of critical entries is reduced
    call HC(row)
    call HC(column)
end do
```

Fig. 8. *Hill climbing for unsymmetric matrices.*

ing $w(i)$. This will tend to move a row with a λ-critical entry in the lower triangle but no λ-critical entry in the upper triangle forward; hopefully, its new leading entry will be nearer the diagonal than the old one was, and its trailing entry will not have moved out so much that it becomes critical. Similar arguments apply to a row with a λ-critical entry in the upper triangle but no λ-critical entry in the lower triangle, which will tend to be moved back. The algorithm is outlined in Figure 9.

Lim, Rodrigues, and Xiao [10] apply a sequence of major steps, each of which consists of two iterations of node centroid ordering followed by one iteration of hill climbing, as illustrated in Figure 10. The decision to perform hill climbing after two steps of the node-centroid algorithm was taken on the basis of numerical experimentation. Using a Cuthill–McKee-type initial ordering with a random starting node, Lim, Rodrigues, and Xiao [10] report encouraging results for the DWT set of symmetric problems from the Harwell–Boeing Sparse Matrix Collection [4].

We have adapted this idea to the unsymmetric case by again alternating between permuting the rows while the column permutation is fixed and permuting the columns

ALGORITHM NC (SYMMETRIC).
choose $\lambda \leq 1$.
do $i = 1, n$
  $w(i) = i$; $c(i) = 1$; form $N_\lambda(i)$
  do for each $j \in N_\lambda(i)$
    $w(i) = w(i) + j$; $c(i) = c(i) + 1$
  end do
  $w(i) = w(i)/c(i)$
end do
sort entries of w into increasing order
reorder nodes in accord with the sorted sequence.

FIG. 9. *Node centroid algorithm for symmetric matrices.*

ALGORITHM NCHC (SYMMETRIC).
choose an initial ordering
do while semibandwidth is reduced
  call NC(symmetric)
  call NC(symmetric)
  call HC(symmetric)
end do

FIG. 10. *Node centroid plus hill climbing for symmetric matrices.*

while the row permutation is fixed. Suppose that the lower bandwidth is l and the upper bandwidth is u. While permuting the rows, only the leading and trailing entries of the rows are relevant, since they will still have these properties after the row permutation. If the leading or trailing entry of row i is λ-critical, it is desirable to move the row. If its leading entry is in column l_i and its trailing entry is in column u_i, the gap between the upper band and the trailing entry is $u + i - u_i$, and the gap between the lower band and the leading entry is $l_i - (i - l) = l_i - i + l$. If we move the row forward to become row $i + \delta$, the gaps become $u + i + \delta - u_i$ and $l_i - i - \delta + l$. If $l > u$, it would seem desirable to make the gap at the trailing end greater than the gap at the leading end. We choose a parameter $\alpha > 1$ and aim for the gap at the trailing end to be α times greater than the gap at the leading end; that is,

$$u + i + \delta - u_i = \alpha(l_i - i - \delta + l) \tag{7.1}$$

or

$$\delta = \frac{(u_i - i - u) + \alpha(l_i - i + l)}{1 + \alpha}. \tag{7.2}$$

Similar calculations for $l = u$ and $l < u$ lead us to conclude that a desirable position for the row is given by the equation

$$w(i) = \begin{cases} i + \frac{(u_i - i - u) + \alpha(l_i - i + l)}{1 + \alpha} & \text{if } l > u, \\ i + \frac{(u_i - i - u) + (l_i - i + l)}{2} & \text{if } l = u, \\ i + \frac{\alpha(u_i - i - u) + (l_i - i + l)}{1 + \alpha} & \text{if } l < u. \end{cases} \tag{7.3}$$

For other rows, we set $w(i) = i$. We sort the rows in increasing order of $w(i)$, $i = 1, 2, \ldots, n$. This is summarized in Figure 11. In our numerical experiments (see section 9), we found that a suitable value for α is 2.

ALGORITHM NC (ROW).
choose $\lambda \leq 1$ and $\alpha > 1$.
compute l, u.
if $(l > u)$ **then**
  $\beta = 1/(1+\alpha)$; $\gamma = \alpha/(1+\alpha)$
else if $(l < u)$ **then**
  $\beta = \alpha/(1+\alpha)$; $\gamma = 1/(1+\alpha)$
else
  $\beta = 1/2$; $\gamma = 1/2$
end if
do $i = 1, n$
  $w(i) = i$
  compute l_i, u_i
  if $u_i - i > \lambda u$ **or** $i - l_i > \lambda l$ **then**
    $w(i) = i + \beta * (u_i - i - u) + \gamma * (l_i - i + l)$
  end if
end do
sort entries of w into increasing order
reorder rows in accord with the sorted sequence.

FIG. 11. *Node centroid algorithm for ordering the rows of an unsymmetric matrix.*

Similar considerations apply to ordering the columns of the matrix with the row order fixed. We apply a sequence of major steps, each consisting of two iterations of the node-centroid row ordering followed by row hill climbing, then two iterations of the node-centroid column ordering followed by column hill climbing. We continue until the total bandwidth ceases to decrease. This is illustrated in Figure 12. We found in our numerical experiments that it is sufficient to limit the number of cycles of the do loop to 10.

ALGORITHM NCHC (UNSYMMETRIC).
choose an initial ordering
do
  call NC(row)
  call NC(row)
  call HC(row)
  if total bandwidth not reduced **exit**
  call NC(column)
  call NC(column)
  call HC(column)
  if total bandwidth not reduced **exit**
end do

FIG. 12. *Node centroid plus hill climbing for unsymmetric matrices.*

8. Relaxed double ordering. Before presenting numerical results for our proposed algorithm, in this section we briefly discuss the recently published algorithm of Baumann, Fleischmann, and Mutzbauer [1] for reducing the bandwidth of an unsymmetric matrix. The sparsity pattern of the matrix is represented by a (0,1)-matrix, that is, a matrix which is the same as the original matrix except that each nonzero

entry is replaced by 1. Each row and column of the (0,1)-matrix then defines a binary number. The algorithm proceeds by alternating between ordering the rows in decreasing order and ordering the columns in decreasing order. The authors [1] show that this converges to a limit and call it a "double ordering." Following reverse Cuthill–McKee, they reverse the converged ordering. Since only the leading entries of the rows or columns affect the bandwidths and profiles, we have implemented an efficient variant in which no attempt is made to order the rows or columns with the same leading entry. We call this a [illegible] (RDO). Results for the RDO algorithm are included in section 9.

Unfortunately, there are huge numbers of double orderings, and Baumann, Fleischmann, and Mutzbauer [1] have no strategy for choosing a good one. For example, a Cuthill–McKee ordering produces a relaxed double ordering regardless of the starting node, since the leading entries of the rows (or columns) form a monotonic sequence. There is scope for the double ordering to reduce the profile of an RCM ordering, but our experience is that the improvement is slight and is often at the expense of the bandwidths (see section 9.2).

9. Numerical experiments. In this section, we first describe the problems that we use for testing the algorithms discussed in this paper and then present numerical results.

TABLE 9.1
The test problems; see text for details.

Identifier	Order	Number of entries	Symmetry index
`4cols`†	11770	43668	0.0159
`10cols`†	29496	109588	0.0167
`bayer01`	57735	277774	0.0002
`bayer03`	6747	56196	0.0031
`circuit_3`	12127	48137	0.7701
`ethylene-1`†	10673	80904	0.2973
`extr1`	2837	11407	0.0042
`g7jac200sc`	59310	837936	0.0323
`fidapm11`	22294	623554	1.0000
`hydr1`	5308	23752	0.0041
`impcol_d`	425	1339	0.0567
`jan99jac020sc`	6774	38692	0.0037
`lhr71c`	70304	1528092	0.0015
`mark3jac140`	64089	399735	0.0740
`poli_large`	15575	33074	0.0035
`radfr1`	1048	13299	0.0537
`rdist1`	4134	94408	0.0588
`sinc15`	11532	568526	0.0138
`Zhao2`	33861	166453	0.9225

9.1. Test problems. The test problems are listed in Table 9.1. Each arises from a real engineering or industrial application. Problems marked with a † are chemical process engineering problems that were supplied to us by Mark Stadtherr of the University of Notre Dame. The remaining problems are available through the University of Florida Sparse Matrix Collection [3]. Most of the test problems were chosen on the grounds of being highly unsymmetric, because working with the symmetrized matrix $A + A^T$ will be satisfactory for near-symmetric matrices. We include two (nearly) symmetric matrices to illustrate this. We have chosen problems of different sizes since, in our experience, for some users it is not just the very large

TABLE 9.2

Details of the block triangular form for our test problems. n_1 and n_2 are the numbers of 1×1 and 2×2 blocks; $n_{>2}$ is the number of larger blocks; n_{off} is the total number of entries in the off-diagonal blocks. For the largest diagonal block A_{kk}, m is the order, me is the number of entries, avg is the average number of entries per row, si is the symmetry index, and me_{kk} is the number of entries in $A_{kk}A_{kk}^T$.

Identifier	n_1	n_2	$n_{>2}$	n_{off}	Largest block A_{kk}				
					m	me	avg	si	me_{kk}
4cols	0	0	1	0	11770	43668	3.71	0.0159	210026
10cols	0	0	1	0	29496	109588	3.72	0.0167	527124
bayer01	8858	0	3	28228	48803	240222	4.92	0.0812	1236678
bayer03	1772	2	6	19575	4776	33555	7.02	0.1066	252236
circuit_3	4520	0	1	9593	7607	34024	4.47	0.5579	76178
ethylene-1	2137	0	7	12865	8336	65375	7.84	0.3000	203920
extr1	424	0	1	464	2413	10519	4.35	0.0935	34118
fidapm11	0	0	1	0	22294	623554	28.0	1.0000	4067228
g7jac200sc	0	0	1	0	59310	837936	14.1	0.0323	6377172
hydr1	968	0	6	1420	2370	11738	4.95	0.0730	47946
impcol_d	226	0	1	551	199	562	2.82	0.0275	1350
jan99jac020sc	0	0	1	0	6774	38692	5.71	0.00376	603720
lhr71c	7038	0	28	95912	7663	173683	22.7	0.07396	2877995
mark3jac140	0	0	1	0	64089	399735	6.24	0.4225	773752
poli_large	15450	12	4	17266	90	286	3.18	0.1633	680
radfr1	97	0	1	969	951	12233	12.9	0.4751	34032
rdist1	198	0	1	3959	3936	90251	22.9	0.4821	280538
sinc15	652	0	1	24343	10880	543531	50.0	0.2615	11063488
Zhao2	0	0	1	0	33861	166453	4.92	0.9225	549692

problems that are important: In their applications they must repeatedly factorize and solve many small or medium-sized problems efficiently, and so spending time and effort on getting a good ordering is essential.

In Table 9.2, we give details of the block triangular form for each of our test matrices. We note that 4cols, 10cols, fidapm11, g7jac200sc, jan99jac020sc, mark3jac140, and Zhao2 are irreducible, while a number of problems (including rdist1 and circuit_3) have only one block of order greater than 1. Most of the remaining problems have fewer than 10 blocks of order greater than 1. As expected, the matrix $A_{kk}A_{kk}^T$ contains many more entries than A_{kk}. We also note that, for the reducible examples, the symmetry index of A_{kk} is usually larger than that of the original matrix.

9.2. Test results. We first present results for applying the HSL [8] implementation of the RCM algorithm (MC60) to the following matrices: (i) $A + A^T$; (ii) $B + B^T$, where $B = PA$ is the permuted matrix after employing the HSL routine MC21 to put entries on the diagonal; (iii) AA^T; (iv) the matrix $\hat{A}$ given by (4.1); and (iv) Unsymmetric RCM code (this is column 8, which is headed A). The total bandwidth for each ordering and for the initial ordering is given in Table 9.3. Results are also given for the RDO algorithm (section 8). A blank entry in the $B + B^T$ column indicates that the matrix A has no zeros on the diagonal, and in these cases, MC21 is not applied. We see that for some problems applying MC21 prior to the reordering with RCM can significantly reduce the bandwidths, but narrower bandwidths are achieved by working with either the row graph (AA^T) or the bipartite graph ($\hat{A}$) or using the unsymmetric RCM. For many of our test examples, the RDO orderings are poorer. However, they are often a significant improvement on the initial ordering, and for a small number of problems (notably radfr1a and rdist1) RDO produces good orderings.

TABLE 9.3
The total bandwidth for the RDO and RCM ordering algorithms.

Identifier	Initial	RDO	RCM				
			$A+A^T$	$B+B^T$	AA^T	$\hat{A}$	A
4cols	11770	4768	846		460	565	541
10cols	29496	13855	1052		546	572	557
bayer01	57735	45332	52201	4117	2232	2236	2331
bayer03	6747	3660	6747	1074	651	651	612
circuit_3	12127	10979	12127	12127	10441	11157	11324
ethylene-1	10664	8301	7797		5114	5093	5306
extr1	2837	1610	2575	298	171	169	169
fidapm11	19515	3315	3189		3299	3211	3202
g7jac200sc1	44611	22822	41198		19554	20384	19248
hydr1	5308	2726	5308	559	337	334	324
impcol_d	425	153	241	219	123	117	102
jan99jac020s	6774	4708	6264		5016	4891	5885
lhr71c	58291	18181	27064	5802	3620	3771	3893
mark3jac140	6825	12126	6550		6106	6112	6159
poli_large	15575	6400	15575		6381	6316	5995
radfr1a	1048	95	970	142	71	98	147
rdist1	4134	195	3421	336	223	215	175
sinc15	11532	9686	11532	11532	11036	9623	10833
Zhao2	33861	33861	1476		1464	1476	565

TABLE 9.4
The total bandwidth for the HC, NCHC, RDO, and RCM ordering algorithms applied to the diagonal blocks of the block triangular form.

Identifier	Initial	HC	NCHC	RDO	RCM			
					$A+A^T$	AA^T(+RDO)	$\hat{A}$	A
4cols	11770	5134	3281	4768	846	**460** (1001)	565	504
10cols	29496	13801	6530	13855	1052	546 (1600)	572	**528**
bayer01	48803	48860	22394	34581	3483	**1768** (3056)	1823	**1776**
bayer03	4776	4792	2279	3617	740	527 (547)	**500**	**506**
circuit_3	7607	7607	3698	4776	1903	**1330** (1394)	**1321**	**1297**
ethylene-1	8336	8336	3540	4967	323	**179** (432)	**184**	230
extr1	2413	2413	1114	1660	240	**145** (266)	**149**	**148**
fidapm11	19515	7168	3880	3315	**3189**	3299 (3987)	**3211**	**3240**
g7jac200sc	44611	**16548**	**16149**	22822	41198	19554 (19279)	20384	19248
hydr1	2370	2370	1151	1640	198	129 (**112**)	134	129
impcol_d	199	194	133	70	98	79 (82)	67	**59**
jan99jac020s	6774	6478	**4046**	4708	6264	5016 (5321)	4891	5652
lhr71c	7663	7663	2911	5135	991	741 (2173)	**727**	**720**
mark3jac140	6825	**5055**	5857	12126	6550	6106 (8846)	6112	6123
poli_large	90	85	**59**	84	90	90 (79)	84	77
radfr1a	621	186	98	132	130	88 (93)	**85**	93
rdist1	341	129	**120**	155	346	188 (193)	189	192
sinc15	10880	10880	**9195**	10880	10880	10491 (10880)	10880	10880
Zhao2	33861	33861	14015	33861	1471	**1454** (2196)	**1467**	**1424**

Table 9.4 shows the effect of applying the ordering algorithms to the diagonal blocks of the block triangular form (2.2). As already noted, the construction of the block triangular form ensures that there are no zeros on the diagonal, so we do not preorder using MC21. Apart from this, the algorithms featured in Table 9.3 are featured here too. We also show results for the hill-climbing algorithm (HC) and hill-climbing plus the node-centroid algorithm (NCHC). For the node-centroid algorithm we have experimented with using values of λ in the range $[0.8, 1]$ and values of α in the range $[1.5, 2.5]$. Our experience was that the bandwidths were not very sensitive

TABLE 9.5

The total bandwidth after hill climbing and the node-centroid algorithm. All are applied to the diagonal blocks of the block triangular form.

Identifier	RCM + HC				RCM + NCHC			
	$A+A^T$	AA^T	$\hat{A}$	A	$A+A^T$	AA^T	$\hat{A}$	A
4cols	718	435	549	481	502	**395**	458	443
10cols	902	498	553	479	625	**448**	462	**447**
bayer01	3241	1739	1742	1756	2243	**1659**	**1675**	**1659**
bayer03	668	446	445	452	411	**381**	**384**	**377**
circuit_3	1715	1228	1227	1123	1356	**1065**	**1074**	**1095**
ethylene-1	271	172	173	216	174	169	**162**	203
extr1	190	119	120	131	130	**115**	119	**116**
fidapm11	**3154**	3261	3183	**3156**	**3123**	3336	3286	**3085**
g7jac200sc	37042	19244	19782	18660	22290	**17383**	**17451**	**17530**
hydr1	133	101	101	120	**89**	**91**	**91**	**89**
impcol_d	66	61	56	55	**50**	51	**49**	52
jan99jac020s	5758	4258	4401	5190	3883	3665	**3249**	3953
lhr71c	862	626	598	576	**540**	572	557	**540**
mark3jac140	6192	**6035**	**6044**	**6053**	**5951**	**5959**	**5946**	**5978**
poli_large	56	70	70	66	54	**50**	61	**52**
radfr1a	**57**	63	72	76	**58**	**58**	**57**	**58**
rdist1	148	133	156	158	**123**	**121**	**124**	**119**
sinc15	10880	8819	10880	10880	**7428**	8866	10648	8097
Zhao2	1471	1454	1467	**1420**	1473	**1446**	**1462**	**1442**

to the precise choice of λ, and for most examples 0.85 gave results that were within three percent of the best. For α, we found that a value of 2 gave slightly narrower bandwidths than either 1.5 or 2.5. We therefore used $\lambda = 0.85$ and $\alpha = 2$.

In Table 9.4 we have highlighted the narrowest bandwidths and those within three percent of the narrowest. As expected, the larger symmetry index for the diagonal blocks of the block triangular form results in an improvement in the performance of RCM applied to $A + A^T$, but it is still better to use the other RCM variants. There appears to be little to choose between RCM applied to the row graph, RCM applied to the bipartite graph, and our Unsymmetric RCM algorithm; for some of the examples, each produces the narrowest total bandwidth. In general, combining hill-climbing with the node-centroid algorithm is better than using hill-climbing alone, but this is not guaranteed. For a small number of problems (including `pol_large` and `rdist1`), the NCHC ordering has the smallest total bandwidth, but for many of the test examples it gives results that are significantly poorer than the RCM variants.

To see whether RDO can be successfully used to refine our RCM orderings, we have experimented with running RDO after RCM applied to AA^T; the results are in parentheses in Table 9.4 in the column headed AA^T(+RDO). For a number of problems (including `poli_large`) the bandwidth is reduced, but for others the results are much worse (for example, `4cols` and `bayer01`). Indeed, using RDO after RCM can be worse than using RDO on the original ordering (for example, `jan99jac020s` and `rdist1`). This illustrates that RDO is extremely sensitive to the initial ordering, and our findings lead us not to recommend its use.

In Table 9.5, we present results for applying the different RCM variants to the block triangular form, followed by applying either hill climbing alone (denoted by RCM + HC) or the node-centroid algorithm plus hill climbing (denoted by RCM + NCHC). Again, the narrowest total bandwidths (and those within three percent of the narrowest) are highlighted. Comparing the results in columns 2–5 of Table 9.5 with the corresponding results in Table 9.4, we see that hill climbing (which never increases the total bandwidth) can significantly improve the RCM orderings. However,

TABLE 9.6
The best and worse total bandwidths using the given ordering and nine random permutations.

Identifier	RCM + NCHC			
	$A + A^T$	AA^T	$\hat{A}$	A
4cols	[457,570]	[384,456]	[385,463]	[380,460]
10cols	[575,625]	[448,463]	[445,462]	[438,461]
bayer01	[1963,2243]	[1659,1680]	[1665,1694]	[1655,1682]
bayer03	[410,463]	[380,400]	[374,401]	[368,397]
circuit_3	[1298,1356]	[1033,1105]	[1041,1159]	[969,1106]
ethylene-1	[169,184]	[156,169]	[158,169]	[157,211]
extr1	[119,134]	[114,127]	[114,131]	[114,125]
fidapm11	[3119,3288]	[3205,3336]	[3172,3290]	[3085,3252]
g7jac200sc	[18849,22290]	[16793,18132]	[16460,18516]	[16396,17827]
hydr1	[89,98]	[89,93]	[89,92]	[88,93]
impcol_d	[50,57]	[43,56]	[42,54]	[46,52]
jan99jac020s	[3396,4107]	[3257,3665]	[3230,3724]	[3209,3953]
lhr71c	[540,633]	[555,600]	[554,578]	[506,700]
mark3jac140	[5951,7385]	[5909,5959]	[5946,6062]	[5941,6011]
poli_large	[49,55]	[50,59]	[55,66]	[49,53]
radfr1a	[56,62]	[58,58]	[58,62]	[58,61]
rdist1	[120,127]	[121,126]	[121,125]	[119,125]
sinc15	[6903,8652]	[7752,8866]	[8265,11532]	[7413,8523]
Zhao2	[1473,2055]	[1446,1467]	[1453,1464]	[1442,1521]

looking also at columns 6–9, it is clear that for all problems except fidapm11 and Zhao2 (the two nearly symmetric problems), the smallest bandwidths are achieved by using RCM + NCHC. For problems with an unsymmetric sparsity structure, the largest improvements resulting from using the node-centroid algorithm are to the orderings obtained using RCM applied to $A + A^T$; for some problems (including the bayer examples and lhr71c) the reductions resulting from including the node-centroid algorithm are more than 30 percent. However, for many of our unsymmetric examples, one of the other variants generally produces orderings with a smaller total bandwidth.

Finally, we note that for a small number of problems, none of our proposed algorithms was successful in significantly reducing the bandwidth. In particular, we were not able to reorder the problems g7jac200sc, jan99jac020s, mark3jac140, and sinc15 to have a small bandwidth. We are not able to predict a priori which problems we are able to reorder to have a small bandwidth using our algorithms.

9.3. The effect of random initial permutations. Finally, we tried applying the algorithms after applying random row and column permutations to the given matrix ordering. The results are shown in Table 9.6. It is indeed the case that better total bandwidths can often be found in this way, which points the way towards finding better algorithms. Meanwhile, if many problems with the same structure are to be solved (so that the cost of reordering may be amortized over the repeated factorizations), it may be worthwhile to perform such random permutations and take the best resulting ordering. The conclusion that we drew from Table 9.5, that there is little to help us choose between the algorithms of the final three columns, is true here too.

10. Concluding remarks. We have considered algorithms for reducing the lower and upper bandwidths l and u of an unsymmetric matrix A, focusing on the total bandwidth, which we have defined as $l + u + \min(l, u)$, because this is relevant for the storage and work when sets of banded linear equations are solved by Gaussian elimination.

The least satisfactory results came from working with the lexicographical method of Baumann, Fleischmann, and Mutzbauer [1] and with the pattern of $A + A^T$, although for unsymmetrically structured matrices the use of the unsymmetric node-centroid algorithm plus hill climbing dramatically improved the results of applying the reverse Cuthill–McKee ordering to $A + A^T$. For the majority of our test problems, we achieved good results by applying the RCM algorithm to the matrices AA^T (whose graph is the row graph) and $\left[\begin{smallmatrix} 0 & A \\ A^T & 0 \end{smallmatrix}\right]$ (whose graph is the bipartite graph). Our unsymmetric variant of RCM gave comparable results. The results were improved by preordering A to block triangular form and applying one of these three RCM-based algorithms to the blocks on the diagonal. The rest of the matrix is used unaltered. The bandwidths were further reduced using our unsymmetric node-centroid and hill-climbing algorithms.

In general, the time taken to reorder an unsymmetric matrix using our algorithms is significantly less than the time required to subsequently factorize the matrix. However, since the codes used to generate the numerical results presented in this paper are prototypes, we have not reported the reordering times. In the future, we plan to include carefully designed efficient implementations of our new algorithms within the mathematical software library HSL [8].

Acknowledgments. We are grateful to Iain Duff of the Rutherford Appleton Laboratory and Yifan Hu of Wolfram Research for helpful comments on a draft of this paper, and to the anonymous referees for their suggestions.

REFERENCES

[1] M. Baumann, P. Fleischmann, and O. Mutzbauer, *Double ordering and fill-in for the LU factorization*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 630–641.

[2] E. Cuthill and J. McKee, *Reducing the bandwidth of sparse symmetric matrices*, in Proceedings of the 24th National Conference of the ACM, Brandon Systems Press, 1969, pp. 157–172.

[3] T. Davis, *University of Florida Sparse Matrix Collection*, NA Digest, 97, 1997; full details from www.cise.ufl.edu/research/sparse/matrices/.

[4] I. S. Duff, R. G. Grimes, and J. G. Lewis, *Sparse matrix test problems*, ACM Trans. Math. Softw., 15 (1989), pp. 1–14.

[5] I. S. Duff and J. Koster, *The design and use of algorithms for permuting large entries to the diagonal of sparse matrices*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 889–901.

[6] A. George, *Computer Implementation of the Finite-Element Method*, Ph.D. thesis, Department of Computer Science, Report STAN CS-71-208, Stanford University, Stanford, CA, 1971.

[7] N. E. Gibbs, W. G. Poole, and P. K. Stockmeyer, *An algorithm for reducing the bandwidth and profile of a sparse matrix*, SIAM J. Numerical Analysis, 13 (1976), pp. 236–250.

[8] *HSL, A collection of Fortran codes for large-scale scientific computation*, 2004; online at http://hsl.rl.ac.uk/.

[9] Y. F. Hu and J. A. Scott, *Ordering techniques for singly bordered block diagonal forms for unsymmetric parallel sparse direct solvers*, Numer. Linear Algebra Appl., 12 (2005), pp. 877–894.

[10] A. Lim, B. Rodrigues, and F. Xiao, *A centroid-based approach to solve the bandwidth minimization problem*, in Proceedings of the 37th Hawaii International Conference on System Sciences, IEEE Press, Piscataway, NJ, 2004, p. 30075a.

[11] B. H. Mayoh, *A graph technique for inverting certain matrices*, Math. Comp., 19 (1965), pp. 644–646.

[12] J. K. Reid and J. A. Scott, *Ordering symmetric sparse matrices for small profile and wavefront*, Internat. J. Numer. Methods Engrg., 45 (1999), pp. 1737–1755.

[13] J. A. Scott, *A new row ordering strategy for frontal solvers*, Numer. Linear Algebra Appl., 6 (1999), pp. 1–23.

[14] J. A. Scott, *Row ordering for frontal solvers in chemical process engineering*, Comput. Chem. Eng., 24 (2000), pp. 1865–1880.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 3, pp. 822–828

SHARP INEQUALITIES FOR SOME OPERATOR MEANS*

DRISS DRISSI†

Abstract. In this paper sharp results on strong domination between the Heinz and logarithmic means are obtained. This leads to sharp operator inequalities extending results given by Bhatia–Davis and Hiai–Kosaki on arithmetic-logarithmic-geometric mean matrix inequalities.

Key words. positive definite matrix, positive definite function, Fourier transform, Bochner's theorem, operator means

AMS subject classifications. 42A82, 47A62, 15A24

DOI. 10.1137/050648444

1. Introduction. There are several means that interpolate the geometric and arithmetic means; see [9], [13], and [14]. One that attracts many researchers is the so-called Heinz mean $H_\alpha(a,b)$ given by

$$H_\alpha(a,b) = \frac{a^{1-\alpha}b^\alpha + a^\alpha b^{1-\alpha}}{2} \text{ for } 0 \le \alpha \le 1.$$

Notice that $H_0(a,b) = H_1(a,b) = \frac{a+b}{2}$ is the arithmetic mean and $H_{\frac{1}{2}}(a,b) = \sqrt{ab}$ is the geometric mean.

In 1951, Heinz [8], in his study of perturbation theory of operators, proved that for the operator norm $\|.\|$, given A, B positive definite, for any X, that

$$\|A^{\frac{1}{2}}XB^{\frac{1}{2}}\| \le \frac{1}{2}\|A^{1-\alpha}XB^\alpha + A^\alpha XB^{1-\alpha}\| \le \frac{1}{2}\|AX + XB\|. \tag{1}$$

In 1993, Bhatia–Davis [1] proved that if A, B, and X are n by n matrices with A and B positive semidefinite, then for every unitarily invariant norm $|||.|||$,

$$|||A^{\frac{1}{2}}XB^{\frac{1}{2}}||| \le \frac{1}{2}|||A^{1-\alpha}XB^\alpha + A^\alpha XB^{1-\alpha}||| \le \frac{1}{2}|||AX + XB|||. \tag{2}$$

Another mean, which is of interest mainly in chemical engineering, statistics, and thermodynamics, is the logarithmic mean defined as

$$L(a,b) = \frac{a-b}{\log a - \log b} = \int_0^1 a^t b^{1-t}dt, \quad (a \ge 0, b \ge 0).$$

It is well known that

$$G(a,b) \le L(a,b) \le A(a,b). \tag{3}$$

In 1999, Hiai–Kosaki [10] obtained the following refinement of the inequality (2) showing:

$$|||A^{\frac{1}{2}}XB^{\frac{1}{2}}||| \le \left|\left|\left|\int_0^1 A^tXB^{1-t}dt\right|\right|\right| \le \frac{1}{2}|||AX + XB|||, \tag{4}$$

*Received by the editors December 27, 2005; accepted for publication (in revised form) by R. Bhatia April 13, 2006; published electronically October 4, 2006. This research was supported by Kuwait University Research Grants SM 02/05.

http://www.siam.org/journals/simax/28-3/64844.html

†Department of Mathematics & Computer Science, Kuwait University, P.O. Box 5969, Safat 13060, Kuwait (drissi@mcs.sci.kuniv.edu.kw).

called the arithmetic-logarithmic-geometric (A-L-G) inequality.

After seeing inequalities (2) and (4) it is hard not to be curious about the relationship between the Heinz and logarithmic means. This was our motivation to investigate this problem.

Assume $M(a,b)$, $N(a,b)$ are symmetric homogeneous means on $(0,\infty)\times(0,\infty)$. M is said to strongly dominate N, in notation $M \ll N$, if and only if the matrix

$$\left[\frac{M(\lambda_i,\lambda_j)}{N(\lambda_i,\lambda_j)}\right]_{i,j=1,\dots,n}$$

is positive semidefinite for any $\lambda_1,\dots,\lambda_n > 0$ with any size n (see [11] for more details). Note that the inequality $M \ll N$ is stronger than the usual order $M \le N$. In [10], Hiai–Kosaki gave an example showing this. Another example was later obtained by Bhatia [4]. Moreover, if A is a positive semidefinite matrix with eigenvalues $\lambda_1,\lambda_2,\dots,\lambda_n$, then $M \ll N$ is equivalent to the operator norm inequality

$$|||M(A,A)\circ X||| \le |||N(A,A)\circ X|||,$$

where $\circ$ is the Schur–Hadamard or the entrywise product, and $M(A,A)$ is the matrix whose ij entry is $M(\lambda_i,\lambda_j)$.

Schur's theorem asserts that the Schur–Hadamard product of two positive matrices is positive. Two matrices A and B are said to be congruent if $B = S^*AS$ for some nonsingular matrix S. If A is positive, then so is every matrix congruent to it. A complex-valued function f on $\mathbb{R}$ is said to be positive definite if the matrix $[f(x_i - x_j)]$ is positive semidefinite for all choices of points $\{x_1,x_2,\dots,x_n\}\subset\mathbb{R}$ and all $n = 1,2,\dots$. Another interesting result that we are going to use is the well-known theorem of Bochner (see [12] for more details) which asserts that a function f in $L^1(\mathbb{R})$ is positive definite if and only if its Fourier transform $\hat{f}(\xi) \ge 0$, for almost all ξ. When calculating Fourier transforms, we ignore constant factors, since the only property of $\hat{f}$ we use is whether it is nonnegative almost everywhere.

In this paper we first present a necessary and sufficient condition for the strong domination of the Heinz mean by the logarithmic mean. This follows from the following theorem, which may be of independent interest, on the positive definiteness of functions; see [2], [3], [4], [5], [6], and [11] for other results on positive definiteness of functions. Second, using a standard result on a norm of the Schur multiplier, we derive norm inequalities extending results given by Bhatia–Davis and Hiai–Kosaki on A-L-G mean matrix inequalities.

2. Main results. THEOREM 1.

$$f(x) = \frac{x\cosh(\beta x)}{\sinh(x)}.$$

f $-\frac{1}{2} \le \beta \le \frac{1}{2}$

The following formulas are known from [7] and we provide the proofs for completeness and the reader's convenience.

LEMMA 1. $|\beta| < 1$

$$\int_0^\infty \frac{\sinh(\beta x)}{\sinh(x)}\cos(\xi x)dx = \frac{\pi\sin(\beta\pi)}{2(\cosh(\xi\pi)+\cos(\beta\pi))}, \tag{5}$$

$$\int_0^\infty \frac{\cosh(\beta x)}{\sinh(x)}\sin(\xi x)dx = \frac{\pi\sinh(\xi\pi)}{2(\cosh(\xi\pi)+\cos(\beta\pi))}. \tag{6}$$

. To compute the above integrals we use the method of residues. We proceed in two steps.

1. Let us consider the complex valued function

$$\phi(z) = \frac{\sinh(\beta z)}{\sinh(z)}e^{i\xi z}.$$

Then ϕ has poles at the points $z_k = ik\pi$, for $k = \pm1, \pm2, \ldots$. Now, consider the contour integral $\int_\Gamma \phi(z)dz$, where Γ is the rectangle with vertices at $(-R, 0)$, $(R, 0)$, $(R, i\pi)$, and $(-R, i\pi)$ described counterclockwise, with an indentation $\gamma_r : z = re^{i\theta} + i\pi$, for $0 \geq \theta \geq -\pi$, so as to avoid the pole at $i\pi$. Since there are no singularities of the integrand inside Γ, we obtain by Cauchy's theorem for analytic functions

$$\int_{-R}^{R}\phi(x)dx + \int_0^\pi \phi(R+iy)idy + \int_R^r \phi(x+i\pi)dx + \int_{\gamma_r}\phi(z)dz + \int_{-r}^{-R}\phi(x+i\pi)dx + \int_\pi^0 \phi(-R+iy)idy = 0.$$

Using the estimation lemma, we obtain along the two vertical lines

$$\left|\int_0^\pi \phi(R+iy)idy\right| \to 0 \text{ and } \left|\int_\pi^0 \phi(-R+iy)idy\right| \to 0 \text{ as } R\to\infty. \tag{7}$$

By Jordan's lemma, we get

$$\lim_{r\to 0}\int_{\gamma_r}\phi(z)dz = i(-\pi-0)(-i\sin(\beta\pi)e^{-\xi\pi} = -\pi\sin(\beta\pi))e^{-\xi\pi}. \tag{8}$$

On the other hand, using the identities

$$\sinh(a\pm ib) = \sinh(a)\cos(b) \pm i\cosh(a)\sin(b),$$

we obtain

$$\int_R^r \phi(x+i\pi)dx = e^{-\xi\pi}\int_r^R \frac{e^{i\xi x}}{\sinh(x)}[\sinh(\beta x)\cos(\beta\pi) + i\cosh(\beta x)\sin(\beta\pi)]dx$$

and

$$\int_{-r}^{-R}\phi(x+i\pi)dx = e^{-\xi\pi}\int_r^R \frac{e^{-i\xi x}}{\sinh(x)}[\sinh(\beta x)\cos(\beta\pi) - i\cosh(\beta x)\sin(\beta\pi)]dx.$$

Combining the two above identities and using Euler's formula, we obtain after simplifications

$$\int_R^r \phi(x+i\pi)dx + \int_{-r}^{-R}\phi(x+i\pi)dx = e^{-\xi\pi}\Big\{\cos(\beta\pi)\int_r^R \frac{\sinh(\beta x)}{\sinh(x)}(2\cos(\xi x))dx + i\sin(\beta\pi)\int_r^R \frac{\cosh(\beta x)}{\sinh(x)}(2i\sin(\xi x))dx\Big\}.$$

Using

$$\int_{-R}^{R}\phi(x)dx=\int_{-R}^{R}\frac{\sinh(\beta x)}{\sinh(x)}\cos(\xi x)dx$$
$$=2\int_{0}^{R}\frac{\sinh(\beta x)}{\sinh(x)}\cos(\xi x)dx$$

and taking $r\to 0$ then after that $R\to\infty$, we obtain

$$2\int_{0}^{\infty}\frac{\sinh(\beta x)}{\sinh(x)}\cos(\xi x)dx+e^{-\xi\pi}\Big\{2\cos(\beta\pi)\int_{0}^{\infty}\frac{\sinh(\beta x)}{\sinh(x)}\cos(\xi x)dx$$
$$-2\sin(\beta\pi)\int_{0}^{\infty}\frac{\cosh(\beta x)}{\sinh(x)}\sin(\xi x)dx-\pi\sin(\beta\pi)\Big\}=0. \tag{9}$$

2. Similarly as in Step 1, we may consider the complex valued function

$$\Psi(z)=\frac{\cosh(\beta z)}{\sinh(z)}e^{i\xi z}.$$

Then Ψ has poles at $z_k=\pm ik\pi$, where $k=0,1,2,\dots$. Consider the contour integral $\int_\Gamma\Psi(z)dz$, where Γ is the same contour as in Step 1 with two indentations $\gamma_{r_1}: z=re^{i\theta}+i\pi$, for $0\geq\theta\geq-\pi$, so as to avoid the pole at $i\pi$, and $\gamma_{r_2}: z=re^{i\theta}$, for $0\geq\theta\geq-\pi$, so as to avoid the pole at 0. By applying Cauchy's theorem, we obtain

$$\int_{-R}^{-r_2}\Psi(x)dx+\int_{\gamma_{r_2}}\Psi(z)dz+\int_{r_2}^{R}\Psi(x)dx+\int_{0}^{\pi}\Psi(R+iy)idy$$
$$+\int_{R}^{r_1}\Psi(x+i\pi)dx+\int_{\gamma_{r_1}}\Psi(z)dz+\int_{-r_1}^{-R}\Psi(x+i\pi)dx$$
$$+\int_{\pi}^{0}\Psi(-R+iy)idy=0.$$

By Jordan's lemma, we get in Step 1

$$\lim_{r_1\to0}\int_{\gamma_{r_1}}\Psi(z)dz=i(-\pi-0)(-\cos(\beta\pi)e^{-\xi\pi})=i\pi\cos(\beta\pi)e^{-\xi\pi}$$

and

$$\lim_{r_2\to0}\int_{\gamma_{r_2}}\Psi(z)dz=i(-\pi-0)(\cosh(0)e^{0})=-i\pi.$$

After similar arguments as in Step 1, with some small changes, by taking limits as $r_2\to0$, $r_1\to0$ and $R\to\infty$, successively, we get

$$2i\int_{0}^{\infty}\frac{\cosh(\beta x)}{\sinh(x)}\sin(\xi x)dx+e^{-\xi\pi}\Big\{2i\cos(\beta\pi)\int_{0}^{\infty}\frac{\cosh(\beta x)}{\sinh(x)}\sin(\xi x)dx$$
$$+2i\sin(\beta\pi)\int_{0}^{\infty}\frac{\sinh(\beta x)}{\sinh(x)}\cos(\xi x)dx+i\pi\cos(\beta\pi)\Big\}-i\pi=0. \tag{10}$$

Let $I=\int_0^\infty\frac{\sinh(\beta x)}{\sinh(x)}\cos(\xi x)dx$, and $J=\int_0^\infty\frac{\cosh(\beta x)}{\sinh(x)}\sin(\xi x)dx$. Then (9) and (10) can be written, successively, as

$$\begin{cases}(2+2e^{-\xi\pi}\cos(\beta\pi))I-2e^{-\xi\pi}\sin(\beta\pi)J-\pi\sin(\beta\pi)e^{-\xi\pi}=0\\ (2+2e^{-\xi\pi}\cos(\beta\pi))J+2e^{-\xi\pi}\sin(\beta\pi)I+\pi\cos(\beta\pi)e^{-\xi\pi}-\pi=0.\end{cases}$$

Solving the above system for I and J, we obtain the desired results.

1. Using Bochner's theorem, the positive definitenesss of the function f can be reduced to showing that the Fourier transform $\hat{f}(\xi)$ is positive. Since f is an even function, its Fourier transform is given by

$$\hat{f}(\xi) = 2\int_0^\infty \frac{x\cosh(\beta x)}{\sinh(x)}\cos(\xi x)dx.$$

The differentiation of the formula (5) in Lemma 1 with respect to β gives

$$\begin{aligned}\int_0^\infty \frac{x\cosh(\beta x)}{\sinh(x)}\cos(\xi x)dx &= \frac{\pi}{2}\frac{\pi\cos(\beta\pi)[\cosh(\xi\pi)+\cos(\beta\pi)] - \sin(\beta\pi)(-\pi\sin(\beta\pi))}{(\cosh(\xi\pi)+\cos(\beta\pi))^2}\\ &= \frac{\pi^2[1+\cos(\beta\pi)\cosh(\xi\pi)]}{2(\cosh(\xi\pi)+\cos(\beta\pi))^2}.\end{aligned}$$

So,

$$\hat{f}(\xi) = \frac{\pi^2[1+\cos(\beta\pi)\cosh(\xi\pi)]}{(\cosh(\xi\pi)+\cos(\beta\pi))^2}.$$

Consequently, if $\frac{-1}{2} \le \beta \le 0$, then $\hat{f}(\xi) \ge 0$. Since ϕ is even in β, the result follows for $\frac{-1}{2} \le \beta \le \frac{1}{2}$.

COROLLARY 1. [illegible] $a, b \ge 0$ [illegible]

$$H_\nu(a,b) << L(a,b) \text{ [illegible] } \frac{1}{4} \le \nu \le \frac{3}{4}. \tag{11}$$

COROLLARY 2. [illegible] A, B [illegible] X [illegible] ν $\frac{1}{4} \le \nu \le \frac{3}{4}$ [illegible]

$$|||A^\nu X B^{1-\nu} + A^{1-\nu} X B^\nu||| \le 2\left|\left|\left|\int_0^1 A^t X B^{1-t} dt\right|\right|\right| \tag{12}$$

[illegible] $|||.|||$

1. We proceed in two steps.

[illegible] 1. By definition, $H_\nu(a,b) << L(a,b)$ if

$$v_{ij} = \left[\frac{H_\nu(\lambda_i,\lambda_j)}{L(\lambda_i,\lambda_j)}\right]_{i,j=1,\dots,n}$$

is positive semidefinite. Put $\lambda_i = e^{x_i}$ and $\lambda_j = e^{x_j}$, with $x_i, x_j \in \mathbb{R}$. Then

$$v_{ij} = \frac{1}{2}e^{\frac{x_i}{2}}\left(\frac{e^{(2\nu-1)(\frac{x_i-x_j}{2})} + e^{(2\nu-1)(\frac{x_j-x_i}{2})}}{e^{\frac{x_i}{2}}(\frac{e^{\frac{x_i-x_j}{2}} - e^{\frac{x_j-x_i}{2}}}{x_i-x_j})e^{\frac{x_j}{2}}}\right)e^{\frac{x_j}{2}}.$$

Thus the matrix $[v_{ij}]$ is congruent to one with entries

$$\frac{(\frac{x_i-x_j}{2})\cosh(\beta(\frac{x_i-x_j}{2}))}{\sinh(\frac{x_i-x_j}{2})},$$

where $\beta = 2\nu - 1$. Hence, the matrix $[v_{ij}]$ is positive semidefinite if and only if the function

$$f(x) = \frac{x\cosh(\beta x)}{\sinh(x)}$$

is positive definite.

2. By Theorem 1, $f(x)$ is positive definite if and only if $-\frac{1}{2} \le \beta \le \frac{1}{2}$, which is equivalent to the condition $\frac{1}{4} \le \nu \le \frac{3}{4}$.

1. The inequality $M << N$ could, in general, be strictly stronger than the usual inequality $M \le N$. That means not every inequality between means of positive numbers leads to a corresponding inequality for positive matrices as shown by the following simple example. For $a, b > 0$ we have

$$H_\alpha(a,b) \le L(a,b) \text{ if and only if } \frac{1 - \frac{1}{\sqrt{3}}}{2} \le \alpha \le \frac{1 + \frac{1}{\sqrt{3}}}{2}. \tag{13}$$

In fact, by taking $a = e^x$ and $b = e^y$ and using Taylor series, it is easy to see that $H_\alpha(a,b) \le L(a,b)$ if and only if

$$\cosh\left((2\alpha - 1)\left(\frac{x-y}{2}\right)\right) \le \frac{\sinh(\frac{x-y}{2})}{\frac{x-y}{2}}.$$

Let $t = \frac{x-y}{2}$, and $\beta = 2\alpha - 1$. Then after simplification

$$1 + \frac{\beta^2 t^2}{2!} + \frac{\beta^4 t^4}{4!} + \cdots \le 1 + \frac{t^2}{3!} + \frac{t^4}{5!} + \cdots.$$

This is true only if $\beta^2 \le \frac{1}{3}$, which leads to the desired result.

2. First assume $A = B$. Since the norms involved are unitarily invariant, we may suppose that A is diagonal with entries $\lambda_1, \lambda_2, \ldots, \lambda_n$. Then we have

$$A^\nu X A^{1-\nu} + A^{1-\nu} X A^\nu = Y \circ \left(\int_0^1 A^t X A^{1-t} dt\right),$$

where Y is the matrix with entries

$$y_{ij} = \frac{2H_\nu(\lambda_i, \lambda_j)}{L(\lambda_i, \lambda_j)}.$$

A well-known result on the Schur multiplier norm (see [12, Theorem 5.5.18 and Theorem 5.5.19]) says that if Y is any positive semidefinite matrix, then for all matrix X,

$$|||Y \circ X||| \le \max_i\{y_{ii}\}|||X|||, \quad \text{for every unitarily invariant norm.} \tag{14}$$

By Corollary 1, Y is a positive semidefinite matrix. Applying (14), we obtain

$$|||A^\nu X A^{1-\nu} + A^{1-\nu} X A^\nu||| \le 2\left|\left|\left|\int_0^1 A^t X A^{1-t} dt\right|\right|\right|. \tag{15}$$

Now, we use the usual trick replacing A and X in the inequality (15) by the 2 by 2 matrices $\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$ and $\begin{pmatrix} 0 & X \\ 0 & 0 \end{pmatrix}$. This gives us the desired inequality (12).

2. Given $a, b > 0$. A natural question arises as to whether the reverse inequality $L(a,b) << H_\nu(a,b)$ is valid.

For $\nu = 0, 1$ we have $L(a,b) << H_\nu(a,b)$ (which is exactly the second part of (4)). On the other hand,

$$L(a,b) \leq H_\nu(a,b)$$

cannot be true for $\nu \in (0,1)$ due to the fact that $f(x) = \frac{\sinh(x)}{x\cosh((2\nu-1)x)}$ goes to infinity as $x \to \pm\infty$. So, f cannot be positive definite.

Acknowledgments. I would like to express my cordial thanks to Kuwait University for their financial support, to Prof. R. Bhatia for the stimulating conversations, and to the referee for his valuable comments.

REFERENCES

[1] R. Bhatia and C. Davis, *More matrix forms of the arithmetic-geometric mean inequality*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 132–136.
[2] R. Bhatia and D. Drissi, *Generalized Lyapunov equations and positive definite functions*, SIAM J. Matrix Anal. Appl., 27 (2005), pp. 103–114.
[3] R. Bhatia and K. R. Parthasarathy, *Positive definite functions and operator inequalities*, Bull. London Math. Soc., 32 (2000), pp. 214–228.
[4] R. Bhatia, *Interpolating the arithmetic-geometric mean inequality and its operator version*, Linear Algebra Appl., 413 (2006), pp. 355–363.
[5] R. Bhatia, *Infinitely divisible matrices*, Amer. Math. Monthly, 113 (2006), pp. 221–235.
[6] R. Bhatia and H. Kosaki, *Mean Matrices and Infinite Divisibilty*, Linear Algebra Appl., in press.
[7] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 6th ed., Academic Press, San Diego, 2000.
[8] E. Heinz, *Beiträge zur Störungstheorie der Spektralzerlegung*, Math. Ann., 123 (1951), pp. 415–438.
[9] F. Hiai and H. Kosaki, *Comparison of various means for operators*, J. Funct. Anal., 163 (1999), pp. 300–323.
[10] F. Hiai and H. Kosaki, *Means for matrices and comparison of their norms*, Indiana Univ. Math. J., 48 (1999), pp. 899–936.
[11] F. Hiai and H. Kosaki, *Means of Hilbert Space Operators*, Lecture Notes in Math. 1820, Springer-Verlag, Berlin, 2003.
[12] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.
[13] H. Kosaki, *Arithmetic-geometric mean and related inequalities for operators*, J. Funct. Anal., 156 (1998), pp. 429–451.
[14] X. Zhan, *Inequalities for unitarily invariant norms*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 466–470.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 3, pp. 829–844

ASYMPTOTICALLY OPTIMAL LOWER BOUNDS FOR THE CONDITION NUMBER OF A REAL VANDERMONDE MATRIX*

REN-CANG LI†

Abstract. Lower bounds on the condition number $\min \kappa_p(V)$ of a real Vandermonde matrix V are established in terms of the dimension n or n and the largest absolute value among all nodes that define the Vandermonde matrix. All bounds here are asymptotically sharp, similar to those in Beckermann (*Numer. Math.*, 85 (2000), pp. 553–577), but bounds here are sharper and cover more cases. Also, qualitative behaviors of $\min \kappa_p(V)$, as well as nearly optimally conditioned real Vandermonde matrices, as functions of the largest absolute value among all nodes are obtained.

Key words. optimal condition number, Vandermonde matrix, Chebyshev polynomials

AMS subject classifications. 15A12, 65F35

DOI. 10.1137/060652737

1. Introduction.

Given n numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$ called , the associated is defined as

$$V \stackrel{\text{def}}{=} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1^{n-1} & \alpha_2^{n-1} & \cdots & \alpha_n^{n-1} \end{pmatrix}. \tag{1.1}$$

It is perhaps one of the best known structured matrices, arising from polynomial interpolation and others [3]. It is invertible if all nodes α_j are distinct, i.e., $\alpha_i \neq \alpha_j$ for $i \neq j$ (Vandermonde matrices are also notoriously known to be ill-conditioned [13, p. 428], [10]). Its condition number can become arbitrarily large, even for modest n. This is not surprising because moving one node arbitrarily close to another will make V arbitrarily close to a singular matrix. Therefore the question of importance about V is not but rather V as far as its condition number is concerned.

Although V is well defined no matter if all or some of α_j are real or complex, this paper is confined to real Vandermonde matrix V only, i.e., α_j . Throughout this paper, some notation is reserved for one assignment, including V and its nodes α_j and $\alpha_{\max} \stackrel{\text{def}}{=} \max_j |\alpha_j|$, along with many others in Table 1.1. V_{sym} is one of those V whose nodes are real symmetric with respect to 0, i.e., $\alpha_i + \alpha_{n-i+1} = 0$.

The major objective of this paper is to bound the ℓ_p-condition number $\kappa_p(V) = \|V\|_p \|V^{-1}\|_p$ from below in terms of n or n and $\alpha_{\max}$. Asymptotically optimal bounds have been established. By we mean those that will give

$$\rho \equiv \text{asymptotic speed} \stackrel{\text{def}}{=} \lim_{n\to\infty} [\min \kappa_p(V)]^{1/n} \tag{1.2}$$

*Received by the editors February 22, 2006; accepted for publication (in revised form) by J. H. Brandts April 18, 2006; published electronically October 4, 2006. Supported in part by the National Science Foundation CAREER award under grant CCR-9875201 and by the National Science Foundation under grant DMS-0510664.

http://www.siam.org/journals/simax/28-3/65273.html

†Department of Mathematics, University of Kentucky, Lexington, KY 40506 (rcli@ms.uky.edu). Current address: Department of Mathematics, University of Texas at Arlington, P.O. Box 19408, Arlington, TX 76019 (rcli@uta.edu).

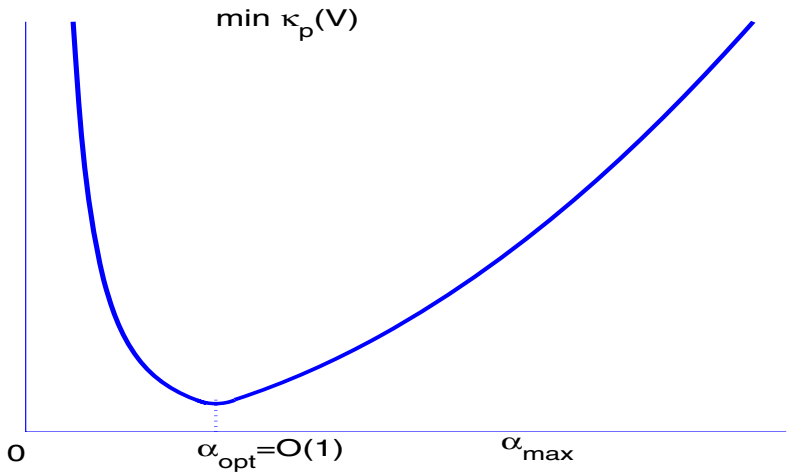


FIG. 1.1. *Qualitative behaviors of* $\min_{\alpha_j} \kappa_p(V)$ *and* $\min_{\alpha_j \ge 0} \kappa_p(V)$ *as* $\alpha_{\max}$ *varies.*

exactly, where min is taken over some prescribed subset or the entire set of real Vandermonde matrices. This is done through establishing bounds like

$$c_1 n^{d_1} \le \min \kappa_p(V)/\,\rho^n \le c_2 n^{d_2}, \tag{1.3}$$

written for short as $\min \kappa_p(V) = \mathcal{O}_n(\rho^n)$, where c_1, c_2, d_1, and d_2 are constants. Particular attention will be given to the case $p = \infty$. In a sense, considering $p = \infty$ is sufficient because of the exponential growth of $\kappa_\infty(V)$ and because

$$n^{-2/p}\kappa_p(V) \le \kappa_\infty(V) \le n^{2/p}\kappa_p(V), \tag{1.4}$$

and thus all $\kappa_p(V)$ have the same asymptotic speed. Nonetheless, whenever it is possible to establish sharper bounds on $\kappa_p(V)$ [illegible] instead of [illegible] through bounds on $\kappa_\infty(V)$ combined with (1.4), we shall go for the sharper ones.

In the past, Gautschi and his coauthor had systematically studied the condition number estimations in [6, 7, 8, 9, 11], where various condition number bounds in terms of the nodes α_j have been established, as well as bounds in terms of the dimension n only. In [11] two lower bounds in terms of n were obtained for positive nodes ($\alpha_j \ge 0$) and real symmetric nodes ($\alpha_j + \alpha_{n+1-j} = 0$). However, bounds in [11] are far from asymptotically optimal. It is Beckermann [2] in 2000 (see also [1]) who obtained asymptotically optimal condition number estimations for all real Vandermonde matrices for the first time.

This paper is based on the technical report [17] which was written before the author came across Beckermann's landmark paper [2]. But we have more detailed and refined analysis and cover more cases, and tighter lower and upper bounds, too. Specifically, the major differences are as follows.

1. We obtain a qualitative plot in Figure 1.1 which shows how $\min_{\alpha_j} \kappa_p(V)$ and $\min_{\alpha_j \ge 0} \kappa_p(V)$ subject to a fixed $\alpha_{\max}$ behave qualitatively as functions of $\alpha_{\max}$. What Figure 1.1 says is that initially as $\alpha_{\max}$ increases, both $\min_{\alpha_j} \kappa_p(V)$ and $\min_{\alpha_j \ge 0} \kappa_p(V)$ decrease until at $\alpha_{\max} = \alpha_{\rm opt}$ when global minimums of $\kappa_p(V)$ are reached, and then they start climbing again. Notice $\alpha_{\rm opt}$ may be different for the two cases, but $\alpha_{\rm opt} = \mathcal{O}(1)$ in both cases.

2. We consider $\min \kappa_p(V)$ under various constraints: (1) all $\alpha_j \in \mathbb{R}$, (2) all $\alpha_j \geq 0$, (3) $\alpha_{\max} = \delta$ or $\alpha_{\max} \leq \delta$ ($\delta \leq 1$ or $\delta > 1$), with or without assuming all $\alpha_j \geq 0$. Essentially only the first two cases were considered in [2], but not the third one which itself has many subcases and is conceivably important in practice. Suppose that we seek polynomial approximations to functions by interpolation on $[\alpha, \beta]$. For the approximations to be any good, most likely, the nodes must be distributed over the entire interval, and in particular $\min \alpha_j \approx \alpha$ and $\beta \approx \max_j \alpha_j$. This will make $\alpha_{\max} \approx \max\{|\alpha|, |\beta|\}$.
3. Our lower and upper bounds are tighter: we have $d_2 - d_1 = 1$ always in (1.3) for $\min \kappa_p(V)$ over real α_j or nonnegative $\alpha_j \geq 0$ (see Remarks 5.1). Although Theorem 4.1 for the same purpose in [2] is for $p = 2$, it was remarked that bounds for the ℓ_p-condition number can also be achieved similarly with[1] $d_2 - d_1 = 2 - 1/p$. Our bounds for $p = \infty$ can be even tighter. In fact, for $p = \infty$ the approach by Beckermann [2] would give $d_2 - d_1 = 2$, while our best results in later sections give $d_2 - d_1 = \sqrt{2}/4$, and therefore smaller upper over low bound ratios for large n; see Tables 5.1 and 6.1.
4. Also for $p = \infty$, we have results that give $d_1 = d_2 = 0$ for $\alpha_{\max} \leq \delta$ or for $\alpha_{\max} \leq \delta$ and all $\alpha_i \geq 0$, where $\delta \leq 1$ is given, while no results as such[2] were presented in [2]; see Tables 5.1 and 6.1. Both in [2] and here it is obtained exactly

$$\rho = 1 + \sqrt{2} \quad \text{for } \min_{\alpha_i} \kappa_p(V), \text{ and} \quad \rho = (1 + \sqrt{2})^2 \quad \text{for } \min_{\alpha_i \geq 0} \kappa_p(V).$$

It is worth mentioning that despite its notorious ill-conditioning, there is a way to compute its singular value decomposition to highly relative accuracy [5, 15], and sometimes very accurate solutions to Vandermonde linear systems [3, 13].

Although our study here does not yield optimally conditioned V, i.e., V that achieve $\min \kappa_p(V)$, it does, however, conclude what nearly optimally conditioned V are for various cases:

1. For nodes in $[-\beta, \beta]$ or for nodes in $[\alpha, \beta]$ with $0 = \alpha < \beta$ (also true for $0 < \alpha$; see [17]), subject to $\alpha_{\max} = \beta$, a nearly optimally conditioned V is the one defined with the translated Chebyshev nodes in a slightly larger interval (so that $\alpha_{\max} = \beta$).
2. If all α_j are allowed to vary freely along the entire real line, a nearly optimally conditioned V is the one defined with Chebyshev nodes (for which $\alpha_{\max} = \cos \frac{\pi}{2n} \approx 1$).
3. If all α_j are forced nonnegative but otherwise free, a nearly optimally conditioned V is the one defined with the translated Chebyshev nodes in the interval $[\alpha, \beta] = [0, 1]$.

Those nearly optimally conditioned V are truly by the word "nearly." That is to say they are just nearly optimal but may not be optimal, according to those few optimally conditioned V computed in [8] under the condition that the optimal V is unique (for any fixed n). Beckermann [2, Theorem 4.1] also implied other nearly optimal conditioned V for the case $\alpha_i \in \mathbb{R}$ or the case $\alpha_i \geq 0$. In particular, Beckermann [1, Theorem 5.9] established that the optimal nodes for $\min_{\alpha_j \geq 0} \kappa_1(V)$ subject to $\alpha_{\max} = \gamma$ are $\alpha_{j+1} = (1 + \cos \frac{j\pi}{n-1})\gamma/2$ for $0 \leq j \leq n - 1$.

[1] V in [2] is V^T here.

[2] As pointed out by an anonymous referee, it is possible to derive asymptotically optimal lower bounds for $\min \kappa_2(V)$ for $-1 \leq \alpha \leq \alpha_j < \beta \leq 1$, using the result about Krylov matrices given in [2, Remarks 3.4 and 3.5], but it was not done explicitly there.

TABLE 1.1
Special notation.

V, α_j, $\alpha_{\max}$	Vandermonde matrix V, its n nodes, and $\alpha_{\max} = \max_j \lvert\alpha_j\rvert$;
$V_{\rm sym}$	V with symmetric nodes: $\alpha_j + \alpha_{n+1-j} = 0$;
$[\alpha, \beta]$	the interval that contains all nodes α_j; see (3.1);
ω, τ	real parameters and whenever there is $[\alpha, \beta]$ in the context, they are defined by (3.2);
$T_n(t)$, $T_n(x;\omega,\tau)$	Chebyshev polynomial, its translation $T_n(x/\omega + \tau)$;
θ_j, t_j	$\theta_j = \frac{2j-1}{2n}\pi$, and $t_j = \cos\theta_j$: zeros of $T_n(t)$, defined by (4.1);
x_j	$x_j = \omega(t_j - \tau)$: zeros of $T_n(x;\omega,\tau)$, defined by (4.2);
$a_{jn} \equiv a_{jn}(\omega,\tau)$	coefficients of $T_n(x;\omega,\tau)$ defined by (2.4);
$S_{n,p}(\omega,\tau)$	$\left(\sum_{j=0}^{n} \lvert a_{jn}\rvert^p\right)^{1/p}$ defined by (2.5).

The rest of this paper is organized as follows. A cornerstone of our study is the use of the absolute sums of coefficients of translated Chebyshev polynomials of the first kind. They are defined and computed for a symmetric interval or a nonnegative interval in section 2. Section 3 proves a general lower bound on $\kappa_p(V)$ with nodes restricted to a given interval $[\alpha, \beta]$. Upper bounds on $\min \kappa_p(V)$ are obtained by the computations for V with the translated Chebyshev nodes. This is done in section 4. Section 5 derives various asymptotically optimal bounds with or without fixing $\alpha_{\max}$, while section 6 considers the case when all $\alpha_j \ge 0$. Finally, section 7 draws a few concluding remarks.

Notation. We shall stick to the global assignments in Table 1.1, unless otherwise explicitly stated. $1 \le p \le +\infty$ and p' is defined by $1/p + 1/p' = 1$. $\mathbb{R}$ is the set of real numbers. $\lceil \xi \rceil$ is the smallest integer that is no less than ξ. For two sequences of numbers a_n and b_n: $a_n \sim b_n$ means $a_n/b_n \to 1$ as $n \to +\infty$; $a_n = \mathcal{O}(b_n)$ means $c_1 \le a_n/b_n \le c_2$ for constants c_1 and c_2; $a_n = \mathcal{O}_n(b_n)$ means $c_1 n^{d_1} \le a_n/b_n \le c_2 n^{d_2}$ for constants c_1, c_2, d_1, and d_2. In this paper, both a_n and b_n grow exponentially in n, and thus the hidden factors n^{d_i} in $a_n = \mathcal{O}_n(b_n)$ are less significant compared to the exponential growth. For notational convenience, by $\min_j$, and $\min_{\alpha_j}$ or min over some constraints on α_j, we mean that j runs from 1 to n.

2. Coefficients of Chebyshev polynomials. The nth Chebyshev polynomial of the first kind is

$$T_n(t) = \cos(n \arccos t) \qquad \text{for } |t| \le 1, \tag{2.1}$$

$$= \frac{1}{2}\left(t + \sqrt{t^2-1}\right)^n + \frac{1}{2}\left(t - \sqrt{t^2-1}\right)^n \qquad \text{for } |t| \ge 1. \tag{2.2}$$

Given real parameters ω and τ, the nth [illegible] is defined by

$$T_n(x;\omega,\tau) \stackrel{\rm def}{=} T_n(x/\omega + \tau). \tag{2.3}$$

Here and in the rest of this paper T_n is overloaded with distinctions according to its argument(s). It can be seen that $T_n(x;\omega,\tau)$ is a polynomial of degree n in x. Write

$$T_n(x;\omega,\tau) = a_{nn}x^n + a_{n-1\,n}x^{n-1} + \cdots + a_{1n}x + a_{0n}, \tag{2.4}$$

where $a_{jn} \equiv a_{jn}(\omega,\tau)$ are functions of ω and τ which, wherever referenced, are all either clear from the context or explicitly stated. Define

$$S_{n,p}(\omega,\tau) \stackrel{\rm def}{=} \left(\sum_{j=0}^{n} |a_{jn}|^p\right)^{1/p}, \tag{2.5}$$

a function of ω and τ, too. Successful computation of $S_{n,p}(\omega,\tau)$ is crucial to our later development. But, in its generality, an explicit formula for $S_{n,p}(\omega,\tau)$ is hard to find. Nevertheless, we still manage to find formulas for $S_{n,1}(\omega,\tau)$ for two different cases $\tau = 0$ and $|\tau| \geq 1$.

THEOREM 2.1.

1. $S_{n,1}(\omega,0) = |T_n(\iota/\omega)|$ [illegible] $\iota = \sqrt{-1}$ [illegible] imaginary unit [illegible][3]

$$S_{n,1}(\omega,0) = |T_n(\iota/\omega)| \sim \frac{1}{2}\left(\frac{1}{|\omega|} + \sqrt{1+\frac{1}{|\omega|^2}}\right)^n.$$

2. [illegible] $|\tau| \geq 1$

$$S_{n,1}(\omega,\tau) = T_n\left(\frac{1}{|\omega|}+|\tau|\right) \sim \frac{1}{2}\left[\left(\frac{1}{|\omega|}+|\tau|\right) + \sqrt{\left(\frac{1}{|\omega|}+|\tau|\right)^2 - 1}\right]^n.$$

For any other p, we may use the inequalities

$$(n+1)^{-1/p'} S_{n,1}(\omega,\tau) \leq S_{n,p}(\omega,\tau) \leq S_{n,1}(\omega,\tau), \tag{2.6}$$

$$\lceil (n+1)/2 \rceil^{-1/p'} S_{n,1}(\omega,0) \leq S_{n,p}(\omega,0) \leq S_{n,1}(\omega,0), \tag{2.7}$$

to get bounds on $S_{n,p}$. Both (2.6) and (2.7) can be proved by using Hölder inequality

$$\sum_{j=1}^{m} |\xi_j \zeta_j| \leq \left(\sum_{j=1}^{m} |\xi_j|^p\right)^{1/p} \left(\sum_{j=1}^{m} |\zeta_j|^{p'}\right)^{1/p'} \tag{2.8}$$

and the fact that $\left(\sum_{j=1}^{m} |\xi_j|^p\right)^{1/p}$ is decreasing in p [12, Lemma 1.1].

THEOREM 2.2. [illegible] $a_n > 0$ [illegible] $\delta = a_n^{1/n}$ [illegible] $a_n \sim cn^\mu$ [illegible] $c > 0$ [illegible] μ [illegible]

$$2S_{n,1}(\delta,0) \sim \frac{(1+\sqrt{2})^n}{(cn^\mu)^{1/\sqrt{2}}}, \quad 2S_{n,1}(\delta/2,1) \sim \frac{(1+\sqrt{2})^{2n}}{(cn^\mu)^{1/\sqrt{2}}}. \tag{2.9}$$

[illegible]. We will prove more general results: [illegible] $(\ln a_n)/n \to 0$ [illegible] $n \to \infty$ [illegible]

$$2S_{n,1}(\delta,0) \sim \frac{(1+\sqrt{2})^n}{a_n^{1/\sqrt{2}}}, \quad 2S_{n,1}(\delta/2,1) \sim \frac{(1+\sqrt{2})^{2n}}{a_n^{1/\sqrt{2}}}. \tag{2.10}$$

Since $a_n \sim cn^\mu$ implies $(\ln a_n)/n \to 0$ as $n \to \infty$, we have (2.9) from (2.10).

The second asymptotical relation in (2.10) follows from the first one because

$$S_{n,1}(\delta/2,1) = |T_n(1+2\delta^{-1})| = |T_{2n}(\iota/\sqrt{\delta})| = S_{2n}(\sqrt{\delta},0)$$

upon noticing that $T_n(2t^2-1) = T_{2n}(t)$, and $\sqrt{\delta} = a_n^{1/(2n)}$. We shall now prove the first relation in (2.10). Notice that $\ln \delta^{-1} \sim -(\ln a_n)/n \equiv \epsilon \Rightarrow \delta^{-1} \sim 1+\epsilon$ to get

$$\delta^{-1} + \sqrt{1+\delta^{-2}} \sim 1 + \epsilon + \sqrt{2}\,(1+\epsilon/2) = (1+\sqrt{2})(1+\epsilon/\sqrt{2}).$$

[3]Going through the proofs in [17], one may see that Theorem 2.1 is valid for complex ω as well. But for the purpose of this paper, ω is real.

Therefore

$$\begin{aligned}\ln\left[2S_{n,1}(\delta,0)\right] &\sim n\ln\left(\delta^{-1}+\sqrt{1+\delta^{-2}}\right)\\ &\sim n\left[\ln(1+\sqrt{2})+\epsilon/\sqrt{2}\right]\\ &= \ln(1+\sqrt{2}\,)^n-(\ln a_n)/\sqrt{2},\end{aligned}$$

which gives the first asymptotical relation in (2.10). □

3. A general lower bound on condition numbers of Vandermonde matrices. Given $1\le p\le\infty$, the ℓ_p-norm of vector $u=(\mu_1,\mu_2,\ldots,\mu_n)^T$ is defined as $\|u\|_p=\left(\sum_{j=1}^n|\mu_j|^p\right)^{1/p}$, and $\|u\|_\infty=\lim_{p\to\infty}\|u\|_p=\max_j|\mu_j|$. The associated ℓ_p-operator norm of an $m\times n$ matrix A is defined as $\|A\|_p=\max_{u\ne0}\|Au\|_p/\|u\|_p$. It can be proved that $\|A\|_p=\|A^T\|_{p'}$, upon noticing that

$$\|A\|_p=\max_{u\ne0,v\ne0}\frac{|v^TAu|}{\|v\|_{p'}\|u\|_p},$$

where $1/p+1/p'=1$ (see also [16]). Superscript "$\cdot^T$" takes the transpose of a matrix or a vector.

We shall start by establishing a general lower bound on $\kappa_p(V)$ for

$$\alpha\le\min_j\alpha_j\le\max_j\alpha_j\le\beta. \tag{3.1}$$

The case $\alpha=\beta$ is of no interest because then V is of rank 1 and thus $\kappa_p(V)=+\infty$ (unless $n=1$). There are many ways to realize (3.1), and it is tempting to always let $\alpha=\min_j\alpha_j$ and $\beta=\max_j\alpha_j$, but that may not always be possible for theorems that require $-\alpha=\beta$. Recall ω and τ defined by (3.2), and let $\alpha_{\max}\stackrel{\rm def}{=}\max_j|\alpha_j|$. Set

$$\omega=\frac{\beta-\alpha}{2}>0,\quad \tau=-\frac{\beta+\alpha}{\beta-\alpha}. \tag{3.2}$$

The linear transformation $t=x/\omega+\tau$ maps $x\in[\alpha,\beta]$ one-to-one and onto $t\in[-1,1]$.

LEMMA 3.1.

$$\max\{n,n\alpha_{\max}^{n-1}\}\ge\|V\|_p\ge\max\{n^{1/p'},\alpha_{\max}^{n-1}\}, \tag{3.3}$$

$$\|V^{-1}\|_p\ge\frac{S_{n-1,p'}(\omega,\tau)}{n^{1/p'}}\ge\begin{cases}\left(\dfrac{n}{\lceil n/2\rceil}\right)^{1/p}\dfrac{S_{n-1,1}(\omega,0)}{n} & \text{if } -\alpha=\beta,\\ \dfrac{S_{n-1,1}(\omega,\tau)}{n} & \end{cases} \tag{3.4}$$

Let e_j be the jth column of the $n\times n$ identity matrix. Then

$$\|V\|_p=\|V^T\|_{p'}\ge\begin{cases}\|V^Te_1\|_{p'}=n^{1/p'},\\ \|V^Te_n\|_{p'}\ge\alpha_{\max}^{n-1}.\end{cases}$$

This yields the second inequality in (3.3). The known formulas for $\|\cdot\|_1$ and $\|\cdot\|_\infty$ [4, page 22] yield $\|V\|_1$, $\|V\|_\infty\le\max\{n,n\alpha_{\max}^{n-1}\}$ and now use [14, page 29]

$$\|V\|_p\le\|V\|_\infty^{1/p'}\|V\|_1^{1/p}$$

to arrive at the first inequality in (3.3).

We now show (3.4). Let v be the vector of the coefficients of $T_{n-1}(x;\omega,\tau) \equiv T_{n-1}(x/\omega+\tau)$, i.e., $v = (a_{0,n-1}\ a_{1,n-1}\ \cdots\ a_{n-1,n-1})^T$. Then

$$V^T v = (T_{n-1}(\alpha_1/\omega+\tau)\ T_{n-1}(\alpha_2/\omega+\tau)\ \cdots\ T_{n-1}(\alpha_n/\omega+\tau))^T,$$

which yields $\|V^T v\|_{p'} \le n^{1/p'}$ because $|T_{n-1}(x/\omega+\tau)| \le 1$ for $x \in [\alpha,\beta]$. We therefore have

$$\|V^{-1}\|_p = \|V^{-T}\|_{p'} \ge \frac{\|v\|_{p'}}{\|V^T v\|_{p'}} \ge \frac{S_{n-1,p'}(\omega,\tau)}{n^{1/p'}}.$$

This is the first inequality in (3.4). Use it, together with (2.6) and (2.7), to get the second inequality. ∎

THEOREM 3.2.

$$\kappa_p(V) \ge \max\left\{ S_{n-1,p'}(\omega,\tau), \frac{\alpha_{\max}^{n-1} S_{n-1,p'}(\omega,\tau)}{n^{1/p'}} \right\} \tag{3.5}$$

$$\ge \max\left\{ \frac{S_{n-1,1}(\omega,\tau)}{n^{1/p}}, \frac{\alpha_{\max}^{n-1} S_{n-1,1}(\omega,\tau)}{n} \right\}. \tag{3.6}$$

Proof. This theorem is an immediate consequence of Lemma 3.1. ∎

This is the most general theorem of this paper for a lower bound on $\kappa_p(V)$. It is its various applications combined with results in section 4 that lead to many interesting asymptotically optimal lower bounds. There are at least two different ways to apply Theorem 3.2 to any given V:

1. Take $\alpha = \min_j \alpha_j$ and $\beta = \max_j \alpha_j$ and then compute the right-hand side of (3.5) or (3.6). But unless $\alpha \ge 0$ or $-\alpha = \beta$, we may have to compute $S_{n-1,1}(\omega,\tau)$ by its definition (2.5) because no explicit formula has yet been found. In this case, both α and β are nodes of V.
2. Take $-\alpha = \beta = \alpha_{\max}$ (and thus $\omega = \alpha_{\max}$ and $\tau = 0$) and then use the explicit formula for $S_{n-1,1}(\alpha_{\max},0)$ to compute the right-hand side of (3.6). In this case, one of α and β is guaranteed to be a node for V.

REMARK 3.1. The lower bounds in [2] were essentially obtained as follows. Let $\omega = \eta\alpha_{\max}$. It follows from $\|V\|_p \ge \max_j \|Ve_j\|_p = \left(\sum_{j=0}^{n-1} \alpha_{\max}^{jp}\right)^{1/p}$ and (3.4) that

$$n^{1/p'}\kappa_p(V) \ge \left(\sum_{j=0}^{n-1} \alpha_{\max}^{jp}\right)^{1/p} S_{n-1,p'}(\omega,\tau).$$

But $S_{n-1,p'}(\omega,\tau) = \left(\sum_j |\omega^{-j} a_{j\,n-1}(1,\tau)|^{p'}\right)^{1/p'}$. By Hölder inequality (2.8), we have $n^{1/p'}\kappa_p(V) \ge \sum_j \eta^{-j}|a_{j\,n-1}(1,\tau)| = S_{n-1,1}(\eta,\tau)$ which gives

$$\kappa_p(V) \ge S_{n-1,1}(\eta,\tau)/\,n^{1/p'}. \tag{3.7}$$

In the case of [2], $p = p' = 2$, either $\eta = 1$ and $\tau = 0$ or $\eta = 1/2$ and $\tau = -1$. This is a pretty decent bound, but it partially collapses the interval information, unlike (3.5) and (3.6) which form the basis for us to eventually arrive at the qualitative behaviors in Figure 1.1.

4. Vandermonde matrices with translated Chebyshev nodes. The zeros of $T_n(t)$ are called

$$t_j = \cos\theta_j \quad \theta_j = \tfrac{2j-1}{2n}\pi \ (1 \le j \le n), \tag{4.1}$$

and the zeros of the translated Chebyshev polynomial $T_n(x;\omega,\tau)$ as in (2.3) are called

$$x_j = \omega(t_j - \tau) \ (1 \le j \le n). \tag{4.2}$$

This section, inspired by Gautschi [7], computes $\kappa_\infty(V)$ for V with the translated Chebyshev nodes for the case $-\alpha = \beta$ and the case $0 \le \alpha < \beta$. But we are still unsure how to deal with the general case $\alpha < 0 < \beta$, $-\alpha \ne \beta$. Recall ω and τ defined in (3.2).

First we compute $\|V\|_\infty$ for V with $\alpha_j = x_j = \omega(\cos\theta_j - \tau)$. This is relatively easy. By [8, Theorem 2.1],

$$\|V\|_\infty = \max\left\{ n, \sum_{j=1}^{n} |\alpha_j|^{n-1} \right\} = \max\left\{ n, \omega^{n-1}\Lambda_n(\tau) \right\}, \tag{4.3}$$

where

$$\Lambda_n(\tau) \stackrel{\text{def}}{=} \sum_{j=1}^{n} |\cos\theta_j - \tau|^{n-1}. \tag{4.4}$$

It can be seen that $\Lambda_n(-\tau) = \Lambda_n(\tau)$. In [17, Appendix B], the following asymptotical behaviors

$$\Lambda_n(0) \sim \sqrt{\frac{2n}{\pi}}, \quad \Lambda_n(1) \sim \sqrt{\frac{n}{\pi}}\, 2^{n-1}. \tag{4.5}$$

were obtained. With (4.5), we have the following theorem.

THEOREM 4.1. [illegible] $\alpha_j = x_j$ $(1 \le j \le n)$ [illegible] (4.2) [illegible] (3.2) [illegible]

$$\|V\|_\infty \sim \max\left\{ n, \sqrt{\frac{2n}{\pi}}\,\omega^{n-1} \right\} \sim \max\left\{ n, \sqrt{\frac{2n}{\pi}}\,\alpha_{\max}^{n-1} \right\} \quad \text{[illegible]} \ -\alpha = \beta > 0,$$

$$\|V\|_\infty \sim \max\left\{ n, \sqrt{\frac{n}{\pi}}\,\beta^{n-1} \right\} \sim \max\left\{ n, \sqrt{\frac{n}{\pi}}\,\alpha_{\max}^{n-1} \right\} \quad \text{[illegible]} \ 0 = \alpha < \beta.$$

In both cases $-\alpha = \beta$ or $0 = \alpha < \beta$, $\sum_{j=1}^{n} |x_j|^{n-1} = \mathcal{O}(\sqrt{n}\,\alpha_{\max}^{n-1})$. But will this also be true for arbitrary interval $[\alpha,\beta]$? We do not know.

We now estimate $\|V^{-1}\|_\infty$ with translated Chebyshev nodes. It is made possible by Gautschi's formulas for $\|V^{-1}\|_\infty$ for V with symmetric nodes or with nonnegative nodes [7]. We have the following theorem.

THEOREM 4.2 (see [17]). [illegible] $\alpha_j = x_j$ $(1 \le j \le n)$ [illegible] (4.2) [illegible] (3.2) [illegible]

$$\omega \min\left\{1, \frac{1+\omega}{1+\omega^2}\right\} \frac{1}{n} \le \frac{\|V^{-1}\|_\infty}{S_{n,1}(\omega,0)} \le \omega \max\left\{1, \frac{1+\omega}{1+\omega^2}\right\} \frac{3^{3/4}}{2n} \quad \text{[illegible]} \ -\alpha = \beta > 0, \tag{4.6}$$

$$\frac{\frac{\beta-\alpha}{2}\cos\frac{\pi}{2n}}{n\left(1+\frac{\beta+\alpha}{2}\right)} \le \frac{\|V^{-1}\|_\infty}{S_{n,1}(\omega,\tau)} \le \frac{\beta-\alpha}{2n\sqrt{(1+\beta)(1+\alpha)}} \quad \text{[illegible]} \ 0 = \alpha < \beta, \tag{4.7}$$

(4.6) $n \geq 3$ V V_{sym} $-\alpha = \beta$

Theorem 4.2 says that for V with translated Chebyshev nodes on $[\alpha, \beta]$, if $-\alpha = \beta$ or $0 \leq \alpha < \beta$ or $\alpha < \beta \leq 0$, then

$$\frac{n\|V^{-1}\|_\infty}{S_{n,1}(\omega,\tau)} = \mathcal{O}(1). \tag{4.8}$$

(The case $[\alpha, \beta]$ for $\alpha < \beta \leq 0$ can be turned into $[-\beta, -\alpha]$, a case that is covered by Theorem 4.2.) But what happens when $\alpha < 0 < \beta$ and $-\alpha \neq \beta$? Is (4.8) still true? We conjecture it would be, but do not have any proof for now.

THEOREM 4.3. $\alpha_j = x_j$ $(1 \leq j \leq n)$ (4.2) (3.2)

$$\min_{-\alpha=\beta} \kappa_\infty(V) \leq \frac{3^{3/4}}{2}\beta_{\text{opt}}\, S_{n,1}(\beta_{\text{opt}}, 0) \sim \frac{3^{3/4}}{2}\left(\frac{2}{\pi}\right)^{\sqrt{2}/4} \frac{(1+\sqrt{2})^n}{2n^{\sqrt{2}/4}} \quad -\alpha = \beta > 0,$$

$$\min_{0=\alpha<\beta} \kappa_\infty(V) \leq \frac{\beta^+_{\text{opt}}}{2\sqrt{1+\beta^+_{\text{opt}}}}\, S_{n,1}(\beta^+_{\text{opt}}/2, 1) \sim \frac{\sqrt{2}\,(1+\sqrt{2})^{2n}}{4(n\pi)^{\sqrt{2}/4}} \quad 0 = \alpha < \beta,$$

$\beta_{\text{opt}} \equiv \omega_{\text{opt}} = (n/\Lambda_n(0))^{1/(n-1)} \sim 1$ $\beta^+_{\text{opt}}/2 \equiv \omega^+_{\text{opt}} = (n/\Lambda_n(1))^{1/(n-1)} \sim 1/2$

. A proof can be found in [17], and the asymptotic relations can be achieved by applying Theorem 2.2. □

5. Condition numbers for V with $\alpha_i \in [\alpha, \beta]$ and $-\alpha = \beta$. In this section, we shall establish lower and upper bounds on

$$\min \kappa_p(V) \text{ subject to } \cdots$$

$\alpha_j \in \mathbb{R}$	$\alpha_{\max} \leq \delta$ or $\alpha_{\max} = \delta$
Theorems 5.3, 5.3′	Theorems 5.4, 5.4′

where for each type of minimization we have two versions of bounds—one for all p (Theorems 5.3 and 5.4) and one just for $p = \infty$ (Theorems 5.3′ and 5.4′, sharper at least asymptotically than by just setting $p = \infty$ in the other version).

LEMMA 5.1.

1. $|\omega|$ $S_{n,p}(\omega,\tau)$ $|\omega|^n S_{n,p}(\omega,\tau)$
2. $\omega S_{n,1}(\omega, 0)$ ω $\omega \leq \max\{\sqrt{n-1}, \sqrt{2}\}$ n

. For item 1, we notice that $[S_{n,p}(\omega,\tau)]^p$ is a polynomial in $|\omega|^{-p}$ while $[|\omega|^n S_{n,p}(\omega,\tau)]^p$ is a polynomial in $|\omega|^p$. Item 2 is proved in [17]. □

LEMMA 5.2.

$$\kappa_p(V) \geq \max\left\{ S_{n-1,p'}(\alpha_{\max}, 0), \frac{\alpha^{n-1}_{\max} S_{n-1,p'}(\alpha_{\max}, 0)}{n^{1/p'}} \right\} \tag{5.1}$$

$$= \begin{cases} S_{n-1,p'}(\alpha_{\max}, 0) & \alpha_{\max} \leq n^{1/[p'(n-1)]}, \\ \alpha^{n-1}_{\max} S_{n-1,p'}(\alpha_{\max}, 0)/\, n^{1/p'} & \alpha_{\max} > n^{1/[p'(n-1)]}. \end{cases} \tag{5.2}$$

. Apply Theorem 3.2 to the case $-\alpha = \beta = \alpha_{\max}$ (and thus $\omega = \alpha_{\max}$ and $\tau = 0$) to get (5.1). By Lemma 5.1, the first quantity within $\max\{\cdots\}$ in (5.1) is decreasing in $\alpha_{\max}$, while the second one is increasing in $\alpha_{\max}$. Therefore the right-hand side of (5.1) achieves its minimum when the two are equal, i.e., $n^{1/p'} = \alpha^{n-1}_{\max}$, which yields (5.2). □

THEOREM 5.3.

$$S_{n-1,p'}(n^{1/[p'(n-1)]},0) \le \min_{\alpha_j} \kappa_p(V) \le \min_{\alpha_j} \kappa_p(V_{\rm sym}) \le n^{1/p}\frac{3^{3/4}}{2}\, S_{n,1}(1,0). \tag{5.3}$$

PROOF. The right-hand side of (5.1), as a function of $\alpha_{\max}$, achieves its minimum at $\alpha_{\max} = n^{1/[p'(n-1)]}$. That gives the first inequality. The second inequality is true because $\{V_{\rm sym}\}$ is a subset of all Vandermonde matrices. We now prove the third one. To this end, consider $V = V_{\rm sym}$ with Chebyshev nodes t_j as in (4.1). Then $\|V\|_p \le n$ by (3.3). Apply Theorem 4.2 to the case $-\alpha = \beta = 1 = \omega$ to get $\|V^{-1}\|_\infty \le \frac{3^{3/4}}{2} \cdot n^{-1} S_{n,1}(1,0)$ and then to get

$$\|V^{-1}\|_p \le n^{1/p}\|V^{-1}\|_\infty \le n^{1/p}\frac{3^{3/4}}{2} \cdot n^{-1} S_{n,1}(1,0).$$

So for this $V_{\rm sym}$, $\kappa_p(V_{\rm sym}) \le n^{1/p}\frac{3^{3/4}}{2} \cdot S_{n,1}(1,0)$, as needed. □

We include $\min_{\alpha_j} \kappa_p(V_{\rm sym})$ in (5.3) mainly because Vandermonde matrices with symmetric nodes were heavily studied by Gautschi [7, 8] and Gautschi and Ingese [11]. Moreover, assuming that the optimally conditioned V is unique, Gautschi [8] showed that the optimally conditioned V must have symmetric nodes.

REMARK 5.1. Upon using (3.7) with $\eta = 1$ and $\tau = 0$, we have

$$S_{n-1,1}(1,0)/\, n^{1/p'} \le \min_{\alpha_j} \kappa_p(V) \le n^{1/p}\frac{3^{3/4}}{2}\, S_{n,1}(1,0), \tag{5.4}$$

which differs from (5.3) only in the leftmost inequalities. The left inequality in (5.4) is due to [2] for $p = 2$, and it is less sharp than the left inequality in (5.3) at least for $p = \infty$ because, by Theorem 2.2,

$$S_{n-1,1}(n^{1/(n-1)},0) \sim \frac{(1+\sqrt{2})^{n-1}}{2\, n^{1/\sqrt{2}}}, \quad \frac{S_{n-1,1}(1,0)}{n} \sim \frac{(1+\sqrt{2})^{n-1}}{2n}.$$

Even so, for any $1 \le p \le \infty$, the ratio of the upper bound in (5.4) over the lower bound is $n\,\frac{3^{3/4}}{2}$, and it gives $\min_{\alpha_j} \kappa_p(V)$ a lower and upper bound like (1.3) with $d_2 - d_1 = 1$, while similar lower and upper bounds in [2] for the same purpose are with $d_2 - d_1 = 2 - 1/p$.

The third inequality in (5.3) was proved by simply picking a special V with Chebyshev nodes. This turns out to be good enough, as we shall see later, in yielding the correct asymptotic speed in our notation $\mathcal{O}_n$, but it does not produce the best possible factor n^d hidden in the notation. For $p = \infty$, however, a tighter upper bound is possible by using the V with the translated Chebyshev nodes in $[-\beta_{\rm opt}, \beta_{\rm opt}]$, where $\beta_{\rm opt} = (n/\Lambda_n(0))^{1/(n-1)}$ as in Theorem 4.3. Of course, one may use this V for all p, but doing so will not only lead to a more complicated bound but also the resulted bound may not be much better due to more complicated estimation of $\|V\|_p$. For this reason, we shall state a sharper version of Theorem 5.3 for $p = \infty$ only as a consequence of Theorem 4.3. The upper bound in (5.5) is sharper because of $1 \sim \beta_{\rm opt} > 1$ and item 2 in Lemma 5.1.

THEOREM 5.3′. Let $\beta_{\rm opt} = (n/\Lambda_n(0))^{1/(n-1)}$ where $\Lambda_n(1)$ is as in (4.4)

$$S_{n-1,1}(n^{1/(n-1)},0) \le \min_{\alpha_j} \kappa_\infty(V) \le \min_{\alpha_j} \kappa_\infty(V_{\rm sym}) \le \frac{3^{3/4}}{2}\beta_{\rm opt}\, S_{n,1}(\beta_{\rm opt},0). \tag{5.5}$$

In what follows, we shall establish theorems that are in the same spirit as Theorems 5.3 and 5.3′, but with $\alpha_{\max}$ subject to a constraint.

THEOREM 5.4. [illegible] $\delta > 0$ [illegible] $\delta' = \delta/\cos\frac{\pi}{2n}$ [illegible] $\delta \le 1$ [illegible]

$$(5.6)\quad S_{n-1,p'}(\delta,0) \le \min_{\alpha_{\max}\le\delta} \kappa_p(V) \le \min_{\alpha_{\max}=\delta} \kappa_p(V) \le n^{1/p}\frac{(\sqrt{2}+1)3^{3/4}}{4}\,\delta'\,S_{n,1}(\delta',0).$$

[illegible] $\delta > 1$ [illegible]

$$(5.7)\quad \frac{\delta^{n-1}S_{n-1,p'}(\delta,0)}{n^{p'}} \le \min_{\alpha_{\max}=\delta} \kappa_p(V) \le n^{1/p}\frac{3^{3/4}\left[\cos\frac{\pi}{2n}\right]^{n-1}}{2}\,(\delta')^n S_{n,1}(\delta',0),$$

$$(5.8)\quad S_{n-1,p'}(n^{1/[p'(n-1)]},0) \le \min_{\alpha_{\max}\le\delta} \kappa_p(V) \le n^{1/p}\frac{3^{3/4}}{2}\cdot S_{n,1}(1,0).$$

[illegible] (5.6) (5.7) [illegible] (5.8) [illegible] V [illegible] V_{sym} [illegible]. (1) Observe that $\{V : \alpha_{\max} = \delta\} \subset \{V : \alpha_{\max} \le \delta\}$ to get the middle inequality in (5.6).

(2) Lemma 3.1 also implies that

$$(5.9)\quad \kappa_p(V) \ge \begin{cases} S_{n-1,p'}(\omega,\tau) & \text{if } \alpha_{\max} \le 1, \\ \alpha_{\max}^{n-1} S_{n-1,p'}(\omega,\tau)/\,n^{1/p'} & \text{if } \alpha_{\max} > 1 \end{cases}$$

upon noticing that $\|V\|_p \ge n^{1/p'}$ if $\alpha_{\max} \le 1$, and $\|V\|_p \ge \alpha_{\max}^{n-1}$ if $\alpha_{\max} > 1$. Apply it to the case $-\alpha = \beta = \alpha_{\max} \le \delta \le 1$ (and thus $\omega = \alpha_{\max}$ and $\tau = 0$) to obtain $\kappa_p(V) \ge S_{n-1,p'}(\alpha_{\max},0) \ge S_{n-1,p'}(\delta,0)$ by Lemma 5.1. This gives the first inequality in (5.6).

(3) Apply (5.9) to the case $-\alpha = \beta = \delta = \alpha_{\max}$ to obtain the first inequality in (5.7).

(4) Take $-\alpha = \beta = \delta/\cos\frac{\pi}{2n} = \delta'$, and $\alpha_j = x_j$ $(1 \le j \le n)$, the translated Chebyshev nodes as in (4.2). Then

$$\tau = 0, \quad \alpha_{\max} = \max|\alpha_j| = \beta\cos\frac{\pi}{2n} = \delta, \quad \delta \le \omega = \beta = \delta'.$$

Theorem 4.2 says that for the V with those nodes

$$\frac{\|V^{-1}\|_\infty}{S_{n,1}(\omega,0)} \le \omega\max\left\{1, \frac{1+\omega}{1+\omega^2}\right\}\frac{3^{3/4}}{2n} \le \begin{cases} \delta'\dfrac{(\sqrt{2}+1)3^{3/4}}{4n} & \text{if } \delta \le 1, \\ \delta'\dfrac{3^{3/4}}{2n} & \text{if } \delta \ge 1, \end{cases}$$

where we have used

$$(5.10)\quad \max_{\omega>0}\left\{1, \frac{1+\omega}{1+\omega^2}\right\} = \frac{1+\omega}{1+\omega^2}\bigg|_{\omega=\sqrt{2}-1} = \frac{\sqrt{2}+1}{2}, \quad \max_{\omega\ge1}\left\{1, \frac{1+\omega}{1+\omega^2}\right\} = 1.$$

Now employ $\|V\|_p \le n$ if $\delta \le 1$, $\|V\|_p \le n\delta^{n-1}$ if $\delta \ge 1$, and $\|V^{-1}\|_p \le n^{1/p}\|V^{-1}\|_\infty$ to get the last inequalities in (5.6) and in (5.7).

(5) A proof of (5.8) can be done in the same way as for Theorem 5.3.

(6) Finally, when V is replaced by $V_{\rm sym}$, the first inequalities in (5.6), (5.7), and (5.8) still hold. The middle inequality in (5.6) also remains valid. The last inequalities in (5.6), (5.7), and (5.8) hold because they all were proved by bounding some $\kappa_p(V_{\rm sym})$. □

There are stronger versions of (5.7) and (5.8) for $p=\infty$, too, just as we did for Theorem 5.3.

THEOREM 5.4′. [illegible] $\delta>1$ [illegible] $\delta'=\delta/\cos\frac{\pi}{2n}$ [illegible]

$$\frac{\delta^{n-1}S_{n-1,1}(\delta,0)}{n}\le\min_{\alpha_{\max}=\delta}\kappa_\infty(V)\le\frac{3^{3/4}}{2}\frac{\max\{n,\Lambda_n(0)(\delta')^{n-1}\}}{n}\delta'\,S_{n,1}(\delta',0), \tag{5.11}$$

$$S_{n-1,1}(n^{1/(n-1)},0)\le\min_{\alpha_{\max}\le\delta}\kappa_\infty(V)\le\frac{3^{3/4}}{2}\omega_1\,S_{n,1}(\omega_1,0), \tag{5.12}$$

[illegible] $\omega_1=\min\{\delta',(n/\Lambda_n(0))^{1/(n-1)}\}$ [illegible] $\omega_1=(n/\Lambda_n(0))^{1/(n-1)}$ [illegible] n [illegible] (5.11) [illegible] (5.12) [illegible] V [illegible] $V_{\rm sym}$ [illegible]. Only the second inequalities in (5.11) and (5.12) need proofs. For (5.11), it follows from the proof of Theorem 5.4, upon using $\|V\|_\infty=\max\{n,\omega^{n-1}\Lambda_n(0)\}$ which for large n is proportional to $\sqrt{n}\,\delta^{n-1}$, better than $\|V\|_\infty\le n\delta^{n-1}$. The second inequality in (5.12) is obtained by minimizing

$$\frac{3^{3/4}}{2}\frac{\max\{n\alpha'_{\max}\,S_{n,1}(\alpha'_{\max},0),\Lambda_n(0)(\alpha'_{\max})^n S_{n,1}(\alpha'_{\max},0)\}}{n}$$

subject to $\alpha_{\max}\le\delta$, where $\alpha'_{\max}=\alpha_{\max}/\cos\frac{\pi}{2n}$. This minimization is solved by noticing Lemma 5.1, which says the first quantity within $\max\{\cdots\}$ is decreasing in $\alpha_{\max}$ when $\alpha_{\max}\le\max\{\sqrt{n-1},\sqrt{2}\}$, while the second one is increasing in $\alpha_{\max}$. That $\omega_1=(n/\Lambda_n(0))^{1/(n-1)}$ for n sufficiently large is due to $(n/\Lambda_n(0))^{1/(n-1)}\sim(2\pi/n)^{1/[2(n-1)]}\sim 1$. □

We shall now investigate the tightness of the upper and the lower bounds we have established so far, as well as the asymptotical speeds of $\kappa_p(V)$ minimized over a certain set of Vandermonde matrices. For this purpose, Li [17] obtained Table 5.1 for the asymptotical behaviors of the ratios of the upper bounds over the corresponding lower bounds. This table is for $p=\infty$. (For any other p, $S_{n-1,p'}$ in the lower bounds will have to be weakened by using (2.7) so as to apply the same lines of arguments in [17].)

Given that $S_{n,1}(\delta,0)$ goes to $+\infty$ exponentially as $n\to+\infty$, our upper bounds and the lower bounds in Theorems 5.3, 5.3′, 5.4, and 5.4′ are very tight. These bounds, together with Lemma 5.1, lead to the qualitative behavior of $\min_{\alpha_j}\kappa_p(V)$ as $\alpha_{\max}$ varies, depicted in Figure 1.1. Examining how we got the upper bounds by these inequalities, we conclude that

(5.13) [illegible] $\alpha_{\max}$ [illegible] V [illegible] $[-\alpha_{\max},\alpha_{\max}]$ [illegible] $\pm\alpha_{\max}$ [illegible]

In addition to Table 5.1, Li [17] also obtained the following corollary on the asymptotical speeds of $\min\kappa_\infty(V)$ as functions of n for various cases.

TABLE 5.1
Ratios of the upper bounds over the lower bounds for $p = \infty$.

$\min \kappa_\infty(V)$ subject to $\cdots$	Ratio (asymptotically dominant term)	Ineq.
$\alpha_j \in \mathbb{R}$	$\frac{(1+\sqrt{2})3^{3/4}}{2} \times n^{1/\sqrt{2}}$	(5.3)
	$\frac{(1+\sqrt{2})3^{3/4}}{2} \left(\frac{2}{\pi}\right)^{\sqrt{2}/4} \times n^{\sqrt{2}/4}$	(5.5)
$\alpha_{\max} \le \delta$ or $\alpha_{\max} = \delta$ for $\delta \le 1$	$\frac{(1+\sqrt{2})3^{3/4}}{2} (1 + \sqrt{\delta^2+1}\,) \times n^0$	(5.6)
$\alpha_{\max} = \delta$ for $\delta > 1$	$\frac{3^{3/4}}{2}(1 + \sqrt{\delta^2+1}\,) \times n^1$	(5.7)
	$\frac{3^{3/4}}{2}\sqrt{\frac{\pi}{2}}(1 + \sqrt{\delta^2+1}\,) \times n^{1/2}$	(5.11)
$\alpha_{\max} \le \delta$ for $\delta > 1$	$\frac{(1+\sqrt{2})3^{3/4}}{2} \times n^{1/\sqrt{2}}$	(5.8)
	$\frac{(1+\sqrt{2})3^{3/4}}{2} \left(\frac{2}{\pi}\right)^{\sqrt{2}/4} \times n^{\sqrt{2}/4}$	(5.12)

COROLLARY 5.5. [illegible] $\delta > 0$ [illegible]

$$\min_{\alpha_j} \kappa_\infty(V) = \mathcal{O}_n\left((1+\sqrt{2}\,)^n\right), \tag{5.14}$$

$$\min_{\alpha_{\max} \le \delta} \kappa_\infty(V), \quad \min_{\alpha_{\max} = \delta} \kappa_\infty(V) = \mathcal{O}\left((\delta^{-1} + \sqrt{1+\delta^{-2}}\,)^n\right) \quad \text{[illegible]}\ \delta \le 1, \tag{5.15}$$

$$\min_{\alpha_{\max} = \delta} \kappa_\infty(V) = \mathcal{O}_n\left((1+\sqrt{1+\delta^2}\,)^n\right) \quad \text{[illegible]}\ \delta > 1, \tag{5.16}$$

$$\min_{\alpha_{\max} \le \delta} \kappa_\infty(V) = \mathcal{O}_n\left((1+\sqrt{2}\,)^n\right) \quad \text{[illegible]}\ \delta > 1. \tag{5.17}$$

[illegible] (5.14)–(5.17) [illegible] V [illegible] V_{sym}

This is a very informative corollary; for example,

$$\min_{\alpha_{\max} \le 1/2} \kappa_\infty(V), \ \min_{\alpha_{\max} = 1/2} \kappa_\infty(V) = \mathcal{O}\left((2+\sqrt{5}\,)^n\right),$$
$$\min_{\alpha_{\max} = 2} \kappa_\infty(V) = \mathcal{O}_n\left((1+\sqrt{5}\,)^n\right).$$

It is worth mentioning that (5.15) is in terms of $\mathcal{O}$, while all other equations in Corollary 5.5 are in terms of $\mathcal{O}_n$.

6. Condition numbers for V with $\alpha_i \in [\alpha, \beta]$ and $0 \le \alpha < \beta$. Notice that the case $\alpha < \beta \le 0$ can be turned into this case by reversing the signs of all α_j while leaving $\|V\|_p$ and $\|V^{-1}\|_p$ unchanged. So the results in what follows apply to the case $\alpha < \beta \le 0$ as well after minor modifications. We shall present lower and upper bounds on

$\min \kappa_p(V)$ subject to $\cdots$

$\alpha_j \ge 0$	$\alpha_j \ge 0$, $\alpha_{\max} \le \delta$ or $\alpha_{\max} = \delta$
Theorems 6.2, 6.2′	Theorems 6.3, 6.3′

Most developments here are parallel to those in the previous section. Proofs share similar lines of arguments as well and thus will be omitted. Also omitted here are various results for the case $0 < \alpha < \beta$, except (6.1) below. The interested reader may find omitted proofs and results in [17].

LEMMA 6.1. [illegible] $\alpha_j \ge 0$ [illegible] $\alpha = \min_j \alpha_j$ [illegible] $\beta = \max_j \alpha_j$ [illegible]

$$\kappa_p(V) \ge \max\left\{S_{n-1,p'}(\omega,\tau),\ \alpha_{\max}^{n-1} S_{n-1,p'}(\omega,\tau)\big/\, n^{1/p'}\right\} \tag{6.1}$$

$$\ge \max\left\{S_{n-1,p'}\left(\alpha_{\max}/2,1\right)\ \alpha_{\max}^{n-1} S_{n-1,p'}\left(\alpha_{\max}/2,1\right)\big/\, n^{1/p'}\right\} \tag{6.2}$$

$$= \begin{cases} S_{n-1,p'}\left(\alpha_{\max}/2,1\right) & \text{if } \alpha_{\max} \le n^{1/[p'(n-1)]}, \\ \alpha_{\max}^{n-1} S_{n-1,p'}\left(\alpha_{\max}/2,1\right)\big/\, n^{1/p'} & \text{if } \alpha_{\max} > n^{1/[p'(n-1)]}. \end{cases} \tag{6.3}$$

THEOREM 6.2.

$$S_{n-1,p'}\left(n^{1/[p'(n-1)]}\Big/\, 2,1\right) \le \min_{\alpha_j \ge 0} \kappa_p(V) \le n^{1/p}\frac{\sqrt{2}}{4}\, S_{n,1}(1/2,1). \tag{6.4}$$

THEOREM 6.2′. [illegible] $\beta_{\rm opt}^{+} = 2\left(n/\Lambda_n(1)\right)^{1/(n-1)}$ [illegible] $\Lambda_n(1)$ [illegible] (4.4) [illegible]

$$S_{n-1,1}\left(n^{1/(n-1)}\Big/\, 2,1\right) \le \min_{\alpha_j \ge 0} \kappa_\infty(V) \le \frac{\beta_{\rm opt}^{+}}{2\sqrt{1+\beta_{\rm opt}^{+}}}\, S_{n,1}(\beta_{\rm opt}^{+}/2,1). \tag{6.5}$$

THEOREM 6.3. [illegible] $\delta > 0$ [illegible] $\delta' = [2/(1+c)]\delta \ge \delta$ [illegible] $c = \cos\frac{\pi}{2n}$ [illegible] $\delta < 1$ [illegible]

$$S_{n-1,p'}(\delta/2,1) \le \min_{0\le\alpha_j\le\delta} \kappa_p(V) \le \min_{0\le\alpha_j,\alpha_{\max}=\delta} \kappa_p(V) \le \frac{n^{1/p}\delta'}{2\sqrt{1+\delta'}} S_{n,1}(\delta'/2,1). \tag{6.6}$$

[illegible] $\delta > 1$ [illegible]

$$\frac{\delta^{n-1} S_{n-1,p'}(\delta/2,1)}{n^{1/p'}} \le \min_{0\le\alpha_j,\alpha_{\max}=\delta} \kappa_p(V) \le \left(\frac{1+c}{2}\right)^{n-1} \frac{n^{1/p}(\delta')^n}{2\sqrt{1+\delta'}}\, S_{n,1}(\delta'/2,1), \tag{6.7}$$

$$S_{n-1,p'}\left(\frac{n^{1/[p'(n-1)]}}{2},1\right) \le \min_{0\le\alpha_j\le\delta} \kappa_p(V) \le n^{1/p}\,\frac{\sqrt{2}}{4}\, S_{n,1}(1/2,1). \tag{6.8}$$

THEOREM 6.3′. [illegible] $\delta > 1$ [illegible] $\delta' = [2/(1+c)]\delta \ge \delta$ [illegible] $c = \cos\frac{\pi}{2n}$ [illegible]

$$\frac{\delta^{n-1} S_{n-1,1}(\delta/2,1)}{n} \le \min_{0\le\alpha_j,\alpha_{\max}=\delta} \kappa_\infty(V) \le \frac{\max\{n, 2^{-(n-1)}\Lambda_n(1)(\delta')^{n-1}\}}{n} \times \frac{\delta'}{2\sqrt{1+\delta'}}\, S_{n,1}(\delta'/2,1), \tag{6.9}$$

$$S_{n-1,1}\left(n^{1/(n-1)}/2,1\right) \le \min_{0\le\alpha_j\le\delta} \kappa_\infty(V) \le \frac{\delta_1}{2\sqrt{1+\delta_1}}\, S_{n,1}(\delta_1/2,1), \tag{6.10}$$

[illegible] $\delta_1 = \min\{\delta', 2\left(n/\Lambda_n(1)\right)^{1/(n-1)}\}$ [illegible] $\delta_1 = 2\left(n/\Lambda_n(1)\right)^{1/(n-1)}$ [illegible] n [illegible]

Table 6.1 from [17] lists the asymptotically dominant terms for the ratios of the upper bounds over the lower bounds. The conclusion is that these bounds are very

TABLE 6.1
Ratios of the upper bounds over the lower bounds for $p = \infty$ and nonnegative nodes.

$\min \kappa_\infty(V)$ subject to $\cdots$	Ratio (asymptotically dominant term)	Ineq.
$\alpha_j \geq 0$	$\frac{4+3\sqrt{2}}{4} \times n^{1/\sqrt{2}}$	(6.4)
	$\frac{\sqrt{2}}{4}\pi^{\sqrt{2}/4} \times n^{\sqrt{2}/4}$	(6.5)
$0 \leq \alpha_j$, $\alpha_{\max} = \delta$ or $\leq \delta$ for $\delta \leq 1$	$\frac{(1+\sqrt{1+\delta})^2}{2\sqrt{1+\delta}} \times n^0$	(6.6)
$0 \leq \alpha_j$, $\alpha_{\max} = \delta$ for $\delta > 1$	$\frac{(1+\sqrt{1+\delta})^2}{2\sqrt{1+\delta}} \times n^1$	(6.7)
	$\frac{(1+\sqrt{1+\delta})^2}{2\sqrt{1+\delta}} \frac{1}{\pi} \times n^{1/2}$	(6.9)
$0 \leq \alpha_j$, $\alpha_{\max} \leq \delta$ for $\delta > 1$	$\frac{4+3\sqrt{2}}{4} \times n^{1/\sqrt{2}}$	(6.8)
	$\frac{\sqrt{2}}{4}\pi^{\sqrt{2}/4} \times n^{\sqrt{2}/4}$	(6.10)

tight. These bounds, together with Lemma 5.1, lead to the qualitative behavior of $\min_{\alpha_j \geq 0} \kappa_\infty(V)$ as $\alpha_{\max}$ varies, depicted in Figure 1.1. Also, proofs in [17] yield

(6.11) $\alpha_{\max}$ V $[0, \alpha_{\max}]$ $\alpha_{\max}$

As was pointed out in section 1, optimal nodes with respect to $\min_{\alpha_j \geq 0} \kappa_1(V)$ subject to $\alpha_{\max} = \gamma$ were obtained by [1, Theorem 5.9].

COROLLARY 6.4. $0 < \delta$

$$\min_{0 \leq \alpha_j \leq \delta} \kappa_\infty(V), \min_{0 \leq \alpha_j, \alpha_{\max} = \delta} \kappa_\infty(V) = \mathcal{O}\left(\left[\delta^{-1/2} + 1 + \sqrt{1+\delta^{-1}}\right]^{2n}\right), \quad \delta \leq 1,$$

$$\min_{0 \leq \alpha_j, \alpha_{\max} = \delta} \kappa_\infty(V) = \mathcal{O}_n\left((1+\sqrt{1+\delta})^{2n}\right), \quad \delta > 1,$$

$$\min_{0 \leq \alpha_j \leq \delta} \kappa_\infty(V) = \mathcal{O}_n\left((3+2\sqrt{2})^n\right), \quad \delta > 1,$$

$$\min_{\alpha_j \geq 0} \kappa_\infty(V) = \mathcal{O}_n\left((3+2\sqrt{2})^n\right).$$

7. Concluding remarks. We have obtained a series of lower and upper bounds on the optimal condition number $\min \kappa_p(V)$ of real Vandermonde matrices. These bounds are proved to be asymptotically optimal, except possibly the one in Theorem 3.2 in the case when interval $[\alpha, \beta]$ is one of the following three kinds: (1) symmetrical ($-\alpha = \beta$), (2) nonnegative ($\alpha = 0$), (3) nonpositive ($\beta = 0$). Asymptotically optimal bounds have been established for the case $\alpha > 0$ and the case $\beta < 0$ (too [17]).

Our results led us to deduce the qualitative behaviors of optimally conditioned Vandermonde matrices as the largest absolute value $\alpha_{\max}$ of all nodes varies, as shown in Figure 1.1 at the beginning of this paper. Our proofs yielded nearly optimally conditioned Vandermonde matrices in various circumstances.

Similar bounds, though unclear about their asymptotical optimality, have been established, too, for confluent Vandermonde matrices [18].

Acknowledgment. The author wishes to thank an anonymous referee for his constructive suggestions that improved and shortened the paper considerably, and for bringing the identity $T_n(2t^2-1) = T_{2n}(t)$ to the author's attention.

REFERENCES

[1] B. Beckermann, *On the numerical condition of polynomial bases: Estimates for the condition number of Vandermonde, Krylov and Hankel matrices*, Habilitationsschrift, Universität Hannover, Germany, http://math.univ-lille1.fr/∼bbecker/abstract/Habilitationsschrift_Beckermann.pdf, 1996.

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

[3] Å. Björck and V. Pereyra, *Solution of Vandermonde systems of equations*, Math. Comp., 24 (1970), pp. 893–903.

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

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

[6] W. Gautschi, *On inverses of Vandermonde and confluent Vandermonde matrices*, Numer. Math., 4 (1962), pp. 117–123.

[7] W. Gautschi, *Norm estimates for inverses of Vandermonde matrices*, Numer. Math., 23 (1975), pp. 337–347.

[8] W. Gautschi, *Optimally conditioned Vandermonde matrices*, Numer. Math., 24 (1975), pp. 1–12.

[9] W. Gautschi, *On inverses of Vandermonde and confluent Vandermonde matrices* III, Numer. Math., 29 (1978), pp. 445–450.

[10] W. Gautschi, *How (un)stable are Vandermonde systems?*, in Asymptotic and Computational Analysis, R. Wong, ed., Lecture Notes in Pure and Appl. Math. 124, Dekker, New York, 1990, pp. 193–210.

[11] W. Gautschi and G. Ingese, *Lower bounds for the condition number of Vandermonde matrices*, Numer. Math., 52 (1988), pp. 241–250.

[12] M. Goldberg and E. G. Straus, *Multiplicativity of l_p norms for matrices*, Linear Algebra Appl., 52–53 (1983), pp. 351–360.

[13] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.

[14] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, Berlin, 1976.

[15] P. Koev, *Accurate and Efficient Computations with Structured Matrices*, Ph.D. thesis, University of California at Berkeley, Berkeley, CA, 2002.

[16] R.-C. Li, *Norms of certain matrices with applications to variations of the spectra of matrices and matrix pencils*, Linear Algebra Appl., 182 (1993), pp. 199–234.

[17] R.-C. Li, *Asymptotically Optimal Lower Bounds for the Condition Number of a Real Vandermonde Matrix*, Technical report 2004-05, Department of Mathematics, University of Kentucky, Lexington, KY, 2004. Available at http://www.ms.uky.edu/∼math/MAreport/.

[18] R.-C. Li, *Lower bounds for the condition number of a real confluent Vandermonde matrix*, Math. Comp., (2006), to appear.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 3, pp. 845–870

FILTERED CONJUGATE RESIDUAL-TYPE ALGORITHMS WITH APPLICATIONS*

YOUSEF SAAD†

Abstract. It is often necessary to filter out an eigenspace of a given matrix A before performing certain computations with it. The eigenspace usually corresponds to undesired eigenvalues in the underlying application. One such application is in information retrieval, where the method of latent semantic indexing replaces the original matrix with a lower-rank one using tools based on the singular value decomposition. Here the low-rank approximation to the original matrix is used to analyze similarities with a given query vector. Filtering has the effect of yielding the most relevant part of the desired solution while discarding noise and redundancies in the underlying problem. Another common application is to compute an invariant subspace of a symmetric matrix associated with eigenvalues in a given interval. In this case, it is necessary to filter out eigenvalues that are not in the interval of the wanted eigenvalues. This paper presents a few conjugate gradient–like methods to provide solutions to these types of problems by iterative procedures which utilize only matrix-vector products.

Key words. conjugate residual, conjugate gradient, polynomial filtering, principal component analysis, interior eigenvalues

AMS subject classifications. 65F10, 65F20, 65F50

DOI. 10.1137/060648945

1. Introduction.

A number of applications in science and engineering require filtering, a process by which a matrix A is replaced by a function $\phi(A)$, where the filter function ϕ has the desirable property of filtering out certain unwanted eigenvalues. For example, this arises when computing the vector $A_k b$, where A_k is a rank-k approximation to A, and b a certain vector. Typically, A_k is the rank-k approximation that is the closest to A in the 2-norm sense, and it can be obtained from the singular value decomposition (SVD) of A. These methods include the techniques based on principal component analysis (PCA), such as, for example, latent semantic indexing (LSI); see [6].

Classical methods based on the SVD consist of approximating A by a rank-k matrix obtained by retaining only the k largest singular values in the SVD. For example, if $A = U\Sigma V^T$ is the SVD of A, where U and V are unitary and Σ is diagonal, then methods based on PCA replace A by $A_k = U\Sigma_k V^T$, where Σ_k is obtained from Σ by setting all singular values $\sigma_i < \sigma_k$ to zero. This truncated SVD (TSVD) technique amounts to replacing Ab by $s(A)b$, where $s(\lambda)$ is a step function that has value 1 for $\lambda \geq \sigma_k$ and zero for $\lambda < \sigma_k$. An obvious limitation of the SVD-based approach is its excessive computational cost for large matrices since in principle, at least, a complete SVD factorization of A is required.

Another important use of filtering is when computing large invariant subspaces. Here, one can think of a Lanczos-type procedure applied to the matrix $p(A)$ instead of A, where p is a low-degree polynomial. This approach has been successfully used

*Received by the editors January 3, 2006; accepted for publication (in revised form) by L. Reichel March 6, 2006; published electronically October 16, 2006. This work was supported by NSF grants ACI-0305120, DMR-0325218, and DMS 0510131; by the DOE under grant DE-FG02-03ER25585; and by the Minnesota Supercomputing Institute.

http://www.siam.org/journals/simax/28-3/64894.html

†Department of Computer Science and Engineering, University of Minnesota, 200 Union Street S.E., Minneapolis, MN 55455 (saad@cs.umn.edu).

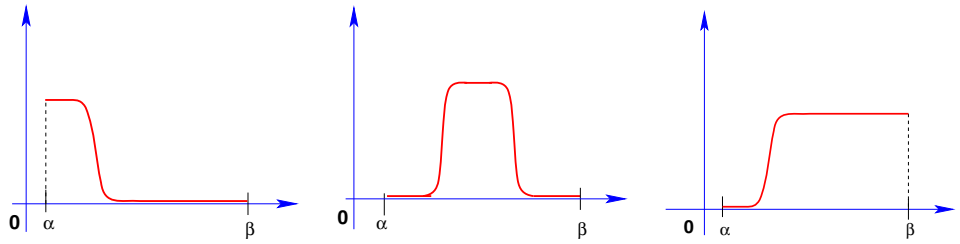


FIG. 1. *Low-pass (left), middle-pass (center), and high-pass (right) filter functions.*

in an application related to quantum mechanics [4], where large invariant subspaces associated with the lowest part of the spectrum are required. A rationale and some details on this approach are given in section 3. An emphasis is placed on invariant subspaces associated with middle eigenvalues, as this was not covered in [4].

The algorithms to be described in the next sections are based on polynomial filtering. Typically, a smooth "base filter" ϕ is selected, and then a sequence of least-squares polynomial approximations to this base function is constructed, from which a sequence of approximate solutions is extracted. One of the main goals of this paper is to express these approximations in a form which resembles the well-known conjugate gradient (CG) or conjugate residual (CR) algorithms. The algorithms to be described can be used to compute solutions of various problems in numerical linear algebra. As an example, by selecting the filter function ϕ to be the exponential function, the algorithms will yield a method for computing approximations to $\exp(A)b$.

2. Polynomial filtering. Given a filter function ϕ, a symmetric matrix A, and a vector b, the problem addressed in this paper is to formulate algorithms for computing approximations of the form $p(A)b$ to the filtered vector $\phi(A)b$, where p is a polynomial. Specifically, we are interested in CG-type algorithms for finding "best" approximations to $\phi(A)b$.

This paper considers only three cases for the filter function ϕ: (a) A high-pass filter function, (b) a low-pass filter, and (c) a middle-pass filter. These three cases are illustrated in Figure 1. A high-pass (low-pass) filter function is one that is close to 1 for large (resp., small) eigenvalues and close to zero for small (resp., large) eigenvalues. A middle-pass filter is close to 1 for eigenvalues in a certain interval of the spectrum and close to zero elsewhere.

Consider a low-pass filter ϕ. We seek to approximate $\phi(A)b$ by vectors of the form $\rho(A)b$, where ρ is a polynomial. We will assume that $\phi(0) = 1$ so that ρ can be sought to also satisfy $\rho(0) = 1$. Thus the polynomial is selected to be in the form

$$\rho(\lambda) = 1 - \lambda s(\lambda). \tag{1}$$

In fact, ρ has the form of residual polynomials used in standard iterative methods such as the CG iteration. We will still often use the term "residual" polynomial for ρ, noting that there is not really a linear system to solve. We would like to minimize, for a certain norm $\|.\|_w$, the difference

$$\|\phi(A)b - \rho(A)b\|_w \tag{2}$$

over all polynomials ρ of degree $\leq k$, such as $\rho(0) = 1$.

The solution corresponding to the case of a high-pass filter can be trivially obtained from that of a low-pass filter. When ψ is a high-pass filter, then we can

minimize $\|(1-\psi)-\rho(\lambda)\|_w$, where now $\phi \equiv 1-\psi$ is a low-pass filter. The best approximation to ψ is $1-\rho(\lambda) = \lambda s(\lambda)$, and therefore the vector $As(A)b$, where ρ minimizes (2), will be the desired solution in this case.

In summary, given a low-pass filter function ϕ, we seek s so that $\|\phi(\lambda) - (1 - \lambda\ s(\lambda))\|_w$ is small—as measured by a certain norm w. The polynomial $\rho(\lambda) = 1 - \lambda s(\lambda)$ approximates the filter ϕ. This will also yield the minimum for $\|(1-\phi) - \lambda\ s(\lambda)\|_w$ with respect to $\lambda s(\lambda)$. The focus is now on the $\lambda s(\lambda)$, and the same process makes this polynomial close to the function

$$\psi \equiv 1 - \phi \tag{3}$$

a high-pass filter, which we will refer to as the complementary filter of ϕ. The case of the middle-pass filter, which will be addressed in detail in section 3.1, resembles the case of the high-pass filter since the filter function is such that $\phi(0) = 0$.

There are many applications of filtering. Low-pass filters are of interest when computing invariant subspaces associated with all eigenvalues $\leq \alpha$. For example, the Lanczos algorithm can be used on the matrix $\rho(A)$, where ρ is a low-degree polynomial with the property that $\rho(\lambda)$ is small for $\lambda > \alpha$. This can be generalized to other situations, and thus in fact eigenspaces associated with eigenvalues located in arbitrary sub-intervals of the spectrum can be computed with the help of filtering. Filtering can also be used to solve highly ill-conditioned linear systems by regularization, but this is not considered in this paper.

An important application is in information retrieval, where one seeks to compute a matrix-vector product $A_k b$, where A_k is a low-rank approximation to A. Here A is not necessarily a square matrix, and we wish to find an approximation of Ab which is accurate in the dominant singular space. This is the situation in LSI, apart from the fact that it is the transpose of A that is considered instead of A. If $A = U\Sigma V^T$ is the SVD of A, then calculating a solution of the form $A\phi(A^TA)b$, we find that

$$\begin{aligned} A\phi(A^TA)b &= (U\Sigma V^T)V\phi(\Sigma^T\Sigma)V^Tb \\ &= U\Sigma\phi(\Sigma^T\Sigma)V^Tb. \end{aligned}$$

The requirement is that $\Sigma\phi(\Sigma^T\Sigma)$ be close to Σ for large σ_i's and close to zero for small σ_i's. So this situation can be handled by a low-pass filter. In [19], a method of this type was used by exploiting expansions of the desired polynomial in a basis of orthogonal polynomials.

2.1. Polynomial filters. In this section, we focus on the problem of filtered iterations. We begin with some notation as well as a rationale for the approach to be taken. We consider a CG-like (actually CR-like) method which uses an arbitrary inner product of functions. The main reason why we seek to write the solution algorithm by exploiting the CR/CG framework is that we already know some of the good algorithmic properties of these methods. In particular, the solution and residual vectors are available at each step, and the solution vector at step k is easily updated from the solution vector at step $k-1$. The numerical properties of the algorithms are also well understood, both in practice and in theory.

Recall that the approximate solution vector obtained at the jth step of a Krylov subspace method is of the form $x_0 + s_j(A)r_0$, where s_j is a polynomial of degree $\leq j$. The corresponding residual vector is $\rho_{j+1}(\lambda) = 1 - \lambda s_j(\lambda)$. This polynomial is of degree $j+1$. It has value 1 at $\lambda = 0$, and it approximates a low-pass filter function ψ.

2.2. CR algorithms in polynomial spaces. In the standard CR algorithm, the solution polynomial s_j minimizes $\|(I - As(A))r_0\|_2$ over all polynomials s of degree $\leq j$. This is nothing but a discrete least-squares norm when expressed in the eigenbasis. Indeed, if the eigenexpansion of r_0 is $r_0 \equiv \sum_{i=1}^{n} \omega_i u_i$, then

$$\|(I - As(A))r_0\|_2 = \left[\sum_{i=1}^{n} \omega_i^2 (1 - \lambda_i s(\lambda_i))^2\right]^{1/2} \equiv \|1 - \lambda s(\lambda)\|_w.$$

It is possible to write a CR-like algorithm which minimizes $\|1 - \lambda s(\lambda)\|_w$ for any 2-norm associated with a (proper) inner product over polynomial spaces:

$$\langle p, q\rangle_w \ .$$

The generic algorithm is given below for reference.

ALGORITHM 2.1.

0. $r_0 := b - Ax_0$ $p_0 := r_0$ $\pi_0 = \rho_0 = 1$
1. $\lambda\pi_0$
2. $j = 0, 1, \ldots,$
3. $\alpha_j := \langle \rho_j, \lambda\rho_j\rangle_w / \langle \lambda\pi_j, \lambda\pi_j\rangle_w$
4. $x_{j+1} := x_j + \alpha_j p_j$
5. $r_{j+1} := r_j - \alpha_j A p_j$ $\rho_{j+1} = \rho_j - \alpha_j \lambda\pi_j$
6. $\beta_j := \langle \rho_{j+1}, \lambda\rho_{j+1}\rangle_w / \langle \rho_j, \lambda\rho_j\rangle_w$
7. $p_{j+1} := r_{j+1} + \beta_j p_j$ $\pi_{j+1} := \rho_{j+1} + \beta_j \pi_j$
8. $\lambda\pi_{j+1}$
9.

It is easy to show that the residual polynomial ρ_j generated by this algorithm minimizes $\|\rho(\lambda)\|_w$ among all polynomials of the form $\rho(\lambda) = 1 - \lambda s(\lambda)$, where s is any polynomial of degree $\leq j - 1$. In other words, ρ_j minimizes $\|\rho(\lambda)\|_w$ among all polynomials ρ of degree $\leq j$ such that $\rho(0) = 1$. It is also easy to show that the polynomials $\lambda\pi_j$ are orthogonal to each other; i.e., $\langle \pi_i, \pi_j\rangle = 0$ for $i \neq j$.

PROPOSITION 2.1. x_{j+1} j 2.1 $x_{j+1} = x_0 + s_j(A)r_0$ s_j j

$$s_j(\lambda) = \alpha_0\pi_0(\lambda) + \cdots + \alpha_j\pi_j(\lambda). \tag{4}$$

π_j $\rho_{j+1}(\lambda)$

$$\langle \lambda\pi_j(\lambda), \lambda\pi_i(\lambda)\rangle_w = \langle \lambda\rho_j(\lambda), \rho_i(\lambda)\rangle_w = 0 \quad i \neq j. \tag{5}$$

$\rho_{j+1} = 1 - \lambda s_j(\lambda)$ $\|1 - \lambda s(\lambda)\|_w$ s $\leq j$

A formal proof is not necessary, but one can exploit the analogy with the usual CR algorithm. In CR (see, e.g., [24]), it is known that the vectors Ap_j are orthogonal to each other. Writing a member of the affine Krylov subspace $x_0 + K_j$ as $x = x_0 + s(A)r_0$, where the degree of s is $\leq j$, the vectors r_{j+1} minimize the 2-norm of all residuals $b - Ax = r_0 - As(A)r_0$ for x in $x_0 + K_j$.

It is useful to comment on implementation aspects. In the usual CR algorithm (see [24]) we would compute Ap_{j+1} in line 8 using the relation which follows from line 7,

$$Ap_{j+1} = Ar_{j+1} + \beta_j Ap_j,$$

in order to avoid an additional matrix-vector product. The vector Ar_{j+1} is computed after line 5 (and saved for the next step to get α_{j+1}), and Ap_{j+1} is then obtained from it using the above formula. Generally, this needs to be done in the situation when the computation of the scalar α_j in line 3 requires the vector Ap_j as well as the vector Ar_j. In the very first step, p and r are the same, so computing Ap_0 in line 1 will suffice. Thereafter, it is necessary to compute Ar_j (before line 3) and update Ap_{j+1}, as was just explained. This strategy is not necessary here because the updates and computations of polynomials require relatively few operations.

We would like to modify the algorithm shown above in order to incorporate filtering. As it is written the algorithm does not lend itself to filtering. Indeed, filtering amounts to minimizing some norm of $\phi(\lambda)-(1-\lambda s(\lambda))$, where ϕ is the filter function, and one must remember that $\phi(A)v$ may be practically difficult to evaluate for a given vector v. In particular, $\phi(A)r_0$ may not be available.

We omit the discussion of CG-type iterations—but it is clear that a CG algorithm in polynomial space can also be written. The residual polynomials will be orthogonal, while the π_js will be conjugate ($\langle \lambda\pi_j, \pi_i\rangle_w = 0$ for $i \neq j$).

2.3. Filtered CR polynomial iterations. Given a certain filter function ϕ, the method to be described in this section consists of finding an approximate solution x_j whose residual polynomial $\rho_j(\lambda)$ approximates the function ϕ, in the least-squares sense. Throughout this section, we consider the dual viewpoint, which is that we are given a filter ψ which is close to zero for λ near zero and close to 1 for large eigenvalues. To make the computation tractable, the function ψ will be chosen to be a piecewise continuous function, though this is not an essential requirement. This will be discussed in more detail in section 2.5. In mathematical terms, we seek a polynomial $s_j(\lambda)$ such that

$$\|\psi(\lambda) - \lambda s_j(\lambda)\|_w = \min_{s \in \mathcal{P}_j} \|\psi(\lambda) - \lambda s(\lambda)\|_w. \tag{6}$$

Here $\mathcal{P}_j$ represents the space of polynomials of degree $\leq j$, and the w-norm is associated with an inner product of the form

$$\langle p, q\rangle_w = \int_0^\beta p(\lambda)q(\lambda)w(\lambda)d\lambda.$$

Note that the left bound of the interval is taken to be zero without loss of generality. For the sake of clarity, the discussion of the choice of the weight function is deferred to a later section. For now, all that needs to be said is that w is selected primarily to enable an easy computation of an inner product of any two functions involved in the algorithms, without resorting to numerical integration.

The condition for the polynomial s_j to be the solution to (6) is that

$$\langle \psi(\lambda) - \lambda s_j(\lambda), \lambda q(\lambda)\rangle_w = 0 \quad \forall q \in \mathcal{P}_j.$$

In order to construct the sequence of approximate solutions, we can generate the sequence of polynomials of the form $\lambda\pi_j$ which are orthogonal. The sequence satisfies a three-term recurrence, and the approximation can be directly expressed in this basis. This was the approach taken in [11, 19].

As a slight alternative, we can try to proceed as in the CR algorithm by updating s_j from s_{j-1} as

$$s_j(\lambda) = s_{j-1}(\lambda) + \alpha_j\pi_j(\lambda). \tag{7}$$

The scalar α_j can be obtained by expressing the condition that $\psi(\lambda) - \lambda s_j(\lambda)$ is orthogonal to $\lambda\pi_j(\lambda)$, or $\langle \psi(\lambda) - \lambda s_j(\lambda), \lambda\pi_j(\lambda)\rangle_w = 0$, which, with the use of (7), leads to

$$\alpha_j = \frac{\langle \psi(\lambda) - \lambda s_{j-1}(\lambda), \lambda\pi_j(\lambda)\rangle_w}{\langle \lambda\pi_j(\lambda), \lambda\pi_j(\lambda)\rangle_w} . \tag{8}$$

The orthogonality of the set $\{\lambda\pi_i\}$ can be exploited to observe that $\lambda s_{j-1}(\lambda)$ is orthogonal to $\lambda\pi_j$. In the end the above expression simplifies to

$$\alpha_j = \frac{\langle \psi(\lambda), \lambda\pi_j(\lambda)\rangle_w}{\langle \lambda\pi_j(\lambda), \lambda\pi_j(\lambda)\rangle_w} . \tag{9}$$

This is a different expression from that obtained from the usual CR algorithm. However, it is possible to express it differently, and this will be explored later for a different algorithm.

After α_j is computed in this manner, we proceed to update the solution x_j and the residual vector r_{j+1} as in steps 4 and 5 of Algorithm 2.1. The polynomial ρ_{j+1} is also updated accordingly. Next, we must compute π_{j+1}. In the usual CG and CR algorithms, π_{j+1} is computed in the form $\pi_{j+1}(\lambda) = \rho_{j+1}(\lambda) + \beta_j\pi_j(\lambda)$, but this will not work here because such an expression exploits the orthogonality of ρ_{j+1} against all $\lambda\pi_i$'s with $i \le j$, which is no longer satisfied. Instead, we could just use a Stieljes-type procedure of the form

$$\beta_{j+1}\pi_{j+1}(\lambda) = \lambda\pi_j(\lambda) - \eta_j\pi_j(\lambda) - \beta_j\pi_j(\lambda).$$

Note that $-\alpha_j\lambda\pi_j(\lambda) = \rho_{j+1}(\lambda) - \rho_j(\lambda)$, and so, if we need the leading coefficients of π_{j+1} and ρ_{j+1} to be the same, we can use the formula

$$\pi_{j+1}(\lambda) = -\alpha_j\left[\lambda\pi_j(\lambda) - \eta_j\pi_j(\lambda) - \beta_j\pi_{j-1}(\lambda)\right] \tag{10}$$

and select the scalars η_j and β_j to make $\lambda\pi_{j+1}$ orthogonal to both $\lambda\pi_j$ and $\lambda\pi_{j-1}$. Assume by induction that the $\lambda\pi_i(\lambda)$'s are orthogonal for $i \le j$. Then, we find that

$$\eta_j = \frac{\langle \lambda^2\pi_j, \lambda\pi_j\rangle_w}{\langle \lambda\pi_j, \lambda\pi_j\rangle_w} \quad \text{and} \quad \beta_j = \frac{\langle \lambda^2\pi_j, \lambda\pi_{j-1}\rangle_w}{\langle \lambda\pi_{j-1}, \lambda\pi_{j-1}\rangle_w} .$$

ALGORITHM 2.2.

0. [illegible] $r_0 := b - Ax_0$ $p_0 := r_0$ $\pi_0 = \rho_0 = 1$
1. [illegible] $\lambda\pi_0$ $\lambda^2\pi_0$
2. [illegible] $j = 0, 1, \ldots,$ [illegible]
3. $\alpha_j := \frac{\langle \psi, \lambda\pi_j\rangle_w}{\langle \lambda\pi_j, \lambda\pi_j\rangle_w}$
4. $x_{j+1} := x_j + \alpha_j p_j$
5. $r_{j+1} := r_j - \alpha_j A p_j$ $\rho_{j+1} = \rho_j - \alpha_j\lambda\pi_j$
6. $\eta_j := \frac{\langle \lambda^2\pi_j, \lambda\pi_j\rangle_w}{\langle \lambda\pi_j, \lambda\pi_j\rangle_w}$ $\beta_j := \frac{\langle \lambda^2\pi_j, \lambda\pi_{j-1}\rangle_w}{\langle \lambda\pi_{j-1}, \lambda\pi_{j-1}\rangle_w}$
7. $p_{j+1} := -\alpha_j[Ap_j - \eta_j p_j - \beta_j p_{j-1}]$ $\pi_{j+1} := -\alpha_j\left[\lambda\pi_j - \eta_j\pi_j - \beta_j\pi_{j-1}\right]$
8. [illegible] $\lambda\pi_{j+1}, \lambda^2\pi_{j+1}$
9. [illegible]

This approach is a slight variation of the one presented in [11, 19]. The main difference is that the algorithms in [11, 19] focus on the solution polynomial instead of the residual polynomial; i.e., they do not explicitly compute or exploit residual

polynomials. However, the two algorithms are mathematically equivalent. Note that when $\psi(\lambda) \equiv 1$, the algorithm should give the same iterates (and same auxiliary vectors) as those of Algorithm 2.1 in exact arithmetic.

The polynomials $\lambda\pi_j$ are orthogonal by construction. On the other hand, the residual polynomials ρ_j do not satisfy any orthogonality relation, but optimality implies that $\langle \psi - \lambda s_j(\lambda), \lambda\pi_i\rangle_w = 0$ for $i \le j$, so we have (recall that $\phi \equiv 1 - \psi$)

$$\langle \phi - \rho_{j+1}, \lambda\pi_i\rangle_w \;=\; 0, \quad i \le j.$$

2.4. Corrected CR algorithm. We now consider an alternative implementation of the above algorithm, which can be viewed as a corrected version of the standard CR algorithm. The derivation is based on the following observation. After line 5 of Algorithm 2.2, the residual vector r_{j+1} is no longer used. This particular residual vector is not all that useful since a convergence test cannot employ it. It would have been more meaningful to compute $[\psi(A) - As(A)]b$, but this is not practically computable. Therefore, instead of r_j we can generate another residual polynomial which will help obtain the p_i's: [illegible] i.e., the same r vectors as those of Algorithm 2.1. It is interesting to note that with this sequence of residual vectors, which will be denoted by $\tilde{r}_j$, it is easy to generate the directions p_i, [illegible] for both algorithms. So the idea is straightforward: obtain the auxiliary residual polynomials $\tilde{\rho}_j$ that are those associated with the [illegible] CR algorithm and exploit them to obtain the π_i's in the same way as in the CR algorithm. The polynomials $\lambda\pi_j$ are orthogonal, and therefore the expression of the desired approximation is the same. The algorithm is described next where now $\tilde{\rho}_j$ is the polynomial associated with the auxiliary sequence $\tilde{r}_j$.

ALGORITHM 2.3. [illegible]

0. [illegible] $\tilde{r}_0 := b - Ax_0$ $p_0 := \tilde{r}_0$ $\pi_0 = \tilde{\rho}_0 = 1$
1. [illegible] $\lambda\pi_0$
2. [illegible] $j = 0, 1, \ldots,$ [illegible]
3. $\tilde{\alpha}_j := \langle \tilde{\rho}_j, \lambda\tilde{\rho}_j\rangle_w / \langle \lambda\pi_j, \lambda\pi_j\rangle_w$
4. $\alpha_j := \langle \psi, \lambda\pi_j\rangle_w / \langle \lambda\pi_j, \lambda\pi_j\rangle_w$
5. $x_{j+1} := x_j + \alpha_j p_j$
6. $\tilde{r}_{j+1} := \tilde{r}_j - \tilde{\alpha}_j A p_j$ $\tilde{\rho}_{j+1} = \tilde{\rho}_j - \tilde{\alpha}_j \lambda\pi_j$
7. $\beta_j := \langle \tilde{\rho}_{j+1}, \lambda\tilde{\rho}_{j+1}\rangle_w / \langle \tilde{\rho}_j, \lambda\tilde{\rho}_j\rangle_w$
8. $p_{j+1} := \tilde{r}_{j+1} + \beta_j p_j$ $\pi_{j+1} := \tilde{\rho}_{j+1} + \beta_j \pi_j$
9. [illegible] $\lambda\pi_{j+1}$
10. [illegible]

It is remarkable that the only difference between this and generic CR-type algorithm (see, e.g., Algorithm 2.1) is that the updates to x_{j+1} use a coefficient α_j different from that of the update to the vectors $\tilde{r}_{j+1}$. Observe that the residual vectors $\tilde{r}_j$ obtained by the algorithm are just auxiliary vectors that do not correspond to the original residuals $r_j = b - Ax_j$. Needless to say, these residuals, the r_j's, can also be generated after line 5 (or 6) from $r_{j+1} = r_j - \alpha_j A p_j$. Depending on the application, it may be necessary to include these computations.

PROPOSITION 2.2. [illegible] x_{j+1} [illegible] j [illegible] 2.3 [illegible] $x_{j+1} = x_0 + s_j(A)r_0$ [illegible] s_j [illegible] j [illegible]

$$s_j(\lambda) = \alpha_0\pi_0(\lambda) + \cdots + \alpha_j\pi_j(\lambda). \tag{11}$$

[illegible] π_j [illegible] $\tilde{\rho}_j(\lambda)$ [illegible]

(12) $$\langle \lambda\pi_j(\lambda), \lambda\pi_i(\lambda)\rangle_w = \langle \lambda\tilde{\rho}_j(\lambda), \tilde{\rho}_i(\lambda)\rangle_w = 0 \quad , \quad i \neq j.$$

$\psi - \lambda s_j(\lambda)$, $\|\psi - \lambda s(\lambda)\|_w$, s , $\leq j-1$

The first observation is that the polynomials $\tilde{\rho}_j$ and π_j are identical with the polynomials ρ_j and π_j of Algorithm 2.1, so the orthogonality property (12) is trivially satisfied. The relation (4) uses scalars α_j that are different from those denoted by $\tilde{\alpha}_j$ of the sequence $\tilde{\rho}_j$. From this relation, we have that $\psi - \lambda s_j(\lambda) = \psi - \sum_{i=0}^{j} \alpha_i \lambda\pi_i(\lambda)$. By the optimality condition, the best polynomial is obtained when the scalars α_i satisfy the relation $\langle \psi - \lambda s_j(\lambda), \lambda\pi_i(\lambda)\rangle_w = 0$, for $i = 1, \ldots, j$. Exploiting (11) and the orthogonality of the system $\{\lambda\pi_i\}_{i=0,\ldots,j}$, this yields

$$\alpha_j = \langle \psi, \lambda\pi_j\rangle_w \;/\; \langle \lambda\pi_j, \lambda\pi_j\rangle_w. \qquad \square$$

It is worth exploring the formula (9), which defines the scalars α_j, a little further. In the standard CR algorithm, the expression (8) is modified by exploiting orthogonality relations to lead to the standard expression of line 3 of Algorithm 2.1. However, this is no longer possible here, essentially because the polynomial s_{j-1} in (9) uses the scalar α_i's (formula (11)), and there are no orthogonality relations satisfied with the corresponding residual polynomials ρ_j. It is, however, possible to express the scalar α_j as a modification to the scalar $\tilde{\alpha}_j$. Indeed, define $\tilde{s}_j \equiv \sum_{i=0}^{j} \tilde{\alpha}_i \pi_i$, which is the solution polynomial of Algorithm 2.1, and observe that $\langle \lambda\tilde{s}_{j-1}, \lambda\pi_j\rangle_w = 0$, because $\lambda\pi_j$ is orthogonal to all polynomials λq_i for polynomials q_i of degree $i \leq j-1$. Then, we can rewrite the numerator of (9) as

$$\langle \psi, \lambda\pi_j\rangle_w = \langle \psi - \lambda\tilde{s}_{j-1}, \lambda\pi_j\rangle_w = \langle (\psi - 1) + 1 - \lambda\tilde{s}_{j-1}, \lambda\pi_j\rangle_w = \langle \tilde{\rho}_j, \lambda\pi_j\rangle_w - \langle 1 - \psi, \lambda\pi_j\rangle_w .$$

Since $\tilde{\rho}_j$ and π_j have the same leading coefficient, by exploiting orthogonality we readily obtain the relation $\langle \tilde{\rho}_j, \lambda\pi_j\rangle_w = \langle \tilde{\rho}_j, \lambda\tilde{\rho}_j\rangle_w$, which yields the following alternative formula for α_j:

(13) $$\alpha_j = \tilde{\alpha}_j - \frac{\langle 1 - \psi, \lambda\pi_j\rangle_w}{\langle \lambda\pi_j, \lambda\pi_j\rangle_w} .$$

The only merit of this expression, as a substitute for (9), is that it clearly establishes Algorithm 2.3 as a "corrected version" of the standard Algorithm 2.1. In the special situation when $\psi \equiv 1$, $\alpha_i = \tilde{\alpha}_i$, and the two algorithms coincide as expected.

2.5. The base filter function. The solutions computed by the algorithms just seen are based on generating polynomial approximations to a certain base filter function ϕ. In the following, we will consider a low-pass filter ϕ. As was already mentioned, it is generally not a good idea to use as ϕ the step function

$$\phi(t) = \begin{cases} 1, & t < \tau_0, \\ 0, & t \geq \tau_0. \end{cases}$$

This is because this function is discontinuous, and approximations to it by high-degree polynomials will exhibit very wide oscillations, known as Gibbs oscillations. It is preferable to take as a "base" filter, i.e., the filter which is ultimately approximated by polynomials, a smooth function such as the one on the left side of Figure 1.

The base filter function can be a piecewise polynomial consisting of two parts: a function which decreases smoothly from 1 to 0 when λ increases from 0 to τ_0, and the constant function zero in the interval $[\tau_0, \beta]$. Alternatively, the function can consist of three parts, one on each of the intervals $[0,\ \tau_0]$, $[\tau_0,\ \tau_1]$, and $[\tau_1,\ \beta]$, with $0 < \tau_0 < \tau_1 < \beta$. It will begin with the constant value 1 in the interval $[0,\ \tau_0]$, then decrease smoothly from 1 to 0 in the second interval $[\tau_0\ \tau_1]$, and finally take the constant value 0 in $[\tau_1,\ \beta]$. The second part of the function (the first part for the first scenario) bridges the values 0 and 1 by a smooth function and was termed a "bridge function" in [11]. In what follows we focus on obtaining bridge functions for the generic case, i.e., for an interval $[\tau_0,\ \tau_1]$.

A systematic way of generating base filter functions is to use bridge functions obtained from Hermite interpolation. The bridge function is an interpolating polynomial (in the Hermite sense) depending on two integer parameters m_0, m_1, and denoted by $\Theta_{[m_0,m_1]}$, which satisfies the following conditions:

$$\begin{aligned} &\Theta_{[m_0,m_1]}(\tau_0) = 1, \qquad \Theta'_{[m_0,m_1]}(\tau_0) = \cdots = \quad \Theta^{(m_0)}_{[m_0,m_1]}(\tau_0) = 0, \\ &\Theta_{[m_0,m_1]}(\tau_1) = 0, \qquad \Theta'_{[m_0,m_1]}(\tau_1) = \cdots = \quad \Theta^{(m_1)}_{[m_0,m_1]}(\tau_1) = 0. \end{aligned} \tag{14}$$

Thus, $\Theta_{[m_0,m_1]}$ has degree $m_0 + m_1 + 1$, m_0, and m_1 define the degree of smoothness at the points τ_0 and τ_1, respectively.

Such polynomials can be easily determined by the usual finite difference tables in the Hermite sense. To find a closed form for the polynomials $\Theta_{[m_0,m_1]}$ it is useful to change variables in order to exploit symmetry. We map the variable onto the interval $[-1, 1]$ and shift the function down by $1/2$. If the corresponding function is denoted by η, then the above conditions become

$$\begin{array}{rclcrcll} \eta(-1) & = & 1/2, & & \eta(+1) & = & -1/2, & \\ \eta^{(i)}(-1) & = & 0 & \text{for } i = 1, \ldots, m_0, & \eta^{(i)}(+1) & = & 0 & \text{for } i = 1, \ldots, m_1\ . \end{array}$$

The derivative function η' can be expressed as $\eta'(t) = c\ (1-t)^{m_1}(1+t)^{m_0}$, and as a result we have a closed form expression of $\eta(t)$:

$$\eta(t) = \frac{1}{2}\ -\ \frac{\int_{-1}^{t}\ (1-s)^{m_1}(1+s)^{m_0}\ ds}{\int_{-1}^{1}\ (1-s)^{m_1}(1+s)^{m_0}\ ds}. \tag{15}$$

The first and second derivatives of η are

$$\eta'(t) = -\frac{(1-t)^{m_1}(1+t)^{m_0}}{\int_{-1}^{1}\ (1-s)^{m_1}(1+s)^{m_0}\ ds}\ , \qquad \eta''(t) = \left[\frac{m_1}{1-t} - \frac{m_0}{1+t}\right]\eta'(t). \tag{16}$$

Thus there is an inflexion point at

$$t = \frac{m_0 - m_1}{m_0 + m_1}.$$

Since the maximum absolute value of the derivative is required for the convergence analysis, it will be useful to determine it. The derivative is negative and decreases from its value at the point -1 to a certain minimum, reached at the inflexion point, and then it increases from there to its final value at the point 1. The peak value and an approximation to it are given by the following lemma.

LEMMA 2.3. [illegible] η [illegible] $[-1,1]$ [illegible]

$$\eta'_{max} = \frac{m_0+m_1+1}{2} \; \frac{m_1^{m_1}\; m_0^{m_0}}{(m_0+m_1)^{m_0+m_1} \times \binom{m_0}{m_0+m_1}} \; . \tag{17}$$

[illegible] m_0 [illegible] m_1 [illegible]

$$\eta'_{max} \approx \frac{m_0+m_1}{2\sqrt{2\pi}} \sqrt{\frac{1}{m_0}+\frac{1}{m_1}} \; . \tag{18}$$

[illegible]. The integral in the denominator of η in (16) can be computed by successive integration by parts to be

$$\int_{-1}^{1} (1-s)^{m_1}(1+s)^{m_0} \; ds = \frac{m_1! m_0!}{(m_0+m_1)!} \times \frac{2^{m_0+m_1+1}}{m_0+m_1+1} \; .$$

Evaluating the negative derivative $-\eta'$ at the inflexion point yields

$$\eta'_{max} = \frac{\frac{(2m_1)^{m_1}\;(2m_0)^{m_0}}{(m_0+m_1)^{m_0+m_1}}}{\frac{m_1!m_0!}{(m_0+m_1)!} \times \frac{2^{m_0+m_1+1}}{m_0+m_1+1}} = \frac{m_0+m_1+1}{2} \; \frac{m_1^{m_1}\; m_0^{m_0}}{(m_0+m_1)^{m_0+m_1} \times \binom{m_0}{m_0+m_1}},$$

which is the first result. This can be rewritten as

$$\eta'_{max} \;=\; \frac{m_0+m_1+1}{2} \; \frac{\frac{m_0^{m_0}}{m_0!} \times \frac{m_1^{m_1}}{m_1!}}{\frac{(m_0+m_1)^{m_0+m_1}}{(m_0+m_1)!}} \; .$$

Using Sterling's formula $m! \approx \sqrt{2\pi m}\; (m/e)^m$ yields (18), after simplifications. ∎

This result must now be translated into the original interval $[\tau_0,\ \tau_1]$. The function Θ (indices m_0, m_1 are omitted) and its derivative in terms of η and η' are

$$\Theta(\lambda) = \frac{1}{2} + \eta\left(2\frac{\lambda-\tau_0}{\tau_1-\tau_0}-1\right), \qquad \Theta'(\lambda) = \frac{2}{\tau_1-\tau_0}\eta'\left(2\frac{\lambda-\tau_0}{\tau_1-\tau_0}-1\right),$$

and so

$$\Theta'_{max} = \frac{2}{\tau_1-\tau_0}\eta'_{max} \; .$$

As an example of a bridge function, the case when $m_0 = m_1 = 2$ yields

$$\eta(t) = \frac{-15}{16} \times \left(t - 2\frac{t^3}{3} + \frac{t^5}{5}\right),$$

which, for the interval $[0,\alpha]$, translates into the function

$$\Theta_{[2,2]}(t) = \frac{1}{2} - \frac{15}{16}\;\left(2\frac{t}{\alpha}-1\right) + \frac{5}{8}\;\left(2\frac{t}{\alpha}-1\right)^3 - \frac{3}{16}\left(2\frac{t}{\alpha}-1\right).^5$$

Similarly, for $m_0 = m_1 = 3$ we find

$$\Theta_{[3,3]}(t) = \frac{1}{2} - \frac{35}{32}\left(2\frac{t}{\alpha}-1\right) + \frac{35}{32}\left(2\frac{t}{\alpha}-1\right)^3 - \frac{21}{32}\left(2\frac{t}{\alpha}-1\right)^5 + \frac{5}{32}\left(2\frac{t}{\alpha}-1\right).^7$$

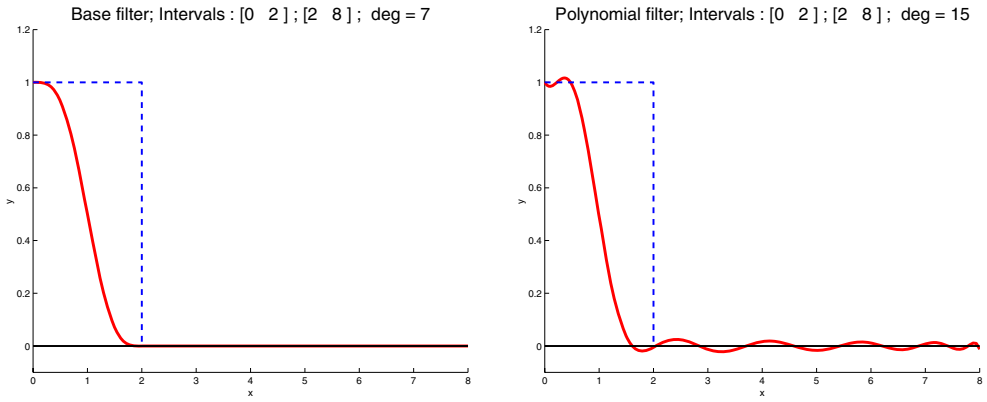


FIG. 2. *Left: Base filter ϕ defined on two intervals: $\Theta_{[4,4]}$ in $[0, 2]$ and zero in $[2, 8]$; Right: its polynomial approximation of degree* 15.

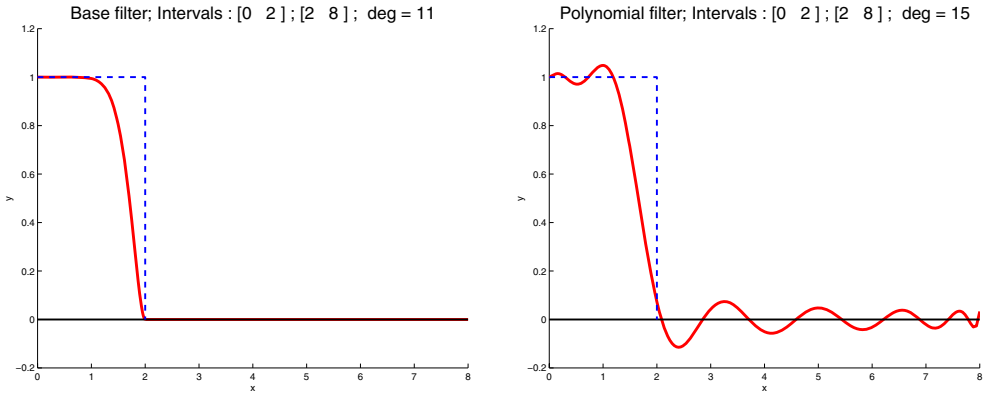


FIG. 3. *Left: Base filter ϕ defined on two intervals: $\Theta_{[10,2]}$ in $[0, 2]$ and zero in $[2, 8]$. Right: Its polynomial approximation of degree* 15.

As was seen, the ratio $\frac{m_1}{m_0}$ determines the localization of the inflexion point. The polynomial can be made to decrease rapidly from one to zero in a small interval by taking high-degree polynomials, but this has the effect of slowing down convergence toward the desired filter, as it tends to cause undesired oscillations.

Two examples of filter functions are shown in Figures 2 and 3. A third example, shown in Figure 4, shows a situation where three intervals are used. In the first interval [0, 1.7] and third interval [2.3, 8], the filter takes the constant values 1 and 0, respectively. In the middle interval [1.7, 2.3], ϕ is defined by the Hermite polynomial $\Theta_{[5,5]}$ in [1.72.3]. This time we plot a higher-degree polynomial approximation to ϕ to show the quality of the resulting polynomial. For higher-degree polynomials (say 80) there is no visible difference between the base filter ϕ and its polynomial approximation. We also computed many other polynomials using Legendre weights in each interval instead of Chebyshev weights and, in all cases, saw no significant difference.

2.6. The weight function w. Denoting the l subintervals of $[0,\ \beta]$ by $[\tau_{i-1}, \tau_i]$, $i = 1, \ldots, l$, we define the inner product on each subinterval (τ_{l-1}, τ_l), using Chebyshev

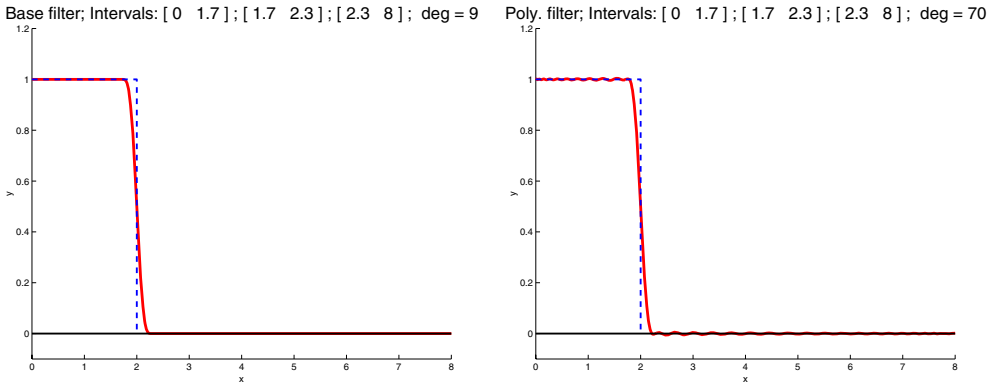


FIG. 4. *Left: Dual base filter ϕ defined on three intervals:* 1 *in* $[0, 1.7]$, $\Theta_{[5,5]}$ *in* $[1.7, 2.3]$, *and* 0 *in* $[2.3, 8]$. *Right: Its polynomial approximation of degree* 70.

weights:

$$\langle \psi_1, \psi_2 \rangle_{\tau_{l-1},\tau_l} = \int_{\tau_{l-1}}^{\tau_l} \frac{\psi_1(t)\psi_2(t)}{\sqrt{(t-\tau_{l-1})(\tau_l - t)}}\, dt.$$

Then the inner product on the interval $[0,\ \beta] \equiv [\tau_0,\ \tau_1] \cup [\tau_1,\ \tau_2] \cdots \cup [\tau_{l-1},\ \tau_l]$ is defined as a weighted sum of the inner products on the smaller intervals,

$$\langle \psi_1, \psi_2 \rangle_w = \sum_{i=1}^{l} \mu_i \int_{\tau_{i-1}}^{\tau_i} \frac{\psi_1(t)\psi_2(t)}{\sqrt{(t-\tau_{i-1})(\tau_i - t)}}\, dt. \tag{19}$$

For example, for two intervals the weight function is defined as

$$\langle \psi_1, \psi_2 \rangle_w = \mu_1 \int_0^{\alpha} \frac{\psi_1(t)\psi_2(t)}{\sqrt{t(\alpha - t)}}\, dt + \mu_2 \int_a^{\beta} \frac{\psi_1(t)\psi_2(t)}{\sqrt{(t-\alpha)(\beta - t)}}\, dt. \tag{20}$$

The μ_i's can be chosen to emphasize or deemphasize specific subintervals. In most of our tests we took the μ_i to be either equal to the constant 1 or to the inverse of the width of each subinterval. Note that we can also use Legendre polynomials, or indeed any other orthogonal polynomials, instead of Chebyshev polynomials. We found very little difference in performance (convergence) between Legendre and Chebyshev polynomials.

The issue of obtaining orthogonal polynomials from sequences of orthogonal polynomials on other intervals was addressed in [12] and [22]. One of the main problems is to avoid numerical integration. In [22] this was achieved by expanding the desired functions in a basis of Chebyshev polynomials on each of the subintervals. Note that the expansions are redundant—but cost is not a major issue. Let $\varsigma^{(l)}$ be the mapping which transforms the interval $[\tau_{l-1}, \tau_l]$ into $[-1, 1]$:

$$\varsigma^{(l)}(\lambda) = \frac{2}{\tau_l - \tau_{l-1}}\lambda - \frac{\tau_l + \tau_{l-1}}{\tau_l - \tau_{l-1}}\ .$$

Denote by C_i the ith degree Chebyshev polynomial of the first kind on $[-1, 1]$, and define

$$C_i^{(l)}(\lambda) = C_i\left(\varsigma^{(l)}(\lambda)\right), \quad i \geq 0.$$

When all polynomials are expanded in the above Chebyshev bases on each interval, then all operations involved in Algorithms 2.1, 2.2, and 2.3 are easily performed with the expansion coefficients. Thus, adding and scaling two expanded polynomials is a trivial operation. Consider now inner products of two polynomials. Recall that on each interval the scaled and shifted Chebyshev polynomials $(C_k^{(l)})_{k\in\mathbb{N}}$ constitute an orthogonal basis since

$$\langle C_i^{(l)}, C_j^{(l)} \rangle_{\tau_{l-1},\tau_l} = \begin{cases} 0 & \text{if } i \neq j, \\ \pi & \text{if } i = j = 0, \\ \frac{\pi}{2} & \text{if } i = j \neq 0. \end{cases}$$

As a result, if two polynomials ψ_1, ψ_2 are expanded in the above Chebyshev bases for each interval, the inner products (19) of these polynomials are trivially obtained from their expansion coefficients in the bases.

The only remaining operation to consider is that of multiplying a polynomial by λ (e.g., line 9 of Algorithm 2.3). A polynomial ψ expanded in the Chebyshev bases can easily be multiplied by the variable λ, by exploiting the following relations:

$$\lambda\, C_i^{(l)}(\lambda) = \frac{\tau_l - \tau_{l-1}}{4} C_{i+1}^{(l)}(\lambda) + \frac{\tau_l + \tau_{l-1}}{2} C_i^{(l)}(\lambda) + \frac{\tau_l - \tau_{l-1}}{4} C_{i-1}^{(l)}(\lambda), \quad i \geq 1,$$
$$\lambda\, C_0^{(l)}(\lambda) = \frac{\tau_l - \tau_{l-1}}{2} C_1^{(l)}(\lambda) + \frac{\tau_l + \tau_{l-1}}{2} C_0^{(l)}(\lambda).$$

These formulations come from the recurrences obeyed by Chebyshev polynomials: $2tC_i(t) = C_{i+1}(t) + C_{i-1}(t)$ for $i > 0$, and $tC_0(t) = C_1(t)$.

2.7. Convergence. It is desirable to know how fast the polynomial ρ_j converges to the low-pass filter function ϕ. Convergence results of this type utilize uniform norm results. We will restrict ourselves to a simple result derived from the Jackson theorems; see [7]. A common notation adopted in the theory of approximation of functions is the following. For a given continuous function f, define the [illegible] of f by

$$E_n(f) = \min_{p \in \mathcal{P}_n} \|f - p\|_\infty,$$

where $\|g\|_\infty$ is the infinity norm of a continuous function g, on the interval $[\alpha, \beta]$,

$$\|g\|_\infty = \max_{t \in [\alpha\ \beta]} |g(t)| \ .$$

The Weierstrass theorem states that any continuous function f can be uniformly approximated by polynomials [7]. In particular this means that $\lim_{n\to\infty} E_n(f) = 0$. In the early 1900s, Jackson proved a number of theorems which give further information on this convergence. The following is the third of the Jackson theorems. Another definition is needed before stating the theorem: The [illegible] of a bounded function f on an interval $[\alpha,\ \beta]$ is defined as

$$\omega_f(\delta) = \sup_{|t_1 - t_2| \leq \delta} |f(t_1) - f(t_2)|. \tag{21}$$

THEOREM 2.4 (Jackson's theorem III). [illegible] $f \in C[0\ 2\pi]$

$$E_n(f) \leq \omega_f\left(\frac{\pi}{n+1}\right) \ . \tag{22}$$

See [7] for proofs and additional details. For an arbitrary interval $[\alpha, \beta]$ the above theorem translates into

$$E_n(f, [\alpha, \beta]) \leq \omega_f \left(\frac{\beta - \alpha}{2(n+1)} \right) . \tag{23}$$

Applying the above result to base filter functions is easy.

LEMMA 2.5.

$$\phi(t) = \begin{cases} 1 & t \in [0, \tau_0), \\ \Theta_{[m_0, m_1]} & t \in [\tau_0, \tau_1), \\ 0 & t \in [\tau_1, \beta]. \end{cases}$$

$$\omega_\phi(\delta) \leq \frac{2\eta'_{max}}{\tau_1 - \tau_0}\delta,$$

η'_{max} (17) (18) m_0, m_1

Substituting this result into Jackson's theorem, we obtain the following bound.

PROPOSITION 2.6. ϕ 2.5

$$E_n(\phi) \leq \frac{\beta\ \eta'_{max}}{(n+1)(\tau_1 - \tau_0)}, \tag{24}$$

η'_{max} (17) (18) m_0, m_1

The above result is about convergence in the ∞-norm. Obtaining a result for the L-2-norm with the weight function w is straightforward and standard because the norms are related to each other in a simple way. Specifically, the following is easily shown:

$$\|g\|_w \leq K\|g\|_\infty \quad \text{with} \quad K = \|1\|_w.$$

For example, if we have l intervals and the μ_i's are equal to 1 in (19), then $K = \sqrt{l\ \pi}$.

3. Applications and extensions. Polynomial filtering has many applications in numerical linear algebra and related areas. In fact, we can argue that the number of these applications is likely to increase because of the growing need to solve problems in reduced dimensions and to apply various forms of PCA. In [19], we have considered the use of polynomial filters in information retrieval. The paper [18] exploits similar ideas for the problem of eigenfaces. Here we examine a few other applications which may also benefit from polynomial filtering. Though we will show a few supporting experiments shortly, the ideas are exposed here only to describe the rationale and the concepts, and some of these ideas will be further explored in forthcoming articles.

3.1. Computing a large invariant subspace. In this section we show how polynomial filtering can be used to compute large invariant subspaces of symmetric real (or Hermitian complex) matrices. Specifically, the following problem is addressed: A

The simplest form of this problem is to compute all eigenvalues of A that are $\leq \tau$. It can be assumed that an upper bound β for the spectrum is available, and, without loss of generality, that all eigenvalues are ≥ 0. Consider this case first. One

solution to the problem is to use the Lanczos algorithm for the matrix $q(A)$, where q is a low-pass filter polynomial such that $q(\lambda) \approx 1$ for $0 \leq \lambda \leq \tau$ and $q(\lambda) \approx 0$ for $\tau < \lambda \leq \beta$. To reduce cost, the polynomial should not be of high degree. What might happen with this approach is that the Lanczos procedure will quickly produce a good invariant subspace associated with the largest eigenvalues of $q(A)$. If enough steps are taken, then clearly this subspace should include the desired subspace, which could be easily extracted by a simple Rayleigh–Ritz projection. The main point is that a shorter basis is required because the Lanczos algorithm will converge faster, and this will lead to a much lower cost due to much less expensive orthogonalization steps. Indeed, it was observed in [4] that, for large invariant subspaces, the high cost of orthogonalization far outweighs the additional cost of the matrix-vector products with $p(A)$. This comes with the added benefit of using less memory.

The procedure described above can be enhanced by filtering the initial vector of the Lanczos procedure. The reason why this could be useful is the observation that if v has a zero component with respect to $\lambda_i > \tau$, then since $q(\lambda_i)$ is close to zero, the components of the Lanczos vectors will also remain close to zero throughout the algorithm. We can use a high-degree polynomial to filter the initial vector and then a low-degree polynomial for the inner loop of the Lanczos procedure. Initial results show that this process works as predicted and may lead to good savings in time when compared with standard approaches.

Next we provide a motivation for this approach based on an application from quantum mechanics (for details, see [4]) and then explore in detail the case of interior eigenvalues.

3.2. Motivation. In electronic structures calculations one is faced with the problem of computing an orthogonal basis of the invariant subspace associated with the k lowest eigenvalues of a Hamiltonian matrix. This particular problem was the original motivation for this work. The Hamiltonian is (real) symmetric. A major difficulty with these calculations is that the dimension k of the subspace can be quite large. A typical example would be that $k = 1,000$ and that n, the dimension of the matrix, is $n \approx 1,000,000$. Methods based on standard restarted Lanczos procedures tend to suffer from the need to save a very large set of basis vectors as well as from the need for a very large number of costly restarts and reorthogonalizations. An alternative considered recently is to forego the restarts and not focus on individual eigenvectors; see, e.g., [4]. This approach is usually faster than the implicit restarted version of Lanczos, but it may require the use of secondary storage as the Lanczos basis can be quite large.

As an illustration consider a hypothetical situation where, for example, $m = 2000$ Lanczos vectors are required by a standard Lanczos procedure to compute a subspace of dimension $k = 100$. The cost of orthogonalization will be $0.5m^2 \times n$, which is $2 \times 10^6 \times n$ operations. In contrast, if polynomial filtering is used in the manner described earlier and if only 200 vectors are needed, the new cost will $10^4 \times n$ plus the additional cost of matrix-vector products. If degree 10 polynomials are used and the matrix has, say, 13 nonzero entries per row, then this additional cost is roughly $200 * 10 * 13n = 26000n$. So the total adds up to $\approx 36,000n$ operations versus $2,000,000n$. Of course this example is hypothetical and somewhat extreme, but it underscores the unacceptable cost of orthogonalization for large bases. One may argue that a much smaller basis might be needed for the restarted Lanczos method. Though the situation is generally difficult to analyze, the point remains that restarting is expensive because eigenvectors are repeatedly (implicitly) computed. It is not the

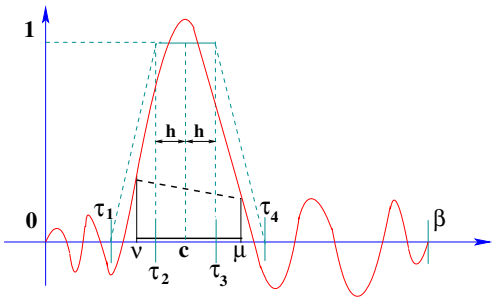


FIG. 5. *Illustration of the procedure to set up the intervals and the base filter function ψ. The polynomial shown is the polynomial which results from approximating ψ with the choice of $\tau_1, \ldots, \tau_4$.*

goal of this paper to compare these approaches. This is done in another article [4], where these comparisons are undertaken for realistic problems arising from electronic structures calculations.

3.2.1. Interior invariant subspaces. A case not considered in [4] is the situation of interior eigenvalues. Though the overall scheme is not too different from that of the computation of the smallest or largest eigenvalues, a few difficulties arise which make the scheme somewhat more complex. One of the main difficulties lies in the selection of the (dual) base filter ψ. Suppose we want all eigenvalues in the subinterval $[\eta,\ \mu] \subset [0\ \beta]$. To construct the base filter ψ we will need to subdivide the interval $[0,\ \beta]$ into five subintervals, $[0,\ \beta] \equiv [\tau_0,\ \tau_1] \cup [\tau_1,\ \tau_2] \cup \cdots \cup [\tau_4,\ \tau_5]$, with $0 = \tau_0 < \tau_1 < \tau_2 < \tau_3 < \tau_4 < \tau_5 = \beta$. In the intervals $[\tau_0\ \tau_1]$ and $[\tau_4\ \tau_5]$, the function ψ takes the value 0. In the central interval, $[\tau_2,\ \tau_3]$, the function ψ has value 1. The other two intervals bridge the values 0 and 1, and so the global function ψ is continuous and sufficiently smooth.

We would like to use the Lanczos algorithm on the matrix $p(A)$, where $p(\lambda) = \lambda s(\lambda)$ is the polynomial approximation to the filter ψ. In order not to miss eigenvalues in the desired interval it is essential that $p(\lambda_i)$ be larger than $p(\lambda_k)$ for each $\lambda_i \in [\eta,\ \mu]$ and $\lambda_k \notin [\eta\ \mu]$. Because of the likely imbalance between the left and right branches of the polynomial, it is not easy to guarantee this without an iterative process for selecting the intervals and ψ. The goal of the iterative process is to guarantee that $p(\mu) = p(\eta)$ and that all eigenvalues inside the interval $[\eta,\ \mu]$ will be mapped to the largest eigenvalues of $p(A)$. To achieve this, a bisection algorithm is applied, whereby τ_2, τ_3 are changed until the relation $p(\mu) = p(\eta)$ is approximately satisfied. A few details on this procedure follow. The discussion is illustrated in Figure 5.

Initially, the values of τ_1 and τ_4 are fixed so that $\tau_1 = \nu - \delta$ and $\tau_4 = \mu + \delta$ for a certain δ (our code uses $\delta = 0.05 * (\mu - \nu)$). Then what is left is to determine τ_2, τ_3. These are set to be of the form $\tau_2 = c - h$ and $\tau_3 = c + h$, where h is a small fraction of the interval width (our code uses $h = (\tau_4 - \tau_1)/10$). This means that the desired "plateau" interval for p is chosen to be of the form $[\tau_2,\ \tau_3] = [c - h,\ c + h]$, where h is fixed and c is to be found. Now the only unknown left is c, which is determined by bisection so that the resulting p satisfies $p(\eta) = p(\mu)$.

The figure reveals another potential problem. Recall that the goal is to use the Lanczos algorithm with the matrix $p(A)$. In order to be able to stop the iteration, it is necessary to know whether the required eigenvalues have converged. Following [4], this is done without computing eigenvectors, but by only considering the tridiagonal

matrix T_m generated from the Lanczos iteration. If the sum of those eigenvalues of T_m which correspond to the wanted eigenvalues of A has converged, then the process is stopped. These eigenvalues are $p(\lambda_i)$ for $\lambda_i \in [\nu, \mu]$. This stopping criterion is modeled after the one in [4], where the restricted trace (sum of desired eigenvalues) has an important physical meaning (total energy) and is therefore the proper quantity to monitor for convergence. In the smallest / largest eigenvalues case, the situation is simple: the eigenvalues $p(\lambda_i)$ of T_m corresponding to the desired λ_i are (in general) the largest eigenvalues of T_m. This facilitates the test. For the case of interior eigenvalues and middle-pass filters, the situation is not as straightforward. Figure 5 illustrates this, since there are points λ at the right of τ_4 whose values $p(\lambda)$ are larger than some values $p(\lambda)$ for λ inside the interval $[\nu,\ \mu]$. We handle this by a heuristic iterative procedure. Once the value of c has been obtained by bisection, it is necessary to check whether the following condition is satisfied:

$$\sup_{\lambda \in [0,\ \nu) \cup (\mu,\ \beta]} |p(\lambda)| \ < \ \min_{\lambda \in [\nu,\ \mu]} |p(\lambda)| \ .$$

If this is not satisfied, then the interval $[\tau_1, \tau_4]$ is expanded by doubling the value of δ. At the same time h is halved, leading to a shrinking of the plateau interval $[\tau_2, \tau_3]$. The process is then repeated until a satisfactory interval is found. In addition to this, the function which determines the interval also returns the maximum value, say γ, of $p(\lambda)$ outside the interval $[\nu,\ \mu]$ so as to recognize which eigenvalues λ of A do belong to the desired interval $[\nu,\ \mu]$: If the eigenvalue $p(\lambda)$ is $> \gamma$, then λ must belong to $[\nu,\ \mu]$. Note that λ is not available; only $p(\lambda)$ is.

The main point of the above discussion is that the polynomials are easy to use, and it is inexpensive to work with them, so some careful preprocessing can be done to ensure that the correct eigenspace is computed.

In the following algorithm, p_0 denotes the "prefilter" polynomial, while p is the filter polynomial used in the iteration. Typically, the prefilter polynomial is of high degree (e.g., 200), while the internal polynomial is of low degree (e.g., 20).

ALGORITHM 3.1.

$A \in \mathbb{R}^{n \times n}$ $q_1,\ \|q_1\|_2 = 1$ m

$[\nu, \mu]$

A $[\nu, \mu]$

0. getIntv
1. $\beta_1 = 0 \quad q_0 = 0$
2. > 0 $q_1 := p_0(A)q_1 \quad q_1 = q_1/\|q_1\|_2$
3. **for** $i = 1, \ldots, m$
4. $\quad w_i = p(A)q_i - \beta_i q_{i-1}$
5. $\quad \alpha_i =< w_i, q_i >$
6. $\quad w_i = w_i - \alpha_i q_i$
7. $\quad \beta_{i+1} = \|w_i\|_2$
8. $\quad$ **if** $(\beta_{i+1} == 0)$ **then** stop
9. $\quad q_{i+1} = w_i/\beta_{i+1}$
10. $\quad T_i = \text{tridiag}(\beta_i, \alpha_i, \beta_{i+1})$
11. $\quad$ T_i
12. $\quad n_{ev}$ $\lambda_j^\star$ T_i $\lambda_j^\star > \gamma$
13. $\quad s_i = \sum_{\lambda_j^\star > \gamma} \lambda_j^\star$
14. $\quad$ **if** $(|s_i - s_{i-1}| < |s_{i-1}| * \text{tol})$ **then** break

15. **end**
16. $j = 1 : n_{ev+2}$
17. z_j T_i
18. $\tilde{\lambda}_j = (Az_j, z_j) \notin [\nu,\ \mu]$ $\tilde{\lambda}_j, z_j$
19. **end**

A few comments are in order. The function `getIntv` referenced in line 0 is the heuristic procedure discussed above, which carefully determines the five subintervals of $[0, \beta]$ and the initial filter ψ. It returns in particular the scalar γ used in line 12, which is such that if $p(\lambda) > \gamma$, then $\lambda \in [\nu,\ \mu]$. In line 16, we compute n_{ev+2} Ritz pairs instead of n_{ev} as a safeguard only. The test in the next lines will keep only the required eigenvalues. The eigenvalues $\lambda_j^\star$ are eigenvalues of T_i, and they approximate eigenvalues of $p(A)$. The convergence test in lines 13–14 need not be executed at each step (this is an $O(i^2)$ process); we can instead perform it at regular intervals. Finally, it is clear that it is essential to include some form or reorthogonalization. In [4] we used partial reorthogonalization.

3.3. Computing $f(A)v$. The procedures described earlier compute approximations to $\phi(A)v$, where ϕ is a specific spline function on up to five intervals. There is, of course, no reason why one should be limited to spline functions which approximate filters. The approach can be extended to other situations where a vector of the form $f(A)v$ is to be computed. The problem of approximating $f(A)v$ has been extensively studied (see, e.g., [25, 23, 5, 16, 15]), though the attention was primarily focussed on the case when f is analytic (e.g., $f(t) = \exp(t)$). Problems which involve noncontinuous functions, such as the step function or the sign function, can also be important. The approach described in this paper can be trivially extended to the case where ϕ is a general spline function. One can certainly imagine situations where a certain vector $f(A)v$ is to be evaluated, where f is some complex function known through an accurate piecewise polynomial approximation. The framework developed in this paper is ideally suited for handling this situation. The only extensions required are to increase the number of intervals (which is now ≤ 5) and to define ψ in each of these intervals by the polynomials of the spline function.

Another interesting application is when approximating $\psi(A)$, where ψ is the sign function. Computing the sign function of a matrix has important applications in QMC (quantum chromo dynamics); see, e.g., [13]. In this case we need to use three intervals, for example, $[a_-\ d_-], [d_-\ d_+], [d_+\ a_+]$, where d_-, a_- are negative and d_+, a_+ are positive. The difficulty here is to compute estimates for the interior values d_- and d_+.

3.4. Estimating the number of eigenvalues in an interval. The most common way to compute the number of eigenvalues inside an interval is to exploit the Sylvester inertia theorem and the LDL^T factorization [14]. However, for large matrices this is not always practically feasible, or it may be too expensive.

It is sometimes useful to obtain a rough idea of the number of eigenvalues of a Hermitian matrix that are located inside a given interval. This information can be used, for example, for the case when the smallest k eigenvalues of A must be computed by using a form of polynomial filtering. In this situation an interval $[0,\ \tau]$ must be found which contains these k eigenvalues. A guess for τ can be given and then refined by answering the question: How many eigenvalues are located on the left of τ?

One possible solution to this can be provided by the Lanczos procedure. One can simply run the Lanczos algorithm without reorthogonalization (the Cullum–Willoughby algorithm; see [8]) or with partial reorthogonalization and record the

number n_τ of all eigenvalues below τ of the tridiagonal matrix T_m obtained from the Lanczos algorithm. When this number stabilizes (i.e., all eigenvalues below τ converge), then n_τ will represent the desired number. The problem with this approach is that it may be very expensive when the number n_τ is large.

A rough approximation of n_τ can be easily obtained from statistical arguments, using polynomial filtering. This technique is an adaptation of methods described elsewhere for estimating the trace of certain operators; see, for example, [17, 21, 3]. Consider a low-pass polynomial filter such as the one shown in Figure 4, and an arbitrary vector v of 2-norm unity. Expand the vector v in the eigenbasis as

$$v = \sum_{i=1}^{n} \xi_i u_i,$$

and consider the inner product of v with $p(A)v$:

$$(v, p(A)v) = \sum_{i=1}^{n_\tau} \xi_i^2 p(\lambda_i) + \sum_{i=n_v+1}^{n} \xi_i^2 p(\lambda_i).$$

If the polynomial p is selected so that it is close to 1 on $[0,\ \tau]$ and to 0 in $(\tau,\ \beta]$, then clearly the second sum in the above expression should be close to zero, and the first close to the sum $\sum_{i=1}^{n_\tau} \xi_i^2$. If the vector v is a random vector, then the ξ_i's are unbiased, and therefore the ratio $\sum_{i=1}^{n_\tau} \xi_i^2 / \sum_{i=1}^{n} \xi_i^2$ should be close to n_τ/n. In the end we can estimate n_τ by

$$n_\tau \approx n \times (v, p(A)v). \tag{25}$$

Of course, a unique sample may not be good enough, and several trials should be taken and the results averaged. The numerical experiments sections explore this approach a little further. It should be emphasized that, as is always the case, it is expensive to obtain an accurate answer by statistical methods in general. Accordingly, this approach may be useful only when a rough estimate of n_τ is wanted and other methods cannot be considered. Two appealing features of the method are its exclusive reliance on matrix-vector products and its highly parallel nature.

4. Numerical tests. Applications of filtered polynomial iterations to information retrieval and face recognition have been reported elsewhere [18, 19]. In addition, the use of these ideas for computing large eigenspaces has recently been successfully exploited; see [4]. Section 4.2 explores this further.

The goals of the tests discussed in this section are (a) to examine the convergence of the process, (b) to show and compare a few of the techniques discussed earlier for computing invariant subspaces, and (c) to demonstrate the use of polynomial filtering for approximating inertia of shifted matrices (see section 3.4).

All tests were performed with Matlab on a Linux workstation (running Debian) and equipped with two 1.7 GHz Xeon processors (with 256kB cache) and 1 GB of main memory.

4.1. Convergence. In this test we generate a matrix obtained from the discretization of a Laplacian using centered differences on a 25×15 mesh. We then compute the vector v, which has all components equal to 1 in the eigenbasis; i.e., v is the sum of all the (normalized) eigenvectors. This vector is then filtered with a chosen

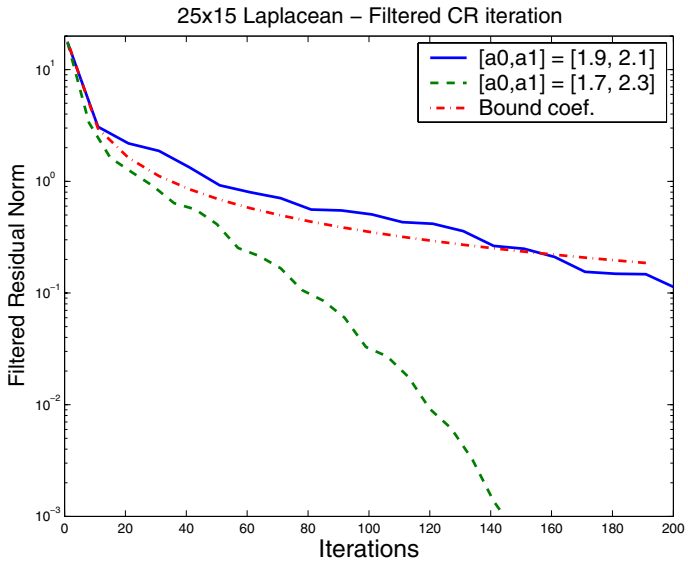


FIG. 6. *Convergence of filtered polynomial CR algorithm for two different cases, and comparison with the coefficient of the bound* (24).

low-pass base filter ϕ, and we plot $\|\phi(A)v - (I - As_{k-1}(A))v\|_2$ for $k = 1, \ldots, 200$. This is referred to as the "filtered residual." The low-pass filter is selected as follows:

$$\phi(t) = \begin{cases} 1 & \text{for} \quad t \in [0, \tau_0), \\ \Theta_{[m_0,m_1]} & \text{for} \quad t \in [\tau_0, \tau_1), \\ 0 & \text{for} \quad t \in [\tau_1, \beta]. \end{cases} \tag{26}$$

A first run used the values $m_0 = m_1 = 10$, $\beta = 8$, $\tau_0 = 1.9, \tau_1 = 2.1$, and the second used the same values for m_0, m_1, and β, and changed τ_0, τ_1 to $\tau_0 = 1.8, \tau_1 = 2.2$. The plot in Figure 6 shows three curves. The first two show the progress of the filtered residual norm for the two runs (solid line and dashed line, respectively). The third one (dash-dot) shows the coefficient in the right-hand side of (24) corresponding to the first test case ($m_0 = m_1 = 10$, $\beta = 8$, $\tau_0 = 1.9$). Here, η'_{max} is estimated by (18), where for $m_0 = m_1 = 10$ we find that $\eta'_{max} \approx \sqrt{m_0/\pi}$. So the third curve shows exactly the sequence

$$\frac{8\sqrt{10/\pi}}{0.4 * (i+1)}, \quad i = 1, \ldots, 200.$$

Two observations can be made. The first is that for the second run, the behavior is not at all predicted by the bounds. It has an exponential character not seen in the bounds obtained in section 2.7. The second observation is the big difference in convergence between two seemingly close situations. If the middle interval increases in width, we can get very fast convergence. However, note that taking a wide middle interval may yield a function that is not desirable from other viewpoints; i.e., there may be situations when this interval must be taken to be small. In information retrieval this is not critical [19]. When computing invariant subspaces, on the other hand, it is undesirable to have a wide gap since it will include eigenvalues that need to be eliminated by some other means.

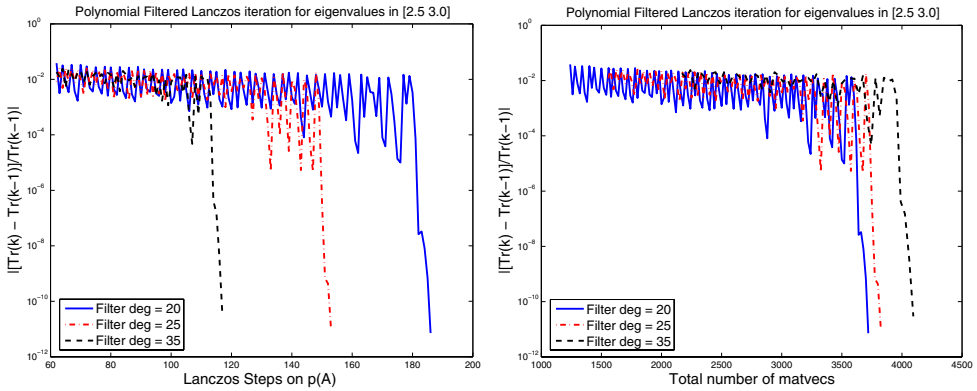


FIG. 7. *Convergence of a filtered Lanczos iteration for computing all eigenvalues of a Laplacian of dimension* 891*, located in the interval* [2.5, 3].

4.2. Computing invariant subspaces. Polynomial filtering can be helpful in the situation when a large invariant subspace associated with all eigenvalues in a given interval is to be computed. In [4] we have shown how the method can be applied to realistic problems arising from electronic structures. The problem there is that of computing an invariant subspace associated with the smallest eigenvalues of a large symmetric real matrix. The ideas used in [4] follow closely those sketched in section 3.1. For matrices such as those that arise in electronic structures, the method works well because the invariant subspace is quite large, and this causes methods which rely too much on orthogonalization to become excessively expensive. This little fact, which has not been adequately addressed by researchers in numerical methods, is well-known to researchers in the physics community. By reducing the frequency as well as the cost of orthogonalization, one can reduce the overall cost dramatically. Thus, a strategy based on polynomial filtering combined with the inexpensive "partial reorthogonalization" resulted in gains close to a factor of 12 in some cases (see [4]) relative to standard existing codes such as ARPACK [20].

Since the case corresponding to the smallest (or largest) eigenvalues has already been covered in detail in [4], we will illustrate next the case of interior eigenvalues. This case is just as important, because there is currently a lack of good algorithms for dealing with it.

In the experiments which follow, we consider a model problem arising from a Laplacian matrix. The matrix corresponds to the discretization of the Laplacian on an $(nx+2) \times (ny+2)$ grid including boundary points. After applying zero Dirichlet boundary conditions, we obtain a matrix of dimension $n = n_x n_y$.

In the first test, we take $n_x = 27, n_y = 33$. This results in a matrix of dimension $n = 891$. We would like to compute all eigenvalues of A in the interval $[\nu,\ \mu] = [2.5,\ 3]$. As it turns out, there are 60 eigenvalues in this interval. We ran three tests with a different degree of the filter polynomial: degree 20, 25, and 35. We did not apply prefiltering. The base filter function uses bridge functions of the form $\Theta_{[10,10]}$. The boundaries for the various intervals defining the base filter function were set up as described in section 3.1. The same initial vector for the Lanczos iteration was used for all three runs, and it was generated randomly. Algorithm 3.1 was run with tol$= 1.e-10$. The plots in Figure 7 show the error measure $|s_i - s_{i-1}|/|s_{i-1}|$ used in lines 13–14 of Algorithm 3.1 to test convergence. These error rates are plotted against the

TABLE 1

Number of Lanczos steps and sum of final eigenvalue errors as expressed by (27) *for the filtered Lanczos procedure using three different filtering polynomials..*

Degrees	20	25	35
Lancz. steps	190	157	120
Error-sums	6.77e-12	4.631e-12	5.570e-11

TABLE 2

Number of Lanczos steps and sum of final eigenvalue errors as expressed by (27) *for the filtered Lanczos procedure using two different filtering polynomials. The matrix is a* 3*-D Laplacian of dimension* $n = 10,051$.

Degrees	75	80
Lancz. steps	364	270
Error-sums	5.684e-14	1.430e-13

number of Lanczos steps (left figure) and against the total number of matrix-vector products (matvecs; right side).

As expected, the number of Lanczos steps decreases as the degree of the polynomial increases. In the case of degree 35, the procedure requires 120 Lanczos steps to compute all 60 eigenvalues. This good performance comes at the cost of 35 matvecs with A per Lanczos step. This amounts to $175n$ operations per Lanczos steps for the matvec, and the total number of matvecs with A is 4200. For the lower degree of 20, we now need 190 steps, so we need a total of 3800 matvecs. Though this is lower than with the degree 35, the comparison favors the higher degree if the cost of orthogonalization is taken into account (a full reorthogonalization is performed). The total number of matvecs may appear to be quite high. However, the alternative of running the Lanczos algorithm to compute eigenvalues from the first to the last one in the interval may be much more costly for realistic cases because of the cost of orthogonalization. This was demonstrated for the computation of smallest eigenvalues in a realistic computation in [4].

In order to verify that the code run does not miss eigenvalues we printed the errors

$$\sum_{\lambda_i \in [\nu,\ \mu]} \min_j |\tilde{\lambda}_j - \lambda_i|, \tag{27}$$

where the $\tilde{\lambda}_j$ are the approximate Ritz values computed in lines 16–18 of Algorithm 3.1. These are printed in Table 1 along with the number of Lanczos steps required for convergence.

The next test proceeds along the same lines but considers a more difficult problem. We take a three-dimensional (3-D) Laplacian with $n_x = 23$, $n_y = 23$, and $n_z = 19$, leading to a problem of size $n = 10,051$. The eigenvalues of A are located in the interval $[0,\ 12]$, so to make the problem more challenging we try to compute eigenvalues around the middle of the interval. Specifically, we seek to compute all eigenvalues in the interval $[\nu,\ \mu] = [6.25,\ 6.30]$. There are 53 eigenvalues of A in this interval. This particular example will require higher-degree polynomials than the smaller example seen above to reach convergence in a small number of Lanczos steps. We take polynomials of degrees 75 and 100. For the filter function the bridge functions are of the form $\Theta_{[25,15]}$. Results similar to the ones seen above are shown in Figure 8 and Table 2.

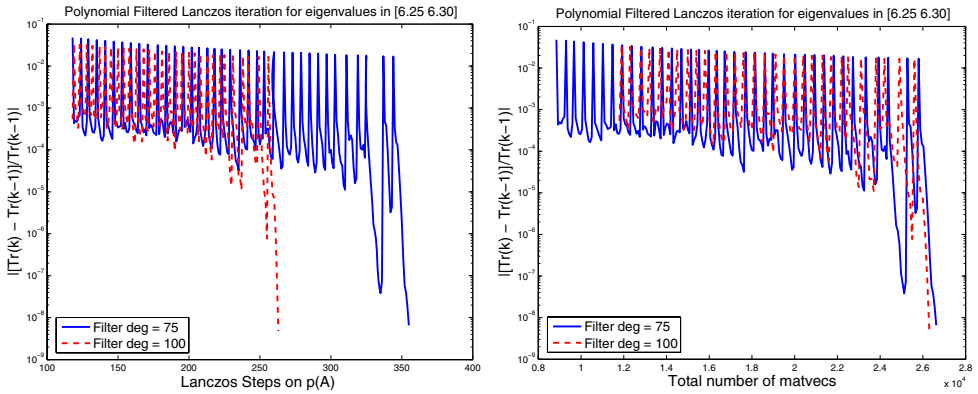


FIG. 8. *Convergence of a filtered Lanczos iteration for computing all eigenvalues of a Laplacian of dimension* $10,051$ *located in the interval* $[6.25,\ 6.30]$.

It is interesting to ask the question: What are the alternatives to this approach for solving this problem? In this case, the best known method is to use a shift-and-invert technique, whereby the Lanczos procedure is applied to $(A-\sigma I)^{-1}$. However, factoring the matrix $A-\sigma I$ can be quite expensive, especially for 3-D problems of this type and when considering the fact that the matrix is highly indefinite. For very large 3-D problems factorization may not even be a feasible option. The other alternative is to employ a Lanczos-type procedure to compute a large number of eigenvalues until those of interest are reached. Orthogonalization and the need to keep a large basis will be two serious problems for large matrices. Polynomial filtering can be attempted in such cases. An approach of this type was suggested in [2, 1] in an algorithm which exploits implicit restarts. The IRBL code (Matlab) presented in [2] uses Leja points for the purpose of acceleration, instead of the least-squares polynomials used in this paper. The other major difference is that IRBL is a block algorithm.

Another idea that is similar in spirit to the one described in this paper is presented in [10]. There, a polynomial is constructed by compounding a quadratic polynomial with a higher degree Chebyshev polynomial in order to obtain a desired filter. In fact, the paper [10] explores several other methods for computing interior eigenvalues. For their problem, called the Anderson model of localization, the authors found that the best approach is the Cullum–Willoughby [8] technique based on the Lanczos algorithm without reorthogonalization. The problem in [10] is somewhat different from the one addressed here in that the number of eigenvalues/ eigenvectors computed is relatively small (all tests were with five eigenpairs).

As pointed out in [10] and elsewhere (see, e.g., [26]), the potential difficulty with any polynomial filtered approach is the high cost of the procedure if large-degree polynomials are required. Though we do not offer comparisons with competing methods, we can say that polynomial filtered Lanczos procedures are likely to be superior to competing techniques in some situations. Specifically, they may offer the best alternative in situations when (a) a large number of eigenvalues and eigenvectors must be computed, (b) matrix-vector products are not expensive, and (c) there are not too many eigenvalues around the interval boundaries ν, μ. Condition (a) is based on the observation that when the subspace is large the cost of the eigenvalue calculation is dominated by orthogonalization. The result is that a big part of this cost can be traded off with filtering, which leads to fewer steps in the Lanczos algorithm. Condi-

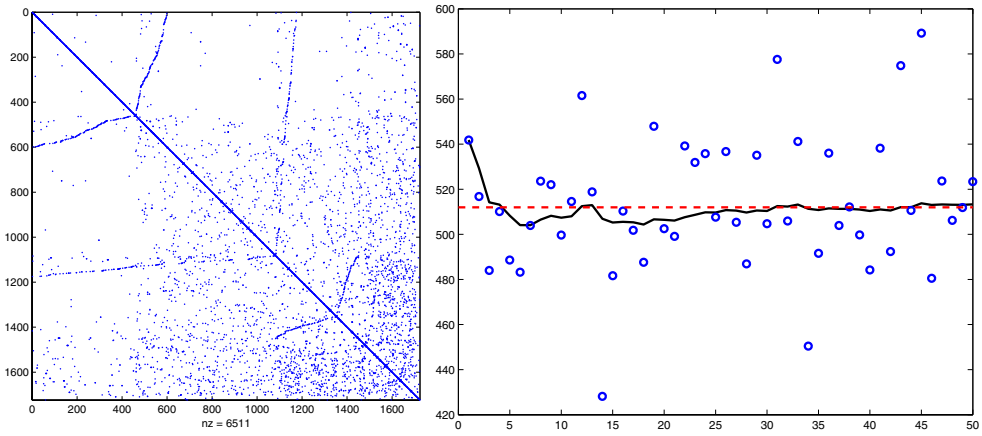


FIG. 9. *Pattern of matrix* `bcspwr09` *(Left) and stochastic estimate of its number of negative eigenvalues (right).*

tion (b) will ensure that convergence will be reached without resorting to a polynomial of too high a degree.

4.3. Estimating the number of eigenvalues in an interval. This section reports on a test with the stochastic estimator of the inertia of a shifted matrix, i.e., the number of eigenvalues of a matrix that are below a certain number α. Section 3.4 suggested a simple algorithm for this calculation for the case when a rough estimate of this number is wanted.

For this test we took a matrix from the Harwell–Boeing collection [9], namely the matrix `bcspwr09`. This matrix is of size $n = 1,723$ and has $nnz = 6,511$ nonzero entries. The sparsity pattern of the matrix is shown on the left side of Figure 9. This matrix has all its eigenvalues in the interval $[-3.117\ldots, 5.971\ldots]$. The question one may ask is: How many eigenvalues are negative? The correct answer is 512. We shifted everything by 3.2 (so A becomes $A + 3.2I$) and we sought the number of eigenvalues of the shifted matrix that are below $\alpha = 3.2$. A dual filter ψ using three intervals, defined as in (26), was used with the parameters: $m_0 = m_1 = 10$. The interval bounds given were $0, \tau_0 = 3.15, \tau_1 = 3.25, \beta = 6$. The degree of the polynomial used was $m = 20$.

The right side of Figure 9 shows a test with 50 runs (each using a polynomial of degree 20 and a different unit random vector v). The number n_α reported for given k in the x-axis is simply the average of the numbers given by formula (25) over all previous k tests:

$$n_\alpha(k) = \frac{n}{k} \times \sum_{i=1}^{k} (v_i, p(A)v_i).$$

The small circles in the figure are the values of $n \times (v_i, p(A)v_i)$ obtained from each (single) sample. The dashed horizontal line represents the correct answer, which is 512. Notice that there are a few outliers, e.g., the smallest single estimate obtained was close to 428 and the largest close to 590, but the average over several runs quickly converges to a reasonable estimate. So after 30 runs (a total of 600 matrix-vector products), a fairly good estimate is reached.

5. Conclusion. Polynomial filtering is a useful and versatile tool in computational linear algebra. It is most appealing in situations where rough solutions to various matrix problems are sought. We have shown a few such applications, and hinted at others, where approximations to the matrix problem are sought which are restricted to be in a small space.

Apart from the methods related to low-rank approximations mentioned above, polynomial filtering has also been tried in the past with limited success in the more traditional areas of matrix computations, for example for the problem of preconditioning. Polynomial filtering is not a panacea, but it can play a significant role in specific cases. Perhaps the most important of these is the computation of large invariant subspaces. A successful use of polynomial filters in a realistic application has already been reported elsewhere [4].

There are many other potential uses of polynomial filtering in numerical linear algebra which remain to be explored. Many computations require the solution of least-squares systems with regularization. We also hinted at the problem of computing $f(A)b$ when f is a spline function, which can itself be an approximation to an arbitrary function.

REFERENCES

[1] J. Baglama, D. Calvetti, and L. Reichel, *IRBL: An implicitly restarted block-Lanczos method for large-scale Hermitian eigenproblems*, SIAM J. Sci. Comput., 24 (2003), pp. 1650–1677.

[2] J. Baglama, D. Calvetti, L. Reichel, and A. Ruttan, *Computation of a few close eigenvalues of a large matrix with application to liquid crystal modeling*, J. Comput. Phys., 146 (1998), pp. 203–226.

[3] C. Bekas, E. Kokiopoulou, and Y. Saad, *An estimator for the diagonal of a matrix*, Appl. Numer. Math., (2007), to appear.

[4] C. Bekas, E. Kokiopoulou, and Y. Saad, *Polynomial Filtered Lanczos Iterations with Applications in Density Functional Theory*, Technical Report umsi-2005-117, Minnesota Supercomputer Institute, University of Minnesota, Minneapolis, MN, 2005; SIAM J. Sci. Comput., submitted.

[5] L. Bergamaschi, M. Caliari, and M. Vianello, *Efficient computation of the exponential operator for discrete 2d advection-diffusion equations*, Numer. Linear Algebra Appl., 10 (2003), pp. 271–289.

[6] M. W. Berry and M. Browne, *Understanding Search Engines: Mathematical Modeling and Text Retrieval*, 2nd ed., Software Environ. Tools 17, SIAM, Philadelphia, 2005.

[7] C. C. Cheney, *Introduction to Approximation Theory*, McGraw-Hill, New York, 1966.

[8] J. K. Cullum and R. A. Willoughby, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations: Vol.* I: *Theory*, Classics in Appl. Math. 41, SIAM, Philadelphia, 2002.

[9] I. S. Duff, R. G. Grimes, and J. G. Lewis, *Sparse matrix test problems*, ACM Trans. Math. Software, 15 (1989), pp. 1–14.

[10] U. Elsner, V. Mehrmann, F. Milde, R. A. Römer, and M. Schreiber, *The Anderson model of localization: A challenge for modern eigenvalue methods*, SIAM J. Sci. Comput., 20 (1999), pp. 2089–2102.

[11] J. Erhel, F. Guyomarc, and Y. Saad, *Least-Squares Polynomial Filters for Ill-Conditioned Linear Systems*, Technical Report umsi-2001-32, Minnesota Supercomputer Institute, University of Minnesota, Minneapolis, MN, 2001.

[12] B. Fischer and G. H. Golub, *How to generate unknown orthogonal polynomials out of known orthogonal polynomials*, J. Comput. Appl. Math., 43 (1992), pp. 99–115.

[13] A. Frommer, T. Lippert, B. Medeke, and K. Shilings, *Numerical Challenges in Lattice Quantum Chromodynamics*, Lectures Notes in Comput. Sci. 15, Springer-Verlag, Berlin, 1999.

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

[15] M. Hochbruck and C. Lubich, *On Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 34 (1997), pp. 1911–1925.

[16] M. Hochbruck, C. Lubich, and H. Selhofer, *Exponential integrators for large systems of differential equations*, SIAM J. Sci. Comput., 19 (1998), pp. 1552–1574.
[17] M. F. Hutchinson, *A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines*, Commun. Statist. Simula., 19 (1990), pp. 433–450.
[18] E. Kokiopoulou and Y. Saad, *PCA without eigenvalue calculations: A case study on face recognition*, in Proceedings of the Fifth SIAM International Conference on Data Mining, Newport, CA, 2005, SIAM, Philadelphia, 2005.
[19] E. Kokiopoulou and Y. Saad, *Polynomial filtering in latent semantic indexing for information retrieval*, in Proceedings of the ACM-SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, 2004, ACM, New York, 2004.
[20] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK User's Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, Software Environ. Tools 6, SIAM, Philadelphia, 1998; see also the software at http://www.caam.rice.edu/software/ARPACK.
[21] G. A. Parket, W. Zhu, Y. Huang, D. K. Hoffman, and D. J. Kouri, *Matrix pseudospectroscopy: Iterative calculation of matrix eigenvalues and eigenvectors of large matrices using a polynomial expansion of the dirac delta function*, Comput. Phys. Comm., 96 (1996), pp. 27–35.
[22] Y. Saad, *Iterative solution of indefinite symmetric linear systems by methods using orthogonal polynomials over two disjoint intervals*, SIAM J. Numer. Anal., 20 (1983), pp. 784–811.
[23] Y. Saad, *Analysis of some Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 29 (1992), pp. 209–228.
[24] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelphia, 2003.
[25] H. A. van der Vorst, *An iterative solution method for solving $f(A)x = b$, using Krylov subspace information obtained for the symmetric positive definite matrix A*, J. Comput. Appl. Math., 18 (1987), pp. 249–263.
[26] K. Wu, *Preconditioned Techniques for Large Eigenvalue Problems*, Ph.D. thesis, Department of Computer Science, University of Minnesota, Twin Cities, Minneapolis, MN, 1997.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 3, pp. 871–889

MULTILEVEL HIERARCHICAL MATRICES*

SABINE LE BORNE†

Abstract. This paper deals with a multilevel construction of hierarchical matrix approximations to the inverses of finite element stiffness matrices. Given a sequence of discretizations $A_\ell x_\ell = f_\ell$, $\ell = 0, \ldots, n$, where $A_0 x_0 = f_0$ denotes the coarse grid problem, we will compute A_0^{-1} exactly and then use interpolation to obtain an $\mathcal{H}$-matrix approximation $A_{\ell+1}^{-\mathcal{H}}$ from the approximate $\mathcal{H}$-matrix inverse $A_\ell^{-\mathcal{H}}$ on the next coarser grid. We develop an exact interpolation scheme for the inverse of tridiagonal matrices as they appear in the finite element discretization of one-dimensional differential equations. We then generalize this approach to two spatial dimensions where these efficiently computed approximations to the inverse may serve as preconditioners in iterative solution methods. We illustrate this approach with some numerical tests for convection-dominated convection-diffusion problems.

Key words. hierarchical matrices, data-sparse approximation, multilevel, inverses of tridiagonal matrices

AMS subject classifications. 65F05, 65F30, 65F50

DOI. 10.1137/040607964

1. Introduction.

In a series of papers, the technique of hierarchical matrices—$\mathcal{H}$-matrices for short—has been introduced [4, 8, 10, 11]. $\mathcal{H}$-matrices provide an inexpensive but sufficiently accurate approximation to fully populated matrices as they appear in boundary element methods. In finite element methods, it is the inverse of the stiffness matrix which is fully populated. An $\mathcal{H}$-matrix approximation can be computed and stored in almost linear complexity, i.e., $\mathcal{O}(n \log_s^\alpha n)$ with moderate parameter α [3]; the constants, however, in these complexity estimates are rather large (they easily lie in the hundreds!). They depend on the structure of the involved $\mathcal{H}$-matrices and have been computed exactly in [8]. In this paper, we will introduce an efficient interpolation-based approach to compute these approximate inverses: Given a sequence of discretizations $A_\ell x_\ell = f_\ell$, $\ell = 0, \ldots, n$, where $A_0 x_0 = f_0$ denotes the coarse grid problem, we will compute A_0^{-1} exactly and then use interpolation to obtain an $\mathcal{H}$-matrix approximation $A_{\ell+1}^{-\mathcal{H}}$ from the approximate $\mathcal{H}$-matrix inverse $A_\ell^{-\mathcal{H}}$ on the next coarser grid. We develop an exact interpolation scheme for the inverses of tridiagonal matrices as they appear in the finite element discretization of one-dimensional differential equations. The inverse of a tridiagonal matrix has been studied extensively in several papers in the past. A review of this topic is given in [15] for symmetric matrices, and some results for the inverses of nonsymmetric tridiagonal matrices can be found in [16] and the references therein. These results motivate data-sparse approximations to the exact inverse, which can be used as preconditioners in iterative methods [6, 20]. We then generalize this approach to two spatial dimensions and use these efficiently computed approximations to the inverse as preconditioners in iterative solution methods (e.g., BiCGstab).

*Received by the editors May 10, 2004; accepted for publication (in revised form) by R. Nabben May 4, 2006; published electronically October 30, 2006. This work was supported in part by the U.S. Department of Energy under grant DE-FG02-04ER25649 and in part by the National Science Foundation under grant DMS-0408950.

http://www.siam.org/journals/simax/28-3/60796.html

†Department of Mathematics, Tennessee Technological University, Box 5054, Cookeville, TN 38505 (sleborne@tntech.edu).

The remainder of this paper is structured as follows: In section 2, we provide a brief introduction to the construction and arithmetic of $\mathcal{H}$-matrices. Section 3 deals with the multilevel construction of $\mathcal{H}$-matrices for one-dimensional problems. In section 4, the generalization to two-dimensional problems is developed, and section 5 concludes this paper with some numerical results.

2. A brief introduction to $\mathcal{H}$-matrices. In this section, we introduce the main concepts of $\mathcal{H}$-matrices to the extent of which they are required for the remainder of this paper. For more detailed introductions, we refer the reader to [4, 8, 10, 11] and the references therein.

An $\mathcal{H}$-matrix approximation to a given (dense) matrix is obtained by replacing certain blocks of the matrix by matrices of a low rank k, stored in so-called Rk-format, as will be further explained below. Given such an $\mathcal{H}$-matrix, the standard matrix operations such as matrix-vector multiplication, matrix-matrix addition and multiplication, as well as (approximate) matrix inversion can be defined for this $\mathcal{H}$-matrix format. Whereas these ($\mathcal{H}$-)matrix operations yield only approximations, they can be performed in almost optimal complexity, i.e., $\mathcal{O}(n \log^{\alpha} n)$ with moderate parameter α. The construction of $\mathcal{H}$-matrices is reviewed in subsection 2.1, and their arithmetic is reviewed in subsection 2.2.

2.1. Construction of $\mathcal{H}$-matrices. The formal definition of an $\mathcal{H}$-matrix depends on appropriate hierarchical partitionings of the index set and also of the product index set, which are organized in (block) cluster trees, as defined next. Instead of fixed partitionings, these trees will provide hierarchies of partitionings, which gives a hierarchical matrix its name.

DEFINITION 2.1 (cluster tree). *Let I be a finite index set and let $T_I = (V, E)$ be a tree with vertex set V and edge set E. For a vertex $v \in V$ we define the set of successors of v as $S(v) := \{w \in V \mid (v, w) \in E\}$. The tree T_I is called a cluster tree of I if its vertices consist of subsets of I and satisfy the following conditions:*

1. *$I \in V$ is the root of T_I and $v \subset I$, $v \neq \emptyset$ for all $v \in V$.*
2. *For all $v \in V$ there holds either $S(v) = \emptyset$ or $v = \dot{\bigcup}_{w \in S(v)} w$.*

In the following, we identify V and T_I, i.e., we write $v \in T_I$ instead of $v \in V$. The nodes $v \in V$ are called clusters.

For regular grids one can construct the cluster tree T_I in a cardinality balanced way, i.e., an index cluster is divided into a certain number of successors of approximately the same size with respect to the number of indices [10, 12]. In this paper we restrict our attention to the case of two (or no) sons per cluster, which is the easiest with respect to the analysis and implementation. For locally refined grids, the results from [8] indicate that the cardinality balanced clustering is *not* optimal with respect to the complexity of adaptive $\mathcal{H}$-matrix updates. Instead, one can use a geometrically balanced approach as described in [9], which is used in the remainder of this paper. In the case of regular grids, however, both approaches yield very similar if not identical results.

A simple example for the construction of a cluster tree is given in Figure 2.1. In this example, the geometrically and cardinality balanced approaches result in the same cluster tree. Here, the regular grid Ω_h contains 64 vertices, i.e., $I = \{1, \ldots, 64\}$ (assuming continuous, piecewise linear elements). The index set is subdivided into two subsets of size 32 each along the middle vertical line. The resulting subsets are both subdivided horizontally, resulting in a total of four subsets of size 16. This process is continued until 64 subsets (or clusters) of size 1 are obtained.

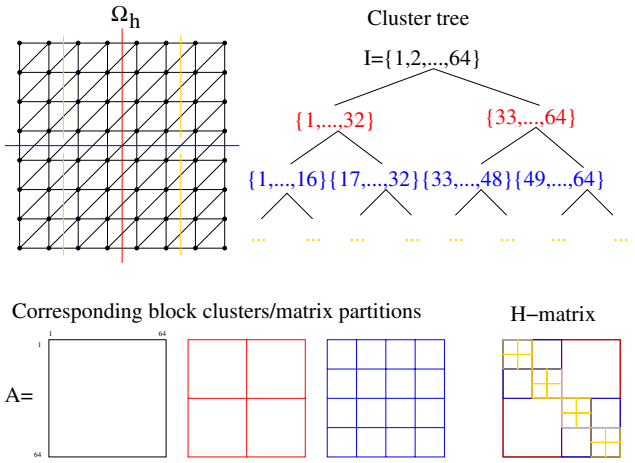


FIG. 2.1. *A mesh Ω_h (top left), a corresponding cluster tree (top right), the resulting hierarchy of matrix partitions (bottom left), and an example of an $\mathcal{H}$-matrix (bottom right).*

DEFINITION 2.2 (leaf, predecessor, level, depth). *Let T_I be a cluster tree. The* set of leaves *of the tree T_I is $\mathcal{L}(T_I) = \{v \in T_I \mid S(v) = \emptyset\}$. The uniquely determined predecessor (father) of a nonroot vertex $v \in T_I$ is denoted by $\mathcal{F}(v)$. The* levels *of the tree T_I are defined by*

$$T_I^{(0)} := \{I\}, \quad T_I^{(\ell)} := \{v \in T_I \mid \mathcal{F}(v) \in T_I^{(\ell-1)}\} \qquad \text{for } \ell \in \mathbb{N},$$

and we write $\mathrm{level}(v) = \ell$ *if $v \in T_I^{(\ell)}$. The depth of T is defined as* $\mathrm{d}(T) := \max\{\ell \in \mathbb{N} \cup \{0\} \mid T_I^{(\ell)} \neq \emptyset\}$.

For any cluster tree T_I there holds $I = \dot{\bigcup}\{v \mid v \in \mathcal{L}(T_I)\}$, i.e., the set of leaves yields a partition of the index set I. A hierarchy of block partitionings of the product index set $I \times I$ is based upon a cluster tree T_I and is organized in a block cluster tree.

DEFINITION 2.3 (block cluster tree). *Let T_I be a cluster tree of the index set I. A cluster tree $T_{I\times I}$ is called a block cluster tree (based upon T_I) if for all $v \in T_{I\times I}^{(l)}$ there exist $s, t \in T_I^{(l)}$ such that $v = s \times t$. The nodes $v \in T_{I\times I}$ are called block clusters.*

Analogously to the cluster tree, for any block cluster tree $T_{I\times I}$ there holds $I \times I = \dot{\bigcup}\{v \mid v \in \mathcal{L}(T_{I\times I})\}$, i.e., the leaves of the block cluster tree provide a block partition of the product index set $I \times I$.

The objective is to construct a block cluster tree from a given cluster tree such that the leaves (of the block cluster tree) correspond to (preferably large) matrix blocks with "smooth" data that can be approximated by low rank matrices in the following Rk-matrix representation.

DEFINITION 2.4 (Rk-matrix representation). *Let $k, n, m \in \mathbb{N} \cup \{0\}$. Let $M \in \mathbb{R}^{n\times m}$ be a matrix of at most rank k. A representation of M in factorized form*

$$M = AB^T, \quad A \in \mathbb{R}^{n\times k}, \quad B \in \mathbb{R}^{m\times k}, \tag{2.1}$$

with A and B stored in full matrix representation, is called an Rk-matrix representation of M, or, in short, we call M an Rk-matrix.

If the rank k is small compared to the matrix size given by n and m, we obtain considerable savings in the storage and work complexities of an Rk-matrix compared

to a full matrix, i.e., $(m+n)k$ versus mn memory cells (or flops). Such a representation has also been used, e.g., in [19], where it is referred to as "skeleton."

In the following construction, we build a block cluster tree iteratively by starting from $I \times I$ and refining the block clusters if they do not satisfy a certain admissibility condition. The choice of the admissibility condition depends on the underlying continuous problem (i.e., the elliptic partial differential equation, in particular its associated Green's function) and shall ensure that all admissible blocks allow a sufficiently accurate Rk-approximation. A typical admissibility condition for uniformly elliptic problems, which we will refer to as the *standard* or *strong* admissibility condition, is as follows:

$$\mathrm{Adm}_s(s \times t) = \mathrm{TRUE} \iff \min(\mathrm{diam}(s), \mathrm{diam}(t)) \leq \eta \ \mathrm{dist}(s,t). \tag{2.2}$$

Here, "diam" and "dist" denote the Euclidean diameter/distance of the (union of the) supports of the basis functions with indices in s, t, respectively. In some cases (e.g., one-dimensional problems), the weaker admissibility condition

$$\mathrm{Adm}_w(s \times t) = \mathrm{TRUE} \iff s \neq t \tag{2.3}$$

turns out to be sufficient [13]. A given cluster tree together with an admissibility condition allows the following canonical construction of a block cluster tree:

Let the cluster tree T_I be given. We define the block cluster tree $T_{I\times I}$ by $\mathrm{root}(T) := I \times I$, and each vertex $s \times t \in T$ has the set of successors

$$S(s \times t) := \begin{cases} \emptyset & \text{if } s \times t \text{ admissible,} \\ \emptyset & \text{if } \min\{\#s, \#t\} \leq n_{\min}, \\ \{s' \times t' \mid s' \in S(s), t' \in S(t);\} & \text{otherwise.} \end{cases} \tag{2.4}$$

The parameter n_{min} ensures that blocks do not become too small where the matrix arithmetic of a full matrix is more efficient than any further subdivision. It is typically set such that $10 \leq n_{min} \leq 100$. The leaves of a block cluster tree obtained through this construction will be used in the definition of an $\mathcal{H}$-matrix.

DEFINITION 2.5 ($\mathcal{H}$-matrix). *Let $k, n_{\min} \in \mathbb{N} \cup \{0\}$, and let $n := \#I$ be the number of indices in an index set I. The set of $\mathcal{H}$-matrices induced by a block cluster tree $T := T_{I\times I}$ with blockwise rank k and minimum block size $n_{\min}$ is defined by*

$$\mathcal{H}(T,k) := \{M \in \mathbb{R}^{n,n} \mid \forall s \times t \in \mathcal{L}(T) : \mathrm{rank}(M|_{s\times t}) \leq k \text{ or } \min\{\#s, \#t\} \leq n_{\min}\}.$$

A matrix $M \in \mathcal{H}(T,k)$ is said to be given in $\mathcal{H}$-matrix representation if the blocks $M|_{s\times t}$ with $\mathrm{rank}(M|_{s\times t}) \leq k$ are in Rk-matrix representation and the remaining blocks with $\min\{\#s, \#t\} \leq n_{\min}$ are stored as full matrices.

An example for the block structure of an $\mathcal{H}$-matrix constructed with the weak admissibility condition (2.3) is shown in Figure 2.1 (bottom right). Both the accuracy and (storage) complexity of an $\mathcal{H}$-matrix approximation to a given matrix depend on the construction of an appropriate cluster tree, i.e., a hierarchy of index set partitionings. Details regarding approximation errors for blocks that satisfy the admissibility condition, i.e., for blocks that have a (relatively) large distance compared to their diameters, as well as storage requirements for full, Rk-, and $\mathcal{H}$-matrices, are given in [8]. The intuitive objective in the construction of a cluster tree, given the standard admissibility condition (2.2), is to partition the index set into clusters of vertices that are geometrically far from each other. As a result, relatively large blocks become

admissible and we obtain an accurate $\mathcal{H}$-matrix approximation that is inexpensive to store (i.e., almost linear complexity).

Whereas the classical $\mathcal{H}$-matrix uses a fixed rank for the Rk-blocks, it is possible to replace it by *variable* (or *adaptive*) *ranks* in order to enforce a desired accuracy within the individual blocks. In particular, for a given admissible block $s \times t$, we set the rank k of the corresponding matrix block $M|_{s\times t}$ as follows:

$$k(M|_{s\times t}) := \min\{k' \mid \sigma_{k'} \leq \delta\sigma_1\}, \tag{2.5}$$

where $\sigma_0 \geq \sigma_1 \geq \cdots$ denote the singular values of $M|_{s\times t}$, and $0 < \delta < 1$ denotes the desired relative accuracy within each block. Numerical tests have shown that adaptive ranks are typically superior to fixed ranks, especially when applied to singularly perturbed problems [14]. A related idea where variable ranks have been assigned depending on the hierarchy level (see Definition 2.2) was also pursued in [19].

2.2. Arithmetic of $\mathcal{H}$-matrices. Given two $\mathcal{H}$-matrices $A, B \in \mathcal{H}(T, k)$ based on the same block cluster tree T, i.e., with the same block structure, the exact sum or product of these two matrices will typically not belong to $\mathcal{H}(T, k)$. In the case of matrix addition, we have $A + B \in \mathcal{H}(T, 2k)$; the rank of an exact matrix product is less obvious. We will use a truncation operator $\mathcal{T}^{\mathcal{H}}_{k\leftarrow k'}$ to define the $\mathcal{H}$-matrix addition $C := A \oplus_{\mathcal{H}} B$ and $\mathcal{H}$-matrix multiplication $C := A \otimes_{\mathcal{H}} B$, where again $C \in \mathcal{H}(T, k)$.

A truncation of a rank k' matrix R to rank $k < k'$ is defined as the best approximation with respect to the Frobenius (or spectral) norm in the set of rank k matrices. In the context of $\mathcal{H}$-matrices, we use such truncations for all admissible (rank k) blocks. Using truncated versions of the QR-decomposition and singular value decomposition, the truncation of a rank k' matrix $R \in \mathbb{R}^{n,m}$ (given in the form $R = AB^T$, where $A \in \mathbb{R}^{n,k'}$ and $B \in \mathbb{R}^{m,k'}$) to a lower rank can be computed with complexity $\mathcal{O}(5(k')^2(n + m) + 23(k')^3)$; further details are provided in [8].

We then define the $\mathcal{H}$-matrix addition and multiplication as follows:

$$\begin{aligned} A \oplus_{\mathcal{H}} B &= \mathcal{T}^{\mathcal{H}}_{k\leftarrow 2k}(A + B);\\ A \otimes_{\mathcal{H}} B &= \mathcal{T}^{\mathcal{H}}_{k\leftarrow k'}(AB), \end{aligned}$$

where $k' \leq c(p+1)k$ is the rank of the exact matrix product, c denotes some constant (which depends on the block cluster tree T), and p denotes the depth of the tree (see Definition 2.2). Estimates show that the $\mathcal{H}$-matrix addition and multiplication have almost optimal complexity and are provided in [8] along with efficient implementations of these operations.

The approximate $\mathcal{H}$-matrix addition and multiplication permit the explicit computation of an approximate matrix inverse in $\mathcal{H}$-matrix format. One possible approach to construct an $\mathcal{H}$-inverse is defined recursively in the block structure which results from the block cluster tree (see Figure 2.1): An approximation to the exact inverse

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}^{-1} = \begin{pmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}S^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}S^{-1} \\ -S^{-1}A_{21}A_{11}^{-1} & S^{-1} \end{pmatrix}, \tag{2.6}$$

with Schur complement $S = A_{22} - A_{21}A_{11}^{-1}A_{12}$, is computed by replacing the exact matrix multiplication, addition, and inversion on the coarser levels by $\mathcal{H}$-arithmetic [10]. Other approaches to compute an approximate inverse may be based on Newton-like iterations.

3. Multilevel $\mathcal{H}$-matrices for one-dimensional problems. In this section we deal with the $\mathcal{H}$-matrix representation of the matrix inverse of a tridiagonal stiffness matrix representing a discretized one-dimensional (convection-diffusion) differential equation. In particular, we will prove that the exact $\mathcal{H}$-matrix representation of a fine grid inverse A_h^{-1} may be obtained by interpolation from the coarse grid inverse A_H^{-1}.

3.1. General setting. In a one-dimensional setting, an (upwind) discretization of the convection-diffusion equation leads to a tridiagonal matrix. Its exact inverse can be represented in $\mathcal{H}$-matrix format as shown in the following.

THEOREM 3.1. *Let A be an irreducible and nonsingular tridiagonal $n \times n$ matrix. Its exact inverse A^{-1} can be represented as an $\mathcal{H}$-matrix using the weak admissibility condition* (2.3) *and rank* 1 *representations in the admissible blocks. The exact storage of the inverse matrix amounts to $(1 + 2\log_2 n)n$ in the case $n = 2^p$.*

Proof. The weak admissibility condition produces an $\mathcal{H}$-matrix format for A^{-1} where the diagonal blocks are successively subdivided until 1×1 blocks are obtained on the diagonal, as illustrated in Figure 3.1.

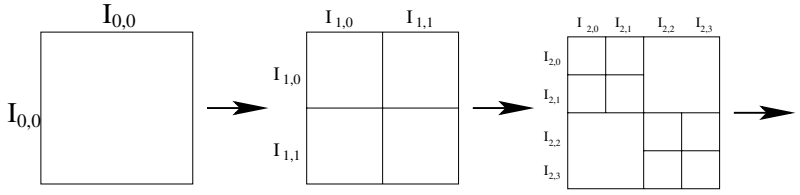


FIG. 3.1. *$\mathcal{H}$-matrix structure for weak admissibility.*

The theorem follows directly from a result on the inverses of tridiagonal matrices [7, 5] which states that A is tridiagonal if and only if there exist four sequences u_i, v_i, x_i, and y_i where $u_i v_i = x_i y_i$ for all i such that $A^{-1} =: C = (c_{ij})$ is given by

$$c_{ij} = \begin{cases} u_i v_j & : \quad i \le j, \\ x_i y_j & : \quad i > j. \end{cases} \tag{3.1}$$

For the storage requirements of an $\mathcal{H}$-matrix with a format as illustrated in Figure 3.1, see Lemma 3.1 in [10]. ∎

Theorem 3.1 has already been stated in the first paper that appeared on $\mathcal{H}$-matrices [10], however, with a different proof. We will use the representation (3.1) together with a matrix-coefficient-based interpolation scheme that has previously been used in the construction of generalized hierarchical bases [2] and wavelets.

We will analyze the case of a one-dimensional constant coefficient boundary value problem, giving rise to a constant coefficient tridiagonal stiffness matrix A_H when discretized using some finite element or finite difference approximation on a uniform mesh of meshwidth H. The terms of the sequences u_i, v_i, x_i, y_i can be associated with the geometric vertex locations, as illustrated in Figure 3.2 (top).

Let $A_H = \text{tridiag}\,[a_H, c_H, b_H]$ be the tridiagonal matrix that arises from a discretization on the coarse mesh with meshwidth H, and let u_i^c, v_i^c, x_i^c, y_i^c be the four sequences for the representation of A_H^{-1} as in (3.1). The superscript "c" indicates that these are the sequences representing the *coarse* grid inverse matrix.

Let $A_h = \text{tridiag}\,[a_h, c_h, b_h]$ be the tridiagonal matrix that arises from a discretization on the regularly refined mesh with meshwidth $h = H/2$. In the following, we will show that the four sequences u_i^f, v_i^f, x_i^f, y_i^f for the representation of the *fine* grid

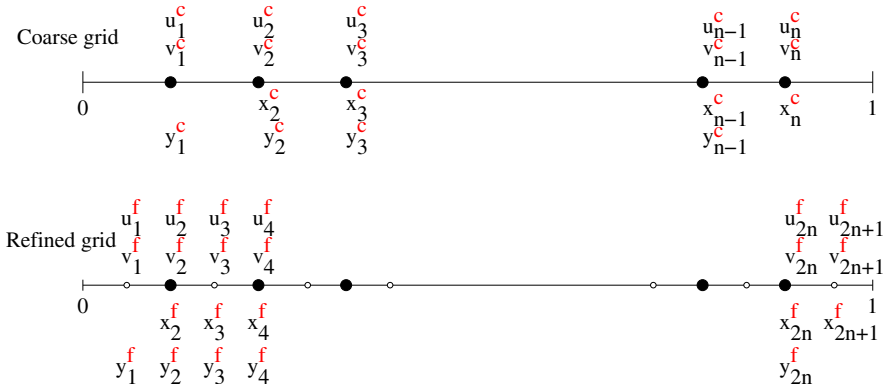


FIG. 3.2. *One-dimensional coarse and refined grids and associated sequences.*

inverse A_h^{-1} can be obtained through interpolation of the terms of the corresponding coarse sequences.

Let $n = 1/H - 1$ denote the number of coarse grid vertices. We compute the fine grid sequences u_i^f and v_i^f by interpolation for the interior fine grid points,

$$u_{2i}^f = u_i^c, \quad u_{2i+1}^f = \theta u_i^c + \vartheta u_{i+1}^c, \qquad i = 1, \ldots, n-1, \tag{3.2}$$

$$v_{2i}^f = v_i^c, \quad v_{2i+1}^f = \vartheta v_i^c + \theta v_{i+1}^c, \qquad i = 1, \ldots, n-1, \tag{3.3}$$

$$u_{2n}^f = u_n^c, \quad v_{2n}^f = v_n^c, \tag{3.4}$$

and extrapolation for the first and last fine grid points,

$$u_1^f = \theta_{ext} u_1^c + \vartheta_{ext} u_2^c, \quad u_{2n+1}^f = \tilde{\theta}_{ext} u_{2n-2}^c + \tilde{\vartheta}_{ext} u_{2n}^c, \tag{3.5}$$

$$v_1^f = \tilde{\vartheta}_{ext} v_1^c + \tilde{\theta}_{ext} v_2^c, \quad v_{2n+1}^f = \vartheta_{ext} v_{2n-2}^c + \theta_{ext} v_{2n}^c. \tag{3.6}$$

In an analogous fashion, we interpolate (and extrapolate) the terms of the sequences x_i^c, y_i^c to obtain x_i^f, y_i^f.

The interpolation coefficients θ, ϑ and the extrapolation coefficients θ_{ext}, ϑ_{ext}, $\tilde{\theta}_{ext}$, $\tilde{\vartheta}_{ext}$ are problem-dependent and will be computed from the corresponding (fine grid) matrix entries in A_h; details will be given in the subsequent application to the one-dimensional convection-diffusion equation.

3.2. Application to the one-dimensional convection-diffusion equation. As an example, we consider the one-dimensional convection-diffusion equation $-(u' + \beta u)' = f$ on an interval I, with Dirichlet boundary conditions and constant β. In the case of a Scharfetter–Gummel discretization [1] on a uniform mesh, the resulting tridiagonal matrix has entries

$$a_h = -\frac{\mathcal{B}(-\beta h)}{h}, \tag{3.7}$$

$$b_h = -\frac{\mathcal{B}(\beta h)}{h}, \tag{3.8}$$

$$c_h = \frac{\mathcal{B}(\beta h) + \mathcal{B}(-\beta h)}{h} \quad (= -a_h - b_h), \tag{3.9}$$

where $\mathcal{B}(\cdot)$ denotes the Bernoulli function $\mathcal{B}(x) = \frac{x}{e^x-1}$. For this case, we choose interpolation factors

$$\theta = \tilde{\vartheta} = -\frac{b_h}{c_h} = \frac{\mathcal{B}(\beta h)}{\mathcal{B}(-\beta h) + \mathcal{B}(\beta h)} = \frac{1-e^{\beta h}}{1-e^{2\beta h}}, \tag{3.10}$$

$$\tilde{\theta} = \vartheta = -\frac{a_h}{c_h} = 1-\theta. \tag{3.11}$$

These are the same coefficients we get by calculating θ from $u(x+h) = \theta u(x) + (1-\theta)u(x+2h)$ and making the interpolation exact for functions of the form $u(x) = \alpha e^{-\beta x} + \gamma$, i.e., for the fundamental solution of the homogeneous equation. For the extrapolation, we set the coefficients to

$$\theta_{ext} = \frac{e^{3\beta h}-1}{e^{2\beta h}-1}, \quad \vartheta_{ext} = 1-\theta_{ext}, \tag{3.12}$$

$$\tilde{\theta}_{ext} = \frac{e^{-\beta h}-1}{e^{2\beta h}-1}, \quad \tilde{\vartheta}_{ext} = 1-\tilde{\theta}_{ext}. \tag{3.13}$$

Again, these are the same coefficients we get by calculating θ_{ext} and $\tilde{\theta}_{ext}$ from $u(x-h) = \theta_{ext}u(x) + (1-\theta_{ext})u(x+2h)$ and $u(x+h) = \tilde{\theta}_{ext}u(x-2h) + (1-\tilde{\theta}_{ext})u(x)$, respectively, and making the extrapolations exact for the fundamental solution of the homogeneous equation. Next we will prove that our interpolation scheme yields the exact sequences for the inverse matrix on the next finer grid.

THEOREM 3.2. *Let $A_H = \text{tridiag}(a_H, c_H, b_H)$ be the coarse grid tridiagonal stiffness matrix with entries a_H, b_H, c_H as in* (3.7), (3.8), *and* (3.9), *respectively, and let $u_i^c, v_i^c, x_i^c, y_i^c$ be the sequences that represent the inverse A_H^{-1} as in* (3.1). *Let $A_h = \text{tridiag}(a_h, c_h, b_h)$ be the fine grid tridiagonal stiffness matrix with entries a_h, b_h, c_h. Using the interpolation coefficients* (3.10), (3.11) *as well as extrapolation coefficients* (3.12), (3.13) *to obtain the fine sequences $u_i^f, v_i^f, x_i^f, y_i^f$ from the coarse sequences u_i^c, v_i^c, x_i^c, y_i^c, we will produce the exact sequences for the inverse stiffness matrix A_h^{-1} for the fine grid, i.e.,*

$$(A_h^{-1})_{ij} = \begin{cases} u_i^f v_j^f & : \quad i \le j, \\ x_i^f y_j^f & : \quad i > j. \end{cases}$$

We will need the following auxiliary results to prove Theorem 3.2.

LEMMA 3.3. *Let $f(\beta, h) := h^{-1}\mathcal{B}(\beta h)$, where $\mathcal{B}(x) = \frac{x}{e^x-1}$ denotes the Bernoulli function. Then there holds*

$$f(\beta, 2h) = \frac{f(\beta,h)^2}{f(\beta,h) + f(-\beta,h)}.$$

Proof.

$$\begin{aligned} f(\beta,2h) = (2h)^{-1}\mathcal{B}(2\beta h) &= \frac{\beta}{e^{2\beta h}-1} = \frac{\beta e^{\beta h}(e^{-\beta h}-1)^2}{e^{\beta h}(e^{2\beta h}-1)(e^{-\beta h}-1)^2} \\ &= \frac{\beta(e^{\beta h}-1)(e^{-\beta h}-1)}{(e^{\beta h}-1)^2[(e^{-\beta h}-1)-(e^{\beta h}-1)]} = \frac{h^{-1}\frac{\beta h}{e^{\beta h}-1}\cdot\frac{\beta h}{e^{\beta h}-1}}{\frac{\beta h}{e^{\beta h}-1} - \frac{\beta h}{e^{-\beta h}-1}} \\ &= \frac{f(\beta,h)^2}{f(\beta,h)+f(-\beta,h)}. \qquad \square \end{aligned}$$

COROLLARY 3.4. *The following relationships hold between matrix entries a_H, c_H, b_H of the coarse grid stiffness matrix and a_h, c_h, b_h of the fine grid stiffness matrix:*

$$a_H = -\frac{a_h^2}{c_h}, \qquad b_H = -\frac{b_h^2}{c_h}, \qquad c_H = \frac{a_h^2 + b_h^2}{c_h}, \tag{3.14}$$

$$a_h = (e^{-\beta h} + 1)a_H, \quad b_h = (e^{\beta h} + 1)b_H. \tag{3.15}$$

Proof. The proof of (3.14) follows directly by applying Lemma 3.3:

$$a_H = -f(-\beta, 2h) = -\frac{f(-\beta, h)^2}{f(-\beta, h) + f(\beta, h)} = -\frac{a_h^2}{-a_h - b_h} = -\frac{a_h^2}{c_h},$$

and analogously for b_H, c_H. The relationship in (3.15) follows from the direct calculation

$$a_H = -\frac{\mathcal{B}(-2\beta h)}{2h} = -\frac{-2\beta h}{2h(e^{-2\beta h} - 1)} = \frac{\beta h}{h} \cdot \frac{1}{(e^{-\beta h} - 1)(e^{-\beta h} + 1)} = \frac{a_h}{e^{-\beta h} + 1},$$

which may be shown analogously for b_H. ∎

Proof of Theorem 3.2. Let $C = (c_{ij})$ with

$$c_{ij} = \begin{cases} u_i^f v_j^f & : \quad i \le j, \\ x_i^f y_j^f & : \quad i > j, \end{cases}$$

be the matrix with entries computed from the fine, interpolated sequences. We will show that $C \cdot A_h = I$, where I denotes the identity matrix. For a complete proof, we need to distinguish the following three cases: (a) $i < j$ (upper diagonal entries involving sequences u_i, v_i); (b) $i > j + 1$ (lower diagonal entries involving sequences x_i, y_i); (c) $i = j$ or $i = j + 1$ (entries on or directly below the diagonal involving all four sequences). In the following, e_i denotes the ith unit vector.

Case (a). Let $i < j$. We need to show that $e_i^T C A_h e_j = 0$. If j is even, i.e., $j = 2\tilde{j}$, and $3 < j < 2n - 1$, then

$$\begin{aligned} e_i^T C A_h e_j &= v_i(u_{j-1}^f b_h + u_j^f c_h + u_{j+1}^f a_h) \\ &= v_i \left((\theta u_{\tilde{j}-1}^c + \vartheta u_{\tilde{j}}^c) b_h + u_{\tilde{j}}^c c_h + (\theta u_{\tilde{j}}^c + \vartheta u_{\tilde{j}+1}^c) a_h \right) \\ &\overset{(3.10),(3.11),(3.14)}{=} v_i \left(u_{\tilde{j}-1}^c b_H + u_{\tilde{j}}^c \underbrace{\frac{-2a_h b_h + c_h^2}{c_h}}_{c_H} + u_{\tilde{j}+1}^c a_H \right) \\ &= 0 \end{aligned}$$

since $u_{\tilde{j}}^c$ are the terms of the sequence in the representation of the exact inverse of A_H. If j is odd, i.e., $j = 2\tilde{j} + 1$, and $j < 2n + 1$, then there holds

$$\begin{aligned} e_i^T C A_h e_j &= v_i(u_{j-1}^f b_h + u_j^f c_h + u_{j+1}^f a_h) \\ &= v_i \left(u_{\tilde{j}}^c b_h + (\theta u_{\tilde{j}}^c + \vartheta u_{\tilde{j}+1}^c) c_h + u_{\tilde{j}+1}^c a_h \right) \\ &= v_i \left(u_{\tilde{j}}^c b_h + \left(\frac{-b_h}{c_h} u_{\tilde{j}}^c + \frac{-a_h}{c_h} u_{\tilde{j}+1}^c \right) c_h + u_{\tilde{j}+1}^c a_h \right) \\ &= 0. \end{aligned}$$

In the cases $j = 2, j = 2n$, and $j = 2n + 1$, we need to use interpolation as well as extrapolation. We will show that $(CA_h)_{12} = 0$; the other two cases can be shown in a similar way.

$$\begin{aligned} e_1^T C A_h e_2 &= v_1(u_1^f b_h + u_2^f c_h + u_3^f a_h) \\ &\overset{(3.5),(3.12)}{=} v_i\left((\theta_{ext}u_1^c + (1-\theta_{ext})u_2^c)b_h + u_1^c c_h + (u_1^c\theta + u_2^c\vartheta)a_h\right) \\ &= v_i\left(u_1^c(\theta_{ext}b_h + c_h + \theta a_h) + u_2^c((1-\theta_{ext})b_h + \vartheta a_h)\right) \\ &= 0 \end{aligned}$$

since

$$\begin{aligned} \theta_{ext}b_h + c_h + \theta a_h &= \theta_{ext}b_h + (-a_h - b_h) + \theta a_h \\ &= \left(\frac{1-e^{3\beta h}}{1-e^{2\beta h}} - 1\right) b_h + \underbrace{(\theta - 1)a_h}_{-a_H} \\ &= \frac{e^{2\beta h} - e^{3\beta h}}{1-e^{2\beta h}} \frac{\beta h}{h(e^{\beta h}-1)} + \frac{\beta H}{H(e^{-\beta H}-1)} = 0 \end{aligned}$$

and

$$\begin{aligned} (1-\theta_{ext})b_h + \vartheta a_h &= -\left(\theta_{ext}b_h - b_h + (1-\theta)a_h\right) \\ &= -\left(\theta_{ext}b_h + c_h + \theta a_h\right) = 0. \end{aligned}$$

Case (b). This case can be proven analogous to Case (a) by replacing the sequences u_i, v_i by x_i, y_i, respectively.

Case (c). If $i = j$ is even, i.e., $i = 2\tilde{i}$, then there holds

$$\begin{aligned} (CA_h)_{i,i} &= x_i^f y_{i-1}^f b_h + v_i^f u_i^f c_h + v_i^f u_{i+1}^f a_h \\ &= x_{\tilde{i}}^c(\theta y_{\tilde{i}-1}^c + (1-\theta)y_{\tilde{i}})b_h + u_{\tilde{i}}^c v_{\tilde{i}}^c c_h + v_{\tilde{i}}^c(\theta u_{\tilde{i}}^c + (1-\theta)u_{\tilde{i}+1})a_h \\ &= x_{\tilde{i}}^c y_{\tilde{i}-1}^c b_H - \frac{a_h b_h}{c_h} x_{\tilde{i}}^c y_{\tilde{i}}^c + u_{\tilde{i}}^c v_{\tilde{i}}^c c_h - \frac{a_h b_h}{c_h} v_{\tilde{i}}^c u_{\tilde{i}}^c + v_{\tilde{i}}^c u_{\tilde{i}+1}^c a_H \\ &\overset{(3.14)}{=} x_{\tilde{i}}^c y_{\tilde{i}-1}^c b_H + u_{\tilde{i}}^c v_{\tilde{i}}^c c_h + v_{\tilde{i}}^c u_{\tilde{i}+1}^c a_H \\ &= (C_H A_H)_{\tilde{i},\tilde{i}} = 1, \end{aligned}$$

where we used $u_i v_i = x_i y_i$ in the fourth equality above. ($u_i v_i = x_i y_i$ follows from the proof of Theorem 3.1.) If $i = j$ is odd, we need to distinguish three subcases $i = j = 1$, $i = j = 2\tilde{i} + 1$ for $1 \le \tilde{i} < n$ and $i = j = 2n + 1$. These proofs are similar to those above and are omitted. ∎

In the case of $\beta = 0$, this example becomes the self-adjoint Laplace problem $-u'' = f$ with Dirichlet boundary conditions. Here, the interpolation coefficients are given by

$$\theta = \vartheta = \frac{1}{2},$$

and the extrapolation coefficients are

$$\theta_{ext} = \frac{3}{2}, \quad \vartheta_{ext} = -\frac{1}{2}.$$

For this particular case, the sequences u_i, v_i that represent the inverse stiffness matrix are known explicitly due to the fact that the exact inverse A_H^{-1} is of the form (see also [16])

$$A_H^{-1} = H^2 \begin{pmatrix} n & n-1 & \cdots & 2 & 1 \\ n-1 & 2(n-1) & \cdots & 4 & 2 \\ \vdots & \vdots & \ddots & & \vdots \\ 2 & 4 & & \ddots & n-1 \\ 1 & 2 & \cdots & n-1 & n \end{pmatrix},$$

i.e., its entries can be directly computed as

$$\left(A_H^{-1}\right)_{ij} = \begin{cases} u_i v_j & \text{if} \quad i \geq j, \\ \left(A_H^{-1}\right)_{ji} & \text{if} \quad j > i, \end{cases} \tag{3.16}$$

where

$$u_i = H(n-i+1) \qquad \text{and} \qquad v_j = Hj. \tag{3.17}$$

In the general case of $\beta \neq 0$, an alternative proof of Theorem 3.2 can be given based on the following nice property of the Scharfetter–Gummel discretization: Here, the entries of the inverse A^{-1} are the values of the Green's function of the corresponding differential operator (with Dirichlet boundary conditions) taken on the mesh. In particular, for

$$G(x,y) = \begin{cases} \frac{x(1-y)\mathcal{B}(\beta)}{\mathcal{B}(\beta-\beta y)\mathcal{B}(\beta x)} & : \quad x \leq y, \\ \frac{y(1-x)\mathcal{B}(-\beta)}{\mathcal{B}(\beta x-\beta)\mathcal{B}(-\beta y)} & : \quad x > y, \end{cases}$$

one may show that $(A^{-1})_{ij} = G(ih, jh)$ [18]. Setting $u_i = \frac{ih}{\mathcal{B}(\beta ih)}$, $v_j = \frac{(1-jh)\mathcal{B}(\beta)}{\mathcal{B}(\beta-\beta jh)}$, $x_i = \frac{(1-ih)\mathcal{B}(-\beta)}{\mathcal{B}(\beta ih-\beta)}$, and $y_j = \frac{jh}{\mathcal{B}(-\beta jh)}$ yields the four sequences required in (3.1). Then the proof is completed by noting that these sequences are of the form $u_i = f(ih)$, where $f(x) = \alpha e^{-\beta x} + \gamma$ are fundamental solutions of the homogeneous equation, for which the interpolation is made exact.

4. Multilevel $\mathcal{H}$-matrices for two-dimensional problems. A generalization of the interpolation approach to several spatial dimensions is relatively straightforward at least from the standpoint of how to construct the low rank (Rk-)blocks of the fine grid inverse. This can be done by interpolation as in the one-dimensional case, at least if the fine grid clustering is adjusted to the coarse grid clustering, as will be explained in detail later. In the higher-dimensional case, however, we will not be able to produce an exact fine grid inverse from a coarse grid inverse. This is due to the unknown values on the boundaries of the blocks (clusters), which will typically lead to a residual at these boundaries and therefore prevent the construction of an arbitrarily accurate approximation of the inverse. These approximate inverses that are inexpensive to construct, however, still provide good preconditioners for Krylov subspace methods for a variety of model problems. In the remainder of this section, we will construct the fine grid cluster tree from the coarse grid cluster tree in subsection 4.1. In subsection 4.2, we will generalize the interpolation scheme for one-dimensional problems to compute prolongations of the Rk-blocks of the inverse of the coarse stiffness matrix to obtain the respective Rk-blocks for the inverse of the fine matrix. Subsection 4.3 deals with the computation of inadmissible, full matrix blocks and a subsequent "correction" step for the previously computed Rk-blocks.

4.1. Multilevel cluster tree construction. In order to apply a clusterwise prolongation scheme, we can no longer construct the cluster tree for the fine grid indices independently of the cluster tree for the coarse grid indices. In fact, we will construct the fine grid cluster tree from the geometric fine grid information (coordinates of vertex locations) and the coarse grid cluster tree. Whereas the standard cluster tree construction (see Figure 2.1 for an illustration) starts from the root (which is the full index set I), we will begin with the construction of the leaves of the fine cluster tree and then construct all coarser levels by appropriate unions of clusters.

Let T_c denote the coarse cluster tree, and let $\mathcal{L}(T_c) = \{v_1^c, v_2^c, \ldots, v_m^c\}$ denote the m leaves of this tree with labels v_i^c. We now construct m leaf clusters v_i^f of the fine cluster tree as follows:

1. If a (fine grid) vertex x corresponds to a coarse grid vertex, then $x \in v_i^f \iff x \in v_i^c$.
2. If a (fine grid) vertex x results from refinement of the edge between coarse grid vertices y and z, then
   (a) if y and z both correspond to degrees of freedom (i.e., are not on the boundary), then $x \in v_i^f$ OR $x \in v_j^f$ where $y \in v_i^c$ and $z \in v_j^c$;
   (b) if only y but not z corresponds to a degree of freedom, then $x \in v_i^f$ where $y \in v_i^c$;
   (c) if neither y nor z correspond to degrees of freedom, then $x \in v_i^f$ where v_i^f is a fine grid cluster that contains at least one neighbor of x in the fine grid.

We note that the fine grid vertices are not associated with clusters in a uniquely determined way. For example, if a (fine grid) vertex x results from the refinement of the edge between coarse grid vertices y and z which belong to different coarse grid clusters, then x is (randomly) inserted into one of the corresponding fine grid clusters. The construction of fine leaf clusters is illustrated by a simple example in Figure 4.1. Here, case 2(a) applies, e.g., to vertex 39, which results from refinement of the edge between vertices 6 and 7, which both belong to cluster v_2^c. Thus vertex 39 has to be placed into cluster v_2^f. Vertex 32, on the other side, can be placed in either one of the two clusters v_1^f or v_2^f. Case 2(b) occurs, e.g, for vertex 16. Since its only coarse grid neighbor (vertex 0) belongs to v_1^c, we conclude $16 \in v_1^f$. Case 2(c) applies, e.g., to vertex 15, which results from refinement of an edge with two boundary endpoints. Vertex 15 has to belong to the same cluster as one of its neighbors (vertices 14 and 19), i.e., $15 \in v_3^f$.

Once all the fine grid vertices are assigned to leaf clusters, they are assigned "degree-of-freedom" names. Here, the order of the leaf clusters determines the ordering of the unknowns up to the ordering within a leaf which is arbitrary. In the example given in Figure 4.1, the clustering usually leads to different degree-of-freedom names on the coarse and fine grid, e.g., vertex (or index) 6 corresponds to degree of freedom 3 on the coarse grid and degree of freedom 15 on the fine grid, respectively (in Table 4.1, $dof2idx_c(3) = dof2idx_f(15) = 6$).

Next we construct the lower level clusters from the leaves up to the root in the following canonical way: Let the subscript ℓ denote the level of a cluster in the coarse cluster tree, and let $v_{i,\ell}^c, v_{j,\ell}^c$ be two coarse grid clusters that have a joint predecessor $v_{k,\ell-1}^c$, i.e., $v_{k,\ell-1}^c = v_{i,\ell}^c \cup v_{j,\ell}^c$. We then define a fine cluster $v_{k,\ell-1}^f$ (corresponding to the coarse cluster $v_{k,\ell-1}^c$) by $v_{k,\ell-1}^f = v_{i,\ell}^f \cup v_{j,\ell}^f$.

In the case of the example given in Figure 4.1, there will be two clusters on level 1,

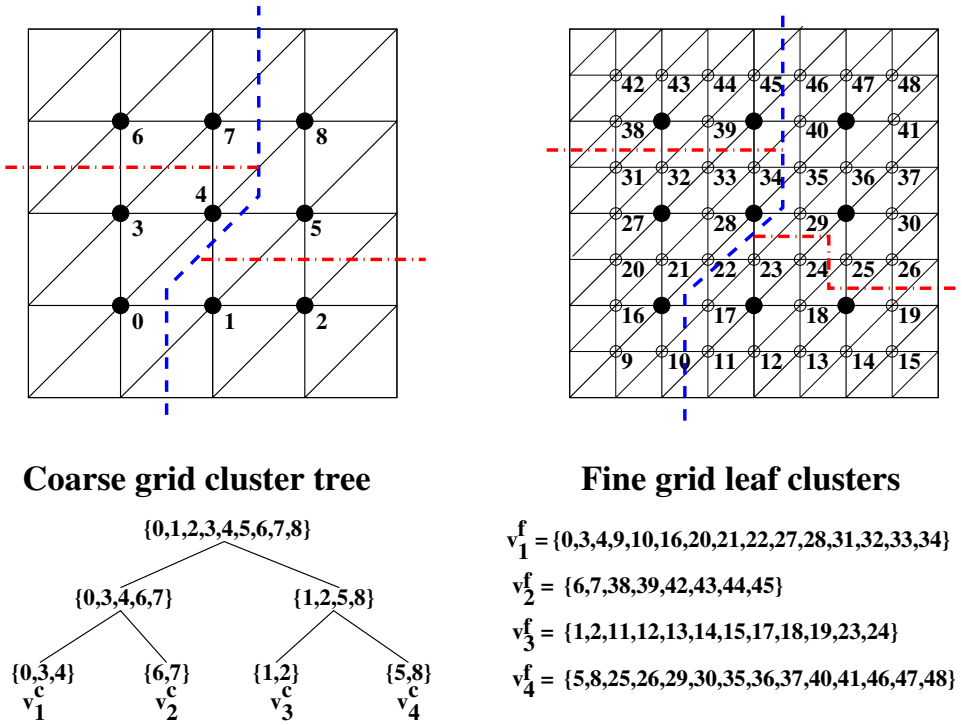


FIG. 4.1. *A coarse grid cluster tree (left) and the (leaves of the) resulting fine grid cluster tree.*

TABLE 4.1
"Degree of freedom to vertex" (dof2idx) arrays for the coarse and fine grids.

DoF	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	...
$dof2idx_c$	0	3	4	6	7	1	2	5	8								
$dof2idx_f$	0	3	4	9	10	16	20	21	22	27	28	31	32	33	34	6	...

namely, $v^f_{1,1} := v^f_1 \cup v^f_2$ and $v^f_{2,1} := v^f_3 \cup v^f_4$. Finally, we obtain the root on level 0 to be the full index set $v^f_{1,0} := v^f_{1,1} \cup v^f_{2,1} = (I = \{0, 1, \ldots, 48\})$.

Once the index cluster tree is given, a block index cluster tree is computed in the canonical way (see (2.4)).

Alternatively, we could construct a coarse cluster tree from a given fine cluster tree by simply deleting vertices that are in the fine grid but not the coarse grid from all clusters. In the case of regularly refined grids, these two approaches yield very similar (or identical) resulting cluster trees. For adaptively refined grids, however, some "unlucky" cases may occur for this approach, e.g., empty coarse clusters or a fine grid vertex belonging to a different cluster than its two coarse grid neighbors.

The multilevel clustering as described in this subsection provides a matrix block structure to approximate the inverse stiffness matrix. The computation of the actual matrix data will be described in the next two subsections for Rk-blocks and full matrix blocks, respectively.

4.2. Prolongation of Rk-blocks. Let $s^c \times t^c$ be an admissible block cluster of the coarse grid, and let $s^f \times t^f$ be the corresponding fine grid cluster. Let

$(A_c^{-\mathcal{H}})_{s^c \times t^c} = \sum_{i=1}^{k} a_i^c (b_i^c)^T$ be the rank k representation of the corresponding (admissible) matrix block in the coarse grid approximation to the inverse stiffness matrix. In order not to overload our presentation with sub- and superscripts, let $c = (\gamma_1, \gamma_2, \ldots, \gamma_{|s^c|})^T$ be one of the vectors a_i^c in the Rk representation of $(A_c^{-\mathcal{H}})_{s^c \times t^c}$. A prolongation of c to $Pc := f := (\phi_1, \phi_2, \ldots, \phi_{|s^f|})^T$ is computed by

$$\phi_i = \begin{cases} \gamma_k & : \quad dof2idx_f(i) = dof2idx_c(k), \\ \theta\gamma_j + \vartheta\gamma_k & : \quad \left(\begin{array}{c} dof2idx_f(i) \text{ results from refinement of} \\ \text{edge between vertices } dof2idx_c(j) \text{ and} \\ dof2idx_c(k), \text{ both of which belong to the} \\ \text{coarse cluster } s^c \end{array} \right), \\ \gamma_k & : \quad \left(\begin{array}{c} dof2idx_c(k) \text{ is the only coarse grid neighbor} \\ \text{of } dof2idx_f(i) \text{ that belongs to } s^c \end{array} \right), \\ 0 & : \quad \left(\begin{array}{c} dof2idx_f(i) \text{ does not have any} \\ \text{coarse grid neighbors in } s^c \end{array} \right). \end{cases} \tag{4.1}$$

Here, "$dof2idx_f$" and "$dof2idx_c$" denote the arrays that store the mappings of the degrees of freedoms to the vertex names for the fine and coarse grids, respectively. An example was given in Table 4.1. θ and ϑ are the problem-dependent interpolation factors. This prolongation is well defined as a result of the multilevel construction of the cluster tree: This construction guarantees $dof2idx_c(s^c) \subset dof2idx_f(s^f)$ (and also $dof2idx_c(t^c) \subset dof2idx_f(t^f)$) so that the prolongation of any entry corresponding to a coarse grid vertex is well defined (first case in (4.1)). A fine grid vertex may have two, one, or no coarse grid neighbors belonging to the respective coarse cluster s^c (second through fourth cases in (4.1)). We note that we expect the fourth case to be the exception. In fact, in a regularly refined grid of the type as shown in Figure 4.1, this case applies to only two corner vertices (vertices 15 and 42 in this example). The prolongation of vectors b_i^c is performed analogously, possibly with different interpolation factors $\tilde{\theta}$ and $\tilde{\vartheta}$. We therefore obtain an approximation (in rank k representation) $(A_f^{-\mathcal{H}})_{s^f \times t^f} = \sum_{i=1}^{k} a_i^f (b_i^f)^T$ to the corresponding fine grid matrix block of the inverse stiffness matrix.

4.3. Computation of full blocks and selected Rk corrections. We assume that the standard admissibility condition (2.2) is used in the construction of the block cluster tree. However, instead of representing nonadmissible, off-diagonal blocks as full matrices, we use the Rk representation with full rank. Therefore, we obtain a block structure with full blocks only on the diagonal of the matrix. We furthermore assume that the prolongations of all off-diagonal Rk-blocks have already been computed. We now compute the full diagonal blocks by solving $A_f A_f^{-\mathcal{H}} = I_f$ exactly for these diagonal blocks, i.e., for every diagonal block $s^f \times s^f$, we solve $(A_f)_{s^f \times J} \cdot (A_f^{-\mathcal{H}})_{J \times s^f} = I_{s^f \times s^f}$ for $(A_f^{-\mathcal{H}})_{s^f \times s^f}$:

$$(A_f^{-\mathcal{H}})_{s^f \times s^f} := ((A_f)_{s^f \times s^f})^{-1} \cdot \left(I_{s^f \times s^f} - (A_f)_{s^f \times (J \setminus s^f)} \cdot (A_f^{-\mathcal{H}})_{(J \setminus s^f) \times s^f} \right). \tag{4.2}$$

Here, A_f denotes the fine stiffness matrix, I_f denotes the identity matrix, and J denotes the set of degrees of freedom on the fine grid.

We provide an illustration in Figure 4.2. Here, we computed approximations to the inverse of the Laplace matrix for a relatively small problem size (441×441). On

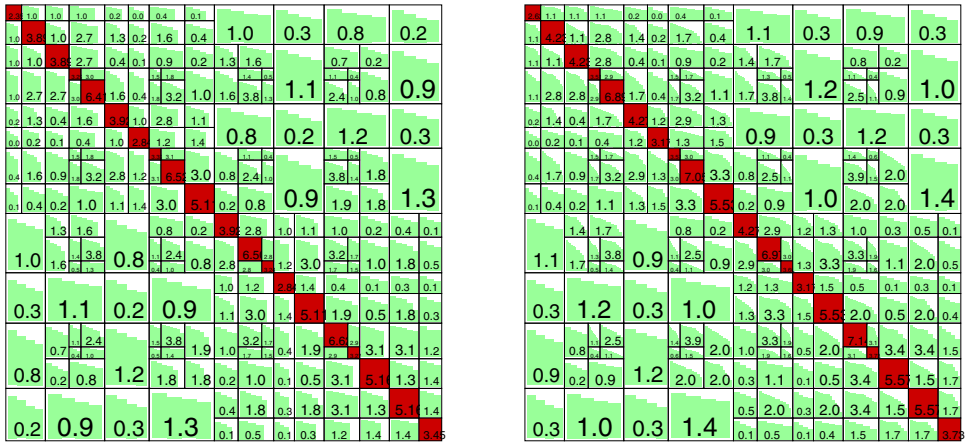


FIG. 4.2. *The multilevel inverse (left) and the $\mathcal{H}$-inverse (right).*

the left, we show the $\mathcal{H}$-matrix obtained by the prolongation. The blocks along the diagonal are full matrix blocks. The number printed in these blocks is the infinity norm of the respective block. The off-diagonal blocks show the first four singular values (on a logarithmic scale) with the largest one being printed in the block. On the right in Figure 4.2, we show the matrix obtained by the $\mathcal{H}$-inversion algorithm applied to the fine matrix (which may differ from a best possible $\mathcal{H}$-approximation to the exact inverse matrix). We note the resemblance of the two matrices.

The particular matrix structure with full blocks only on the diagonal is motivated by the ease of computation for these remaining full matrix blocks. In fact, it is unclear how the full blocks should be computed (efficiently) in the case of several off-diagonal full blocks. However, a clear disadvantage with respect to the approximation accuracy can be observed as a result of this block structure: Let $(A_c^{-\mathcal{H}})_{s^c \times t^c}$ be such an off-diagonal, nonadmissible matrix block in Rk representation with rank $k_c \leq \min\{|s^c|, |t^c|\} (\leq n_{min})$, i.e., full rank. Since $s^c \times t^c$ does not satisfy the admissibility condition, neither do we expect the corresponding fine block $s^f \times t^f$ to be admissible (it may, however, be admissible in exceptional cases) so that an accurate approximation to $(A_f^{-\mathcal{H}})_{s^f \times t^f}$ could require up to a full rank $k_f \leq \min\{|s^f|, |t^f|\}$. The Rk prolongation, however, will provide an approximation of only rank k_c ($\approx \frac{1}{4} k_f$).

We expect the largest errors in entries $(A_f^{-\mathcal{H}})_{ij}$ for which $(A_f)_{ij} \neq 0$, but the respective vertices $dof2idx_f(i)$ and $dof2idx_f(j)$ belong to different leaves in the fine cluster tree. Let A_{corr} be a "correction" matrix of such a sparsity structure. Its nonzero entries $(A_{corr})_{ij}$ are computed by

$$(A_{corr})_{ij} := -\frac{(e_i^T A_f) \cdot (A_f^{-\mathcal{H}} e_j)}{(A_f)_{ii}}$$

to obtain $e_i^T A_f (A_f^{-\mathcal{H}} + A_{corr}) e_j = 0$. Here, e_i again denotes the ith unit vector. We then compute the correction

$$\tilde{A}_f^{-\mathcal{H}} := A_f^{-\mathcal{H}} \oplus_{\mathcal{H}} A_{corr}. \tag{4.3}$$

In section 5 we will show numerical results for both matrices $A_f^{-\mathcal{H}}$ and $\tilde{A}_f^{-\mathcal{H}}$.

We close this section with a remark on how the construction of a multilevel $\mathcal{H}$-inverse differs from the standard geometric multigrid method: In the coarse grid correction step of the geometric two-grid method, we compute an approximate error $e^f \approx PA_c^{-1}Rr^f$ where r^f, e^f denote the fine grid residual and error, respectively, and P and R denote the prolongation and restriction operators. The matrix $PA_c^{-1}R$ may be interpreted as an approximate fine grid inverse, obtained by prolongation from the coarse grid inverse. Our multilevel inverse differs from $PA_c^{-1}R$ since we compute prolongations only restricted to block clusters, and entries on diagonal blocks are not obtained from prolongation but by the explicit computation given in (4.2).

5. Numerical results. We will provide numerical results for the iterative solution of the (discretized) convection-diffusion equation

$$\begin{aligned} -\epsilon\Delta u + b\cdot\nabla u &= f \qquad \text{in } \Omega = (-1,1)^2,\\ u &= g \qquad \text{on } \partial\Omega \end{aligned}$$

for varying values for ϵ and various convections $b \in \{b_{none}, b_{xline}, b_{diag}, b_{circ}, b_{recirc}\}$, where

$$\begin{aligned} b_{none}(x,y) &= (0,0)^T && \text{(no convection)},\\ b_{xline}(x,y) &= (1,0)^T && \text{(constant convection along the } x\text{-axis)},\\ b_{diag}(x,y) &= \sqrt{2}^{-1}(1,1)^T && \text{(diagonal convection)},\\ b_{circ}(x,y) &= \left(\frac{1}{2}-y,\ x-\frac{1}{2}\right)^T && \text{(circular convection)},\\ b_{recirc}(x,y) &= (4x(x-1)(1-2y),\ -4y(y-1)(1-2x))^T && \text{(recirculating convection)}. \end{aligned}$$

A finite element discretization using Tabata's upwind triangle scheme [17, Chap. III, sect. 3.1.1] leads to systems of linear equations $A_h x_h = f_h$, where h denotes the grid width of the underlying (regular) triangulation. Throughout we use the adaptive $\mathcal{H}$-arithmetic (2.5) which is typically superior to fixed ranks, in particular for highly non-symmetric problems [14]. All the numerical results given subsequently use problem *independent* interpolation factors $\theta = \vartheta = \frac{1}{2}$ since the problem-dependent factors did not yield significant improvements. All tests were performed on a Dell Precision 470n Workstation with a Xeon 3.2GHz processor using the standard $\mathcal{H}$-matrix library HLIB (cf. http://www.hlib.org). We choose $x_0 = 100\cdot(1,\ldots,1)^T$ as the initial vector to solve the discrete system by a preconditioned BiCGstab iteration. We iterate until either the maximum number of 100 iterations has been reached, or until the residual has been reduced by a factor of 10^{-6}. We compute an averaged convergence rate $\sqrt[n]{r_n/r_0}$, where $r_n = \|b - Ax_n\|_2$ denotes the norm of the nth residual.

In our first set of experiments (see Table 5.1), we test the dependence of the multilevel $\mathcal{H}$-inverse on the adaptive accuracy δ (2.5) used in the $\mathcal{H}$-arithmetic. These tests have been performed for a fixed coarse problem size $n_c = 10000$ and refined fine grid with $n_f = 40401$ unknowns. In the upper part of the table, we record results obtained for the Laplace equation, and the lower block shows the results for a convection-dominated problem with circular convection. The times for the computation of the coarse $\mathcal{H}$-inversion and prolongation are given in seconds. For comparison, we also record the time for the $\mathcal{H}$-inversion of the fine stiffness matrix. We observe that the coarse $\mathcal{H}$-inversion and prolongation together take about half as much time as the fine $\mathcal{H}$-inversion (at least for adaptive $\mathcal{H}$-accuracy $\delta \le 10^{-3}$). However, the

TABLE 5.1
Dependence on adaptive accuracy, $n_f = 40401$, $n_c = 10000$.

$\mathcal{H}$-accuracy δ	10^{-1}	10^{-2}	10^{-3}	10^{-4}
$\epsilon = 1.0$, $b = b_{none}$ (Laplace)				
Coarse $\mathcal{H}$-inversion	21.7	32.1	44.5	55.4
Prolongation	4.2	4.4	4.7	4.9
Fine $\mathcal{H}$-inversion	35.4	56.7	88.1	124.6
Storage in MB	64	96	145	191
Convergence rates	0.97/0.98/0.96	0.96/0.96/0.71	0.95/0.95/3.5e-4	0.95/0.95/3.7e-10
Steps	100/100/100	100/100/43	100/100/3	100/100/2
Iteration time	7/7/10	10/11/6	14/16/1	17/19/0.5
$\epsilon = 10^{-8}$, $b = b_{circ}$				
Coarse $\mathcal{H}$-inversion	15.1	17.3	20.1	22.5
Prolongation	5.3	4.8	5.4	4.9
Fine $\mathcal{H}$-inversion	25.4	37.1	48.7	62.0
Storage in MB	76.8	102.4	125.8	146.5
Convergence rate	0.48/0.50/0.38	0.31/0.29/4.7e-3	0.31/0.29/1.1e-8	0.31/0.29/0
Steps	23/22/17	13/14/4	13/14/2	13/14/1
Iteration time	2/2/1	1.3/1.7/0.6	1.7/1.9/0.4	1.7/3.2/0.2

TABLE 5.2
Dependence on problem size (n_f, n_c), fixed $\mathcal{H}$-accuracy $\delta = 1e-3$.

n_f	10201	20449	40401	80089	160801
$\epsilon = 1.0$, $b = b_{none}$ (Laplace)					
Coarse $\mathcal{H}$-inversion	7.8	12.4	44.5	67.1	226.8
Prolongation	1.0	2.4	4.7	11.3	21.9
Storage in MB	27.8	59.1	145.7	295.9	719.7
Convergence rate	0.94/0.65	0.94/0.77	0.95/0.95	0.95/0.95	0.95/0.98
Steps	100/37	100/63	100/100	100/100	100/100
Iteration time	2.7/1.2	5.3/3.9	14/16	28/32	70/80
$\epsilon = 10^{-8}$, $b = b_{circ}$					
Coarse $\mathcal{H}$-inversion	3.8	6.1	20.1	31.2	96.8
Prolongation	0.9	2.3	5.4	11.1	21.5
Storage in MB	23.0	50.6	125.8	271.9	654.6
Convergence rate	0.27/0.28	0.25/0.27	0.31/0.29	0.34/0.37	0.45/0.47
Steps	13/12	12/12	13/14	14/16	21/20
Iteration time	0.3/0.4	0.6/0.7	1.7/2.1	3.5/4.6	12.9/14.4

convergence rates are disappointing, at least for the Laplace problem: Listed are the three convergence rates obtained for the multilevel $\mathcal{H}$-inverse without Rk correction (see section 4.3), with Rk correction, and using the $\mathcal{H}$-inverse computed by the standard $\mathcal{H}$-inversion of the fine stiffness matrix. Whereas the latter one yields an almost exact method for adaptive accuracy $\delta = 10^{-4}$, the multilevel inverses with or without Rk correction only yield convergence rates of 0.95 independent of the $\mathcal{H}$-accuracy. However, our main motivation is to find a good preconditioner not for the Laplace problem but rather for highly convection-dominated problems. And here, the convergence rates of 0.30 reported in the bottom part of Table 5.1 look rather satisfactory.

In Table 5.2, we choose a fixed $\mathcal{H}$-accuracy of $\delta = 10^{-3}$ and perform tests for varying problem sizes ranging from 10201 to 160801 unknowns on the fine grid. We notice that even though the convergence rate shows some dependence on the problem size, this dependence is very moderate. The convergence rates obtained with Rk correction only differ from those obtained without this correction for the smaller problem sizes for the Laplace problem. They were almost identical for the convection-dominated circular problem.

TABLE 5.3
Dependence on convection direction and dominance ($n_f = 80089, n_c = 19881$).

ϵ	1.0	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
Setup times (coarse $\mathcal{H}$-inverse & prolongation)							
b_{xline}	67/12	70/12	79/12	70/12	53/12	43/11	37/12
$b_{diagonal}$	68/12	71/12	78/12	70/12	58/13	51/12	43/12
b_{circ}	68/11	70/12	75/12	64/12	50/11	42/11	37/11
b_{recirc}	68/11	71/12	69/12	55/11	43/11	36/11	32/11
Convergence rates & iteration times							
b_{xline}	0.95/32	0.50/7	0.17/3.6	0.36/6	0.77/20	0.85/30	0.87/30
$b_{diagonal}$	0.96/32	0.45/6.5	0.10/2.6	0.22/4.7	0.76/22	0.92/36	0.93/28
b_{circ}	0.95/32	0.86/31	0.44/6.7	0.22/3.7	0.33/4.5	0.36/5	0.36/4.8
b_{recirc}	0.95/33	0.76/19	0.35/5.3	0.32/4.5	0.51/7.2	0.74/13	0.78/17

Next, we perform tests for various convection directions and dominance, leaving the problem size fixed ($n_f = 80089$) and also the adaptive accuracy $\delta = 10^{-3}$ (see Table 5.3). All times are measured in seconds. We notice that the setup times for the $\mathcal{H}$-inversion of the coarse matrix become smaller as the convection dominance increases (i.e., as ϵ decreases). It is interesting to note that the best convergence is observed for an ϵ between 10^{-2} and 10^{-1}, i.e., initially the convergence behavior improves as ϵ becomes smaller, but eventually the rates increase again as $\epsilon \to 0$.

Conclusions. A very efficient construction of an approximate $\mathcal{H}$-inverse from a given coarser level $\mathcal{H}$-inverse via interpolation has been presented. Whereas the approximation accuracy of such an inverse cannot be made arbitrarily good (as it can, e.g., for the explicit $\mathcal{H}$-inverse by increasing the local ranks), it is still sufficiently accurate to obtain efficient preconditioners for iterative solution methods such as BiCGstab. The numerical results have been presented for a two-level method applied to a two-dimensional convection-diffusion problem. Extensions to more than two levels as well as applications to three-dimensional problems are relatively straightforward.

REFERENCES

[1] R. BANK, J. BÜRGLER, W. FICHTNER, AND R. K. SMITH, *Some upwinding techniques for finite element approximations of convection-diffusion equations*, Numer. Math., 58 (1990), pp. 185–202.

[2] R. E. BANK AND S. GUTSCH, *Hierarchical basis for the convection-diffusion equation on unstructured meshes*, in Proceedings of the 9th International Conference on Domain Decomposition Methods, D. K. P. Bjørstad and M. Espedal, eds., DDM.org, 1998, pp. 251–265; available online at www.ddm.org/DD9/index.html.

[3] M. BEBENDORF AND W. HACKBUSCH, *Existence of $\mathcal{H}$-matrix approximants to the inverse FE-matrix of elliptic operators with L^∞-coefficients*, Numer. Math., 95 (2003), pp. 1–28.

[4] S. BÖRM, L. GRASEDYCK, AND W. HACKBUSCH, *Hierarchical Matrices*, Lecture Notes 21, Max-Planck-Institute for Mathematics in the Sciences, Leipzig, Germany, 2003; available online at www.mis.mpg.de/preprints/ln/.

[5] M. CAPOVANI, *Sulla determinazione della inversa delle matrici tridiagonal a blocchi*, Calcolo, 7 (1970), pp. 295–303.

[6] P. CONCUS, G. H. GOLUB, AND G. MEURANT, *Block preconditioning for the conjugate gradient method*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 220–252.

[7] F. GANTMACHER AND M. KREIN, *Sur les matrices completement non negatives et oscillatoires*, Compositio Math., 4 (1937), pp. 445–470.

[8] L. GRASEDYCK AND W. HACKBUSCH, *Construction and arithmetics of $\mathcal{H}$-matrices*, Computing, 70 (2003), pp. 295–334.

[9] L. GRASEDYCK, W. HACKBUSCH, AND S. LE BORNE, *Adaptive geometrically balanced clustering of $\mathcal{H}$-matrices*, Computing, 73 (2003), pp. 1–23.

[10] W. Hackbusch, *A sparse matrix arithmetic based on $\mathcal{H}$-matrices. Part* I: *Introduction to $\mathcal{H}$-matrices*, Computing, 62 (1999), pp. 89–108.
[11] W. Hackbusch, L. Grasedyck, and S. Börm, *An introduction to hierarchical matrices*, Math. Bohem., 127 (2002), pp. 229–241.
[12] W. Hackbusch and B. Khoromskij, *A sparse $\mathcal{H}$-matrix arithmetic. Part* II: *Application to multi-dimensional problems*, Computing, 64 (2000), pp. 21–47.
[13] W. Hackbusch, B. Khoromskij, and R. Kriemann, *Hierarchical matrices based on a weak admissibility criterion*, Computing, 73 (2004), pp. 207–243.
[14] S. Le Borne and L. Grasedyck, *$\mathcal{H}$-matrix preconditioners in convection-dominated problems*, SIAM J. Matrix Anal. Appl., 27 (2006), pp. 1172–1183.
[15] G. Meurant, *A review on the inverse of symmetric tridiagonal and block tridiagonal matrices*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 707–728.
[16] R. Nabben, *Decay rates of the inverse of nonsymmetric tridiagonal and band matrices*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 820–837.
[17] H. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations: Convection Diffusion and Flow Problems*, Comput. Math. 24, Springer, Berlin, 1996.
[18] S. L. Sobolev, *Partial Differential Equations of Mathematical Physics*, Dover, Mineola, NY, 1989.
[19] E. E. Tyrtyshnikov, *Incomplete cross approximation in the mosaic-skeleton method*, Computing, 64 (2000), pp. 367–380.
[20] P. Vassilevski, *On some ways of approximating inverses of banded matrices in connection with deriving preconditioners based on incomplete block factorizations*, Computing, 43 (1990), pp. 277–296.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 3, pp. 890–905

EXTREMAL RANKS OF SOME SYMMETRIC MATRIX EXPRESSIONS WITH APPLICATIONS*

YONGGE TIAN† AND YONGHUI LIU‡

Abstract. Suppose $A - BXB^*$, $A - BX - X^*B^*$, and $A - BX + X^*B^*$ are three linear matrix expressions over the field of complex numbers, where A is an Hermitian or skew-Hermitian matrix. In this paper, we consider how to choose an Hermitian or skew-Hermitian matrix X such that $A - BXB^*$ have the maximal and minimal possible ranks, and how to choose X such that $A - BX \pm X^*B^*$ attain the minimal possible ranks. Some applications to Hermitian or skew-Hermitian solutions of matrix equations with symmetric patterns are also given.

Key words. matrix expression, maximal rank, minimal rank, generalized inverse, Moore–Penrose inverse, Hermitian matrix, skew-Hermitian matrix, matrix equation, Hermitian solution, skew-Hermitian solution

AMS subject classifications. 15A03, 15A09, 15A24

DOI. 10.1137/S0895479802415545

1. Introduction.

Throughout this paper, $\mathbb{C}$ denotes the field of complex numbers; the symbols A^*, $r(A)$, and $\mathscr{R}(A)$ stand for the conjugate transpose, the rank, and the range (column space) of matrix $A \in \mathbb{C}^{m\times n}$, respectively; $[\,A,\,B\,]$ denotes a row block matrix consisting of A and B.

For an $m \times n$ matrix A, the Moore–Penrose inverse $A^\dagger$ of A is defined to be the unique solution X to the four Penrose equations

$$\text{(i) } AXA = A, \quad \text{(ii) } XAX = X, \quad \text{(iii) } (AX)^* = AX, \quad \text{(iv) } (XA)^* = XA.$$

A matrix X is called a generalized inverse of A, denoted by A^-, while the collection of all possible g-inverses of A is denoted by $\{A^-\}$. For convenience, the symbols E_A and F_A stand for the two orthogonal projectors $E_A = I_m - AA^\dagger$ and $F_A = I_n - A^\dagger A$. General properties of g-inverses of matrices can be found in [2, 4, 12].

In matrix theory and applications, there are various matrix expressions that involve variable entries. For example,

$$A - BXC, \quad A - B_1X_1C_1 - B_2X_2C_2, \tag{1.1}$$

where X, X_1, and X_2 are variable matrices. In many situations, it is necessary to know the maximal and minimal possible ranks of the matrices with respect to X, X_1, and X_2. These extremal ranks can be used to characterize nonsingularity, rank invariance, range inclusion of the corresponding matrix expressions, as well as solvability conditions of matrix equations.

In addition to (1.1), there are many matrix expressions that have symmetric patterns or involve Hermitian matrices. Some simpler cases are given by

$$A - BXB^*, \quad A - BY - Y^*B^*, \quad A - BY + Y^*B^*, \tag{1.2}$$

*Received by the editors October 1, 2002; accepted for publication (in revised form) by H. J. Werner April 4, 2006; published electronically October 30, 2006.
http://www.siam.org/journals/simax/28-3/41554.html
†School of Economics, Shanghai University of Finance and Economics, Shanghai 200433, China (yongge@mail.shufe.edu.cn).
‡Department of Applied Mathematics, Shanghai Finance University, Shanghai, 201209, China (liuyh@shfc.edu.cn).

where $A \in \mathbb{C}^{m\times m}$ and $B \in \mathbb{C}^{m\times n}$ are given, and $X \in \mathbb{C}^{n\times n}$ and $Y \in \mathbb{C}^{n\times m}$ are variable matrices. In this paper, we consider how to choose X and Y with $X = X^*$ and $Y = Y^*$, or $X = -X^*$ and $Y = -Y^*$, such that the matrix expressions have the maximal and minimal ranks. A direct motivation for this consideration arises from some previous work on solving the matrix equations $BXB^* = A$ and $BY \pm Y^*B^* = A$; see, e.g., [1, 3, 5, 6, 22]. As applications, we shall give the extremal ranks of Hermitian and skew-Hermitian solutions of some well-known linear matrix equations.

Suppose $A \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{m\times k}$, and $C \in \mathbb{C}^{l\times n}$ are given. Then a valuable formula for the rank of the matrix expression $A - BXC$ is

$$r(\,A - BXC\,) = r\begin{bmatrix} A \\ C \end{bmatrix} + r[\,A,\,B\,] - r(M) + r[\,E_{T_1}(\,X + TM^{\dagger}S\,)F_{S_1}\,], \tag{1.3}$$

where

$$M = \begin{bmatrix} A & B \\ C & 0 \end{bmatrix}, \quad T = [\,0,\,I_k\,], \quad S = \begin{bmatrix} 0 \\ I_l \end{bmatrix}, \quad T_1 = TF_M, \quad S_1 = E_M S.$$

The proof of (1.3) can be found in [19]. Observe that the first three terms on the right-hand side of (1.3) are the ranks of the three partitioned matrices consisting of A, B, and C. Hence the variation of the rank of $A - BXC$ with respect to X is determined by the rank of $E_{T_1}(\,X + TM^{\dagger}S\,)F_{S_1}$ with respect to X. It is obvious that there exists a matrix X such that $E_{T_1}(\,X + TM^{\dagger}S\,)F_{S_1} = 0$. This fact enables us to find the maximal and minimal ranks of $E_{T_1}(\,X + TM^{\dagger}S\,)F_{S_1}$ with respect to X. Some previous results on the extremal ranks of $A - BXC$ and related problems can be found in [16, 19].

In order to find the extremal ranks of $A - BXB^*$ with respect to $X = \pm X^*$ through (1.3), we need the following results on ranks of partitioned matrices and solutions of matrix equations.

LEMMA 1.1 (see [9]). *Let $A \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{m\times k}$, and $C \in \mathbb{C}^{l\times n}$ be given. Then*

$$r[\,A,\,B\,] = r(A) + r(\,B - AA^{\dagger}B\,) = r(B) + r(\,A - BB^{\dagger}A\,), \tag{1.4}$$

$$r\begin{bmatrix} A \\ C \end{bmatrix} = r(A) + r(\,C - CA^{\dagger}A\,) = r(C) + r(\,A - AC^{\dagger}C\,), \tag{1.5}$$

$$r\begin{bmatrix} A & B \\ C & 0 \end{bmatrix} = r(B) + r(C) + r[\,(\,I_m - BB^{\dagger}\,)A(\,I_n - C^{\dagger}C\,)\,]. \tag{1.6}$$

LEMMA 1.2 (see [10, 11]). *Let $A \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{n\times m}$, and $C \in \mathbb{C}^{m\times m}$. Then the matrix equation $AXB = C$ has an Hermitian solution if and only if the pair of matrix equations $AYB = C$ and $B^*YA^* = C^*$ has a common solution. In this case, the general Hermitian solution to $AXB = C$ can be written as*

$$X = \frac{1}{2}(\,Y + Y^*\,), \tag{1.7}$$

where Y is the general common solution to the pair $AYB = C$ and $B^*YA^* = C^*$.

Applying Lemma 1.2 to $AXA^* = B$ yields the following well-known result.

LEMMA 1.3 (see [8]). *Let $A \in \mathbb{C}^{m\times n}$ and $B = B^* \in \mathbb{C}^{m\times m}$. Then the matrix equation $AXA^* = B$ has an Hermitian solution if and only if $\mathscr{R}(B) \subseteq \mathscr{R}(A)$. In this case, the general Hermitian solution can be written as*

$$X = A^{\dagger}B(A^{\dagger})^* + F_A V + V^*F_A, \tag{1.8}$$

where $V \in \mathbb{C}^{n\times n}$ is arbitrary. The Hermitian solution to $AXA^* = B$ is unique if and only if $A^\dagger A = I_n$, i.e., $r(A) = n$.

The following result can be shown similarly.

LEMMA 1.4. *Let $A \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{n\times m}$, and $C \in \mathbb{C}^{m\times m}$. Then the matrix equation $AXB = C$ has a skew-Hermitian solution if and only if the pair of matrix equations $AYB = C$ and $B^*YA^* = -C^*$ has a common solution. In this case, the general skew-Hermitian solutions to $AXB = C$ can be written as*

$$X = \frac{1}{2}(Y - Y^*), \tag{1.9}$$

where Y is the general common solution to the pair $AYB = C$ and $B^*YA^* = -C^*$.

Applying Lemma 1.4 to $AXA^* = B$ with $B = -B^*$ yields the following result.

LEMMA 1.5. *Let $A \in \mathbb{C}^{m\times n}$ and $B = -B^* \in \mathbb{C}^{m\times m}$. Then the matrix equation $AXA^* = B$ has a skew-Hermitian solution if and only if $\mathscr{R}(B) \subseteq \mathscr{R}(A)$. In this case, the general skew-Hermitian solution can be written as*

$$X = A^\dagger B(A^\dagger)^* + F_A V - V^* F_A, \tag{1.10}$$

where $V \in \mathbb{C}^{n\times n}$ is arbitrary. The skew-Hermitian solution to $AXA^* = B$ is unique if and only if $A^\dagger A = I_n$, i.e., $r(A) = n$.

2. Ranks of $A - BXB^*$ with respect to Hermitian and skew-Hermitian matrix X. It is well known that if A is Hermitian, then $A^\dagger$ is Hermitian, too, and $AA^\dagger = A^\dagger A$. Let $A \in \mathbb{C}^{m\times m}$, $B \in \mathbb{C}^{m\times n}$ and let

$$M = \begin{bmatrix} A & B \\ B^* & 0 \end{bmatrix}, \quad S = \begin{bmatrix} 0 \\ I_n \end{bmatrix}, \quad S_1 = S - MM^\dagger S. \tag{2.1}$$

If A is Hermitian, then M is Hermitian, too. In this case, applying (1.3) to

$$p(X) = A - BXB^*, \tag{2.2}$$

where $X \in \mathbb{C}^{n\times n}$, and simplifying give the following rank equality:

$$r(A - BXB^*) = 2r[A, B] - r(M) + r[F_{S_1}(X + S^*M^\dagger S)F_{S_1}]. \tag{2.3}$$

Notice that there exists an $X \in \mathbb{C}^{n\times n}$ such that

$$F_{S_1}(X + S^*M^\dagger S)F_{S_1} = 0. \tag{2.4}$$

Hence we have the following result.

THEOREM 2.1. *Let $p(X)$ be as given in (2.2) with $A = A^*$, and let M, S, and S_1 be as given in (2.1). Then the following hold:*

(a) *The maximal and minimal ranks of $p(X)$ with respect to $X = X^*$ are given by*

$$\max_{X=X^*\in\mathbb{C}^{n\times n}} r(A - BXB^*) = r[A, B], \tag{2.5}$$

$$\min_{X=X^*\in\mathbb{C}^{n\times n}} r(A - BXB^*) = 2r[A, B] - r\begin{bmatrix} A & B \\ B^* & 0 \end{bmatrix}. \tag{2.6}$$

(b) *The general expression of Hermitian matrix X satisfying (2.5) can be written as*

$$X = -S^*M^\dagger S + U, \tag{2.7}$$

where $U = U^* \in \mathbb{C}^{n\times n}$ *is chosen such that* $r(F_{S_1}UF_{S_1}) = r(F_{S_1})$, *say,* $U = F_{S_1}$.

(c) *The general expression of Hermitian matrix* X *satisfying* (2.6) *can be written as*

$$X = -S^*M^\dagger S + S_1^*V^* + VS_1, \tag{2.8}$$

where $V \in \mathbb{C}^{n\times(m+n)}$ *is arbitrary.*

Proof. We first see from (2.3) that

$$\max_{X=X^*} r(\,A - BXB^*\,) = 2r[\,A,\,B\,] - r(M) + \max_{X=X^*} r[\,F_{S_1}(\,X + S^*M^\dagger S\,)F_{S_1}\,], \tag{2.9}$$

$$\min_{X=X^*} r(\,A - BXB^*\,) = 2r[\,A,\,B\,] - r(M) + \min_{X=X^*} r[\,F_{S_1}(\,X + S^*M^\dagger S\,)F_{S_1}\,]. \tag{2.10}$$

Note that

$$\max_{X=X^*} r[\,F_{S_1}(\,X + S^*M^\dagger S\,)F_{S_1}\,] = \max_{U=U^*} r(F_{S_1}UF_{S_1}) = r(F_{S_1}). \tag{2.11}$$

The matrix X satisfying (2.9) can be written as (2.7). Substituting this into (2.9) gives

$$\max_{X=X^*} r(\,A - BXB^*\,) = 2r[\,A,\,B\,] - r(M) + r(F_{S_1}). \tag{2.12}$$

Note that $r(F_{S_1}) = r(\,I_n - S_1^\dagger S_1\,) = n - r(S_1)$ and

$$r(S_1) = r(\,S - MM^\dagger S\,) = r[\,M,\,S\,] - r(M) = n + r[\,A,\,B\,] - r(M) \quad \text{(by (1.4))}.$$

Hence $r(F_{S_1}) = r(M) - r[\,A,\,B\,]$. Substituting this into (2.12) gives (2.5). Also note that

$$\min_{X=X^*} r[\,F_{S_1}(\,X + S^*M^\dagger S\,)F_{S_1}\,] = \min_{U=U^*} r(F_{S_1}UF_{S_1}) = 0.$$

Hence we have (2.6). Solving the matrix equation $F_{S_1}(\,X + S^*M^\dagger S\,)F_{S_1} = 0$ for X by Lemma 1.3 gives (2.8). □

Suppose the Hermitian matrix A is nonnegative definite, i.e., A can be written as $A = NN^*$ for some N. In this case,

$$r\begin{bmatrix} A & B \\ B^* & 0 \end{bmatrix} = r[\,A,\,B\,] + r(B);$$

see Rao and Mitra [12]. Hence if A is nonnegative definite, then (2.6) reduces to

$$\min_{X=X^*} r(\,A - BXB^*\,) = r[\,A,\,B\,] - r(B). \tag{2.13}$$

Corollary 2.2. *Let* $p(X)$ *be as given in* (2.2) *with* $A = A^*$. *Then the rank of* $p(X)$ *is invariant with respect to the choice of Hermitian matrix* X, *i.e.,* $r(\,A - BXB^*\,) = r(A)$ *for any* $X = X^*$, *if and only if*

$$\mathscr{R}\begin{bmatrix} B \\ 0 \end{bmatrix} \subseteq \mathscr{R}\begin{bmatrix} A \\ B^* \end{bmatrix}. \tag{2.14}$$

In particular, suppose A *is nonnegative definite. Then the rank of* $p(X)$ *is invariant for any Hermitian* X *if and only if* $B = 0$.

COROLLARY 2.3. *Let $p(X)$ be as given in* (2.2) *with $A = A^*$, and let M and S be as given in* (2.1). *Then the matrix satisfying* (2.6) *is unique if and only if*

$$r(B) = n \quad \text{and} \quad r\begin{bmatrix} A & B \\ B^* & 0 \end{bmatrix} = r[\,A,\,B\,] + r(B). \tag{2.15}$$

In this case, the unique matrix satisfying (2.6) *is*

$$X = -S^*M^\dagger S. \tag{2.16}$$

Proof. It can be seen from (2.10) that the Hermitian X satisfying (2.6) is unique if and only if the Hermitian solution to (2.4) is unique. From Lemma 1.3, the Hermitian solution of the equation is unique if and only if $r(F_{S_1}) = n$, which is equivalent to $r(M) = r[\,A,\,B\,] + n$. Also note that $r(M) \le r[\,A,\,B\,] + r(B)$. Hence $r(M) = r[\,A,\,B\,] + n$ is equivalent to (2.15). □

If A is skew-Hermitian, i.e., $A = -A^*$, then $A^\dagger$ is skew-Hermitian, too, and $AA^\dagger = A^\dagger A$. Let

$$M = \begin{bmatrix} A & B \\ -B^* & 0 \end{bmatrix}, \quad S = \begin{bmatrix} 0 \\ I_n \end{bmatrix}, \quad S_1 = S - MM^\dagger S. \tag{2.17}$$

Then $M = -M^*$ and $MM^\dagger = M^\dagger M$. In such a case, applying (1.3) to $p(X)$ in (2.2) with $A = -A^*$ yields the following rank identity

$$r(\,A - BXB^*\,) = 2r[\,A,\,B\,] - r(M) + r[\,F_{S_1}(\,X + S^*M^\dagger S\,)F_{S_1}\,]. \tag{2.18}$$

THEOREM 2.4. *Let $p(X)$ be as given in* (2.2) *with $A = -A^*$, and let M, S, and S_1 be as given in* (2.17). *Then the following hold:*

(a) *The maximal and minimal ranks of $p(X)$ with respect to $X = -X^* \in \mathbb{C}^{n\times n}$ are given by*

$$\max_{X=-X^*\in\mathbb{C}^{n\times n}} r(\,A - BXB^*\,) = r[\,A,\,B\,], \tag{2.19}$$

$$\min_{X=-X^*\in\mathbb{C}^{n\times n}} r(\,A - BXB^*\,) = 2r[\,A,\,B\,] - r\begin{bmatrix} A & B \\ -B^* & 0 \end{bmatrix}. \tag{2.20}$$

(b) *The general expression of matrix X satisfying* (2.19) *can be written as*

$$X = -S^*M^\dagger S + U, \tag{2.21}$$

where $U = -U^* \in \mathbb{C}^{n\times n}$ is chosen such that $r(F_{S_1}UF_{S_1}) = r(F_{S_1})$, say, $U = iF_{S_1}$ with $i^2 = -1$.

(c) *The general expression of the matrix X satisfying* (2.20) *can be written as*

$$X = -S^*M^\dagger S + S_1^*V^* - VS_1, \tag{2.22}$$

where $V \in \mathbb{C}^{n\times(m+n)}$ is arbitrary.

Proof. We see from (2.18) that

$$\max_{X=-X^*} r(\,A - BXB^*\,) = 2r[\,A,\,B\,] - r(M) + \max_{X=-X^*} r[\,F_{S_1}(\,X + S^*M^\dagger S\,)F_{S_1}\,], \tag{2.23}$$

$$\min_{X=-X^*} r(\,A - BXB^*\,) = 2r[\,A,\,B\,] - r(M) + \min_{X=-X^*} r[\,F_{S_1}(\,X + S^*M^\dagger S\,)F_{S_1}\,]. \tag{2.24}$$

Note that

$$\max_{X=-X^*} r[\,F_{S_1}(\,X+S^*M^\dagger S\,)F_{S_1}\,] = \max_{U=-U^*} r(F_{S_1}UF_{S_1}) = r(F_{S_1}). \tag{2.25}$$

The matrix X satisfying (2.25) can be written as (2.21). Substituting this into (2.23) gives

$$\max_{X=-X^*} r(\,A-BXB^*\,) = 2r[\,A,\,B\,]-r(M)+r(F_{S_1}). \tag{2.26}$$

It is easy to verify that $r(F_{S_1}) = r(M) - r[\,A,\,B\,]$. Substituting this into (2.26) gives (2.19). Also note that

$$\min_{X=-X^*} r[\,F_{S_1}(\,X+S^*M^\dagger S\,)F_{S_1}\,] = \min_{U=-U^*} r(F_{S_1}UF_{S_1}) = 0.$$

Hence we have (2.20). Solving the matrix equation $F_{S_1}(\,X+S^*M^\dagger S\,)F_{S_1}=0$ for X by Lemma 1.5 gives (2.22). □

COROLLARY 2.5. *Let $p(X)$ be as given in* (2.2) *with $A=-A^*$. Then the rank of $p(X)$ is invariant with respect to the choice of skew-Hermitian matrix X, i.e., $r(\,A-BXB^*\,)=r(A)$ for any $X=-X^*$, if and only if*

$$\mathscr{R}\begin{bmatrix} B \\ 0 \end{bmatrix} \subseteq \mathscr{R}\begin{bmatrix} A \\ B^* \end{bmatrix}. \tag{2.27}$$

COROLLARY 2.6. *Let $p(X)$ be as given in* (2.2) *with $A=-A^*$, and let M and S be as given in* (2.17). *Then the matrix satisfying* (2.20) *is unique if and only if*

$$r(B)=n \quad \textit{and} \quad r\begin{bmatrix} A & B \\ -B^* & 0 \end{bmatrix} = r[\,A,\,B\,]+r(B). \tag{2.28}$$

In this case, the unique matrix satisfying (2.21) *is*

$$X=-S^*M^\dagger S. \tag{2.29}$$

An extension of $A-BXB^*$ is

$$p(X,\,Y)=A-BXB^*-CYC^*, \tag{2.30}$$

where $A\in\mathbb{C}^{m\times m}$, $B\in\mathbb{C}^{m\times n}$, and $C\in\mathbb{C}^{m\times k}$ are given. In such a case, it is of interest to seek analytical expressions of the extremal ranks of $p(X,\,Y)$ with respect to $X=\pm X^*$ and $Y=\pm Y^*$. It has been shown in [15, 17] that

$$\begin{aligned}\max_{X_1\in\mathbb{C}^{p_1\times q_1},\,X_2\in\mathbb{C}^{p_2\times q_2}} & r(\,A-B_1X_1C_1-B_2X_2C_2\,) \\ &= \min\left\{ r[\,A,\,B_1,\,B_2\,],\ \ r\begin{bmatrix} A \\ C_1 \\ C_2 \end{bmatrix},\ \ r\begin{bmatrix} A & B_1 \\ C_2 & 0 \end{bmatrix},\ \ r\begin{bmatrix} A & B_2 \\ C_1 & 0 \end{bmatrix} \right\},\end{aligned} \tag{2.31}$$

$$\begin{aligned}\min_{X_1\in\mathbb{C}^{p_1\times q_1},\,X_2\in\mathbb{C}^{p_2\times q_2}} r(\,A-B_1X_1C_1-B_2X_2C_2\,) &= r\begin{bmatrix} A \\ C_1 \\ C_2 \end{bmatrix} + r[\,A,\,B_1,\,B_2\,] \\ &\quad + \max\{s_1,\ s_2\},\end{aligned} \tag{2.32}$$

where

$$s_1 = r\begin{bmatrix} A & B_1 \\ C_2 & 0 \end{bmatrix} - r\begin{bmatrix} A & B_1 & B_2 \\ C_2 & 0 & 0 \end{bmatrix} - r\begin{bmatrix} A & B_1 \\ C_1 & 0 \\ C_2 & 0 \end{bmatrix},$$

$$s_2 = r\begin{bmatrix} A & B_2 \\ C_1 & 0 \end{bmatrix} - r\begin{bmatrix} A & B_1 & B_2 \\ C_1 & 0 & 0 \end{bmatrix} - r\begin{bmatrix} A & B_2 \\ C_1 & 0 \\ C_2 & 0 \end{bmatrix}.$$

Notice that $p(X, Y)$ in (2.30) is a special case of $A - B_1X_1C_1 - B_2X_2C_2$. Hence we have the following conjecture.

CONJECTURE 2.7. *Let $p(X, Y)$ be as given in* (2.30) *with $A = \pm A^*$. Then*

$$\max_{X=\pm X^*,\, Y=\pm Y^*} r(\,A - BXB^* - CYC^*\,) = \min\left\{ r[\,A,\,B,\,C\,],\quad r\begin{bmatrix} A & B \\ C^* & 0 \end{bmatrix} \right\},$$

$$\min_{X=\pm X^*,\, Y=\pm Y^*} r(\,A - BXB^* - CYC^*\,)$$

$$= 2r[\,A,\,B,\,C\,] + r\begin{bmatrix} A & B \\ C^* & 0 \end{bmatrix} - r\begin{bmatrix} A & B & C \\ B^* & 0 & 0 \end{bmatrix} - r\begin{bmatrix} A & B & C \\ C^* & 0 & 0 \end{bmatrix}.$$

3. The minimal ranks of $A - BX - X^*B^*$ with respect to X. Let $A \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{m\times k}$, and $C \in \mathbb{C}^{l\times n}$ be given. It is shown in [18] that the minimal rank of $A - BX - YC$ with respect to $X \in \mathbb{C}^{k\times n}$ and $Y \in \mathbb{C}^{m\times l}$ is given by the formula

$$\min_{X\in\mathbb{C}^{k\times n},\, Y\in\mathbb{C}^{m\times l}} r(\,A - BX - YC\,) = r\begin{bmatrix} A & B \\ C & 0 \end{bmatrix} - r(B) - r(C); \tag{3.1}$$

a pair of matrices X and Y satisfying (3.1) are given by

$$X = B^\dagger A + UC + (\,I_k - B^\dagger B\,)U_1, \tag{3.2}$$

$$Y = (\,I_m - BB^\dagger\,)AC^\dagger - BU + U_2(\,I_l - CC^\dagger\,), \tag{3.3}$$

where $U \in \mathbb{C}^{k\times l}$, $U_1 \in \mathbb{C}^{k\times n}$, and $U_2 \in \mathbb{C}^{m\times l}$ are arbitrary matrices. Formula (3.1) indicates that there exist two matrices X and Y such that $BX + YC = A$ if and only if

$$r\begin{bmatrix} A & B \\ C & 0 \end{bmatrix} = r(B) + r(C),$$

that is, $E_BAF_C = 0$. This result is well known; see Roth [14]. In this case, (3.2) and (3.3) are the general solutions to $BX + YC = A$.

From (2.31), the maximal rank of $A - BX - YC$ with respect to X and Y is

$$\max_{X\in\mathbb{C}^{k\times n},\, Y\in\mathbb{C}^{m\times l}} r(\,A - BX - YC\,) = \left\{ m,\quad n,\quad r\begin{bmatrix} A & B \\ C & 0 \end{bmatrix} \right\}. \tag{3.4}$$

Letting $A \in \mathbb{C}^{m\times m}$ and $B \in \mathbb{C}^{m\times n}$ and letting $Y = X^*$ and $C = B^*$ in $A - BX - YC$ leads to the matrix expression

$$p(X) = A - BX - X^*B^*. \tag{3.5}$$

We have seen from (1.8) that the general Hermitian solution of $AXA^* = B$ is a special case of (3.5). In this section, we show how to choose X such that $p(X)$ in (3.5) attains the minimal rank.

THEOREM 3.1. *Let $p(X)$ be as given in* (3.5) *with $A = A^*$. Then the minimal rank of $p(X)$ with respect to X is given by*

$$\min_{X\in\mathbb{C}^{n\times m}} r(\,A - BX - X^*B^*\,) = r\begin{bmatrix} A & B \\ B^* & 0 \end{bmatrix} - 2r(B). \tag{3.6}$$

A matrix X satisfying (3.6) *is given by*

$$X = B^\dagger A - \frac{1}{2}B^\dagger ABB^\dagger + UB^* + (\,I_n - B^\dagger B\,)V, \tag{3.7}$$

where both $U = -U^* \in \mathbb{C}^{n\times n}$ and $V \in \mathbb{C}^{n\times m}$ are arbitrary.

Proof. Let $M = \left[\begin{smallmatrix} A & B \\ B^* & 0 \end{smallmatrix}\right]$. Then the rank of M satisfies the inequality

$$r(M) \le r(A) + 2r(B). \tag{3.8}$$

Replacing A in (3.8) with $p(X)$ in (3.5) yields the rank inequality

$$r\begin{bmatrix} A - BX - X^*B^* & B \\ B^* & 0 \end{bmatrix} \le r(\,A - BX - X^*B^*\,) + 2r(B). \tag{3.9}$$

It is easy to find by elementary block matrix operations that

$$r\begin{bmatrix} A - BX - X^*B^* & B \\ B^* & 0 \end{bmatrix} = r\begin{bmatrix} A & B \\ B^* & 0 \end{bmatrix} \quad \text{for any } X.$$

Thus we see from (3.9) that

$$r(\,A - BX - X^*B^*\,) \ge r(M) - 2r(B) \quad \text{for any } X. \tag{3.10}$$

Observe that the right-hand side of (3.10) involves no X. Thus $r(M) - 2r(B)$ is a lower bound for the rank of $p(X)$ with respect to X. On the other hand, substituting (3.7) into $p(X)$ in (3.5) yields

$$\begin{aligned} p(X) &= A - BB^\dagger A + \frac{1}{2}BB^\dagger ABB^\dagger - BUB^* - ABB^\dagger + \frac{1}{2}BB^\dagger ABB^\dagger + BUB^* \\ &= A - BB^\dagger A - ABB^\dagger + BB^\dagger ABB^\dagger \\ &= (\,I_m - BB^\dagger\,)A(\,I_m - BB^\dagger\,). \end{aligned}$$

In this case, the rank of $p(X)$ by (1.6) is

$$r[\,p(X)\,] = r[\,(\,I_m - BB^\dagger\,)A(\,I_m - BB^\dagger\,)\,] = r(M) - 2r(B). \tag{3.11}$$

Combining (3.10) with (3.11), we see that $r(M) - 2r(B)$ is the minimal rank of $p(X)$ with respect to X, and a matrix X satisfying (3.6) is given by (3.7). □

If A is nonnegative definite, then (3.6) reduces to

$$\min_{X\in\mathbb{C}^{n\times m}} r(\,A - BX - X^*B^*\,) = r[\,A,\, B\,] - r(B).$$

Comparing this with (2.13) leads to

$$\min_{X=X^*\in\mathbb{C}^{n\times n}} r(\,A - BXB^*\,) = \min_{Y\in\mathbb{C}^{n\times m}} r(\,A - BY - Y^*B^*\,).$$

The matrix equation associated with (3.5) is $BX + X^*B^* = A$. Braden [3] investigated the equation when both A and B are real.

COROLLARY 3.2. *Let $p(X)$ be as given in* (3.5) *with $A = A^*$.*

(a) *The matrix equation $BX + X^*B^* = A$ has a solution if and only if $r\left[\begin{smallmatrix} A & B \\ B^* & 0 \end{smallmatrix}\right] = 2r(B)$, i.e., $E_BAE_B = 0$. In this case, the general solution is given by* (3.7).

(b) *Suppose that A is nonnegative definite. Then the matrix equation $BX + X^*B^* = A$ has a solution if and only if $\mathscr{R}(A) \subseteq \mathscr{R}(B)$. In this case, the general solution is*

$$X = \frac{1}{2}B^\dagger A + UB^* + (\, I_m - B^\dagger B\,)V,$$

where both $U = -U^* \in \mathbb{C}^{n\times n}$ and $V \in \mathbb{C}^{n\times m}$ are arbitrary.

(c) *Suppose that $B \in \mathbb{C}^{m\times m}$ is nonsingular. Then the matrix equation $BX + X^*B^* = A$ always has a solution, and the general solution can be written as*

$$X = \frac{1}{2}B^{-1}A + UB^*,$$

where $U = -U^* \in \mathbb{C}^{n\times n}$ is arbitrary.

Proof. It is easy to verify that under $E_BAE_B = 0$, (3.6) satisfies $BX+X^*B^* = A$. Further suppose X_0 is any matrix satisfying $BX_0 + X_0^*B^* = A$. In this case, let $U = -B^\dagger X_0^* B^\dagger B + \frac{1}{2}B^\dagger A(B^\dagger)^*$ and $V = X_0$ in (3.7). Then

$$\begin{aligned} U + U^* &= -B^\dagger X_0^* B^\dagger B - B^\dagger BX_0(B^\dagger)^* + B^\dagger A(B^\dagger)^* \\ &= -B^\dagger(\, X_0^*B^* + BX_0 - A\,)(B^\dagger)^* = 0 \end{aligned}$$

and

$$\begin{aligned} X &= B^\dagger A - \frac{1}{2}B^\dagger ABB^\dagger + \left[\, -B^\dagger X_0^* B^\dagger B + \frac{1}{2}B^\dagger A(B^\dagger)^* \,\right] B^* + (\, I_m - B^\dagger B\,)X_0 \\ &= B^\dagger A - \frac{1}{2}B^\dagger ABB^\dagger - B^\dagger X_0^* B^* + \frac{1}{2}B^\dagger ABB^\dagger + X_0 - B^\dagger BX_0 \\ &= B^\dagger A - B^\dagger X_0^* B^* + X_0 - B^\dagger(\, A - X_0^*B^*\,) = X_0. \end{aligned}$$

This indicates that any solution of $BX + X^*B^* = A$ can be represented by (3.7). Thus (3.7) is the general solution of $BX + X^*B^* = A$. Parts (b) and (c) follow from (a). □

Suppose $A - BX$ is a matrix expression with $A \in \mathbb{C}^{m\times m}$, $B \in \mathbb{C}^{m\times n}$, and $X \in \mathbb{C}^{n\times m}$. In this case, it is of interest to find X such that $A - BX$ is Hermitian or skew-Hermitian. These kinds of problems are called matrix completion problems of partial matrices in the literature. Many completion problems on determinants, ranks, inverses and generalized inverses, nonnegative definiteness, and eigenvalues of partial matrices, and their applications have been investigated; see, e.g., [7, 13, 20]. Applying Theorem 3.1 to $A - BX$, we obtain the following result.

COROLLARY 3.3. *Let $A \in \mathbb{C}^{m\times m}$ and $B \in \mathbb{C}^{m\times n}$ be given, and let $X \in \mathbb{C}^{n\times m}$ be a variable matrix. Then*

$$\min_{X\in\mathbb{C}^{n\times m}} r[\,(\, A - BX\,) + (\, A - BX\,)^*\,] = r\begin{bmatrix} A + A^* & B \\ B^* & 0 \end{bmatrix} - 2r(B). \tag{3.12}$$

A matrix X satisfying (3.12) *is given by*

$$X = B^\dagger(\, A + A^*\,) - \frac{1}{2}B^\dagger(\, A + A^*\,)BB^\dagger + UB^* + (\, I_n - B^\dagger B\,)V, \tag{3.13}$$

where both $U = -U^* \in \mathbb{C}^{n\times n}$ *and* $V \in \mathbb{C}^{n\times m}$ *are arbitrary. In particular, there exists an* X *such that* $A - BX$ *is skew-Hermitian if and only if*

$$r\begin{bmatrix} A + A^* & B \\ B^* & 0 \end{bmatrix} = 2r(B),$$

in which case a matrix X *such that* $A - BX$ *is skew-Hermitian is given by* (3.13).

From (3.4) we can also give a conjecture on the maximal rank of $A - BX - X^*B^*$ with $A = A^* \in \mathbb{C}^{m\times m}$:

$$\max_{X\in\mathbb{C}^{n\times m}} r(\,A - BX - X^*B^*\,) = \left\{ m,\ \ r\begin{bmatrix} A & B \\ B^* & 0 \end{bmatrix} \right\}. \tag{3.14}$$

We shall show (3.14) through the generalized singular value decompositions of matrices in a forthcoming paper.

Moreover for the matrix expression $A - BXC - (BXC)^*$ with $A^* = A$, we have the following conjectures on its maximal and minimal ranks:

$$\max_{X\in\mathbb{C}^{p\times q}} r[\,A - BXC - (BXC)^*\,] = \min\left\{ r[\,A,\,B,\,C^*\,],\ r\begin{bmatrix} A & B \\ B^* & 0 \end{bmatrix},\ r\begin{bmatrix} A & C^* \\ C & 0 \end{bmatrix} \right\},$$
$$\min_{X\in\mathbb{C}^{p\times q}} r[\,A - BXC - (BXC)^*\,] = 2r[\,A,\,B,\,C^*\,] + \max\{s_1,\ \ s_2\},$$

where

$$s_1 = r\begin{bmatrix} A & B \\ B^* & 0 \end{bmatrix} - 2r\begin{bmatrix} A & B & C^* \\ B^* & 0 & 0 \end{bmatrix},\ \ s_2 = r\begin{bmatrix} A & C^* \\ C & 0 \end{bmatrix} - 2r\begin{bmatrix} A & B & C^* \\ C & 0 & 0 \end{bmatrix}.$$

4. The minimal rank of $A - BX + X^*B^*$ with respect to X. A variation of $A - BX - X^*B^*$ in (3.4) is

$$p(X) = A - BX + X^*B^*, \tag{4.1}$$

where $A \in \mathbb{C}^{m\times m}$ and $B \in \mathbb{C}^{m\times n}$ are given and $X \in \mathbb{C}^{n\times m}$ is a variable matrix. By an approach similar to that of section 3, we can show the following several results.

THEOREM 4.1. *Let* $p(X)$ *be as given in* (4.1) *with* $A = -A^*$. *Then*

$$\min_{X\in\mathbb{C}^{n\times m}} r(\,A - BX + X^*B^*\,) = r\begin{bmatrix} A & B \\ B^* & 0 \end{bmatrix} - 2r(B). \tag{4.2}$$

A matrix X *satisfying* (4.2) *is given by*

$$X = B^\dagger A - \frac{1}{2}B^\dagger ABB^\dagger + UB^* + (\,I_n - B^\dagger B\,)V, \tag{4.3}$$

where both $U = U^* \in \mathbb{C}^{n\times n}$ *and* $V \in \mathbb{C}^{n\times m}$ *are arbitrary.*

COROLLARY 4.2. *Let* $A = -A^* \in \mathbb{C}^{m\times m}$ *and* $B \in \mathbb{C}^{m\times n}$ *be given. Then the matrix equation* $BX - X^*B^* = A$ *has a solution if and only if* $r\left[\begin{smallmatrix} A & B \\ B^* & 0 \end{smallmatrix}\right] = 2r(B)$, *i.e.,* $E_BAE_B = 0$. *In this case, the general solution is given by*

$$X = B^\dagger A - \frac{1}{2}B^\dagger ABB^\dagger + UB^* + (\,I_n - B^\dagger B\,)V,$$

where both $U = U^* \in \mathbb{C}^{n\times n}$ *and* $V \in \mathbb{C}^{n\times m}$ *are arbitrary. In particular, suppose* $B \in \mathbb{C}^{m\times m}$ *is nonsingular. Then the equation* $BX + X^*B^* = A$ *is consistent, and the general solution can be written as*

$$X = \frac{1}{2}B^{-1}A + UB^*,$$

where $U = U^* \in \mathbb{C}^{n\times n}$ is arbitrary.

The following two corollaries show how to find X and Y such that $A - BX$ and $A - BX - YC$ are Hermitian.

COROLLARY 4.3. *Let $A \in \mathbb{C}^{m\times m}$ and $B \in \mathbb{C}^{m\times n}$ be given. Then*

$$\min_{X\in\mathbb{C}^{n\times m}} r[\,(A - BX) - (A - BX)^*\,] = r\begin{bmatrix} A - A^* & B \\ B^* & 0 \end{bmatrix} - 2r(B). \tag{4.4}$$

A matrix X satisfying (4.4) *is given by*

$$X = B^\dagger(A - A^*) - \frac{1}{2}B^\dagger(A - A^*)BB^\dagger + UB^* + (I_n - B^\dagger B)V, \tag{4.5}$$

where both $U = U^* \in \mathbb{C}^{n\times n}$ and $V \in \mathbb{C}^{n\times m}$ are arbitrary. In particular, there exists an X such that $A - BX$ is Hermitian if and only if

$$r\begin{bmatrix} A - A^* & B \\ B^* & 0 \end{bmatrix} = 2r(B),$$

in which case a matrix X such that $A - BX$ is Hermitian is given by (4.5).

COROLLARY 4.4. *Let $A \in \mathbb{C}^{m\times m}$, $B \in \mathbb{C}^{m\times n}$, and $C \in \mathbb{C}^{p\times m}$ be given. Then*

$$\begin{aligned} &\min_{X\in\mathbb{C}^{n\times m},\, Y\in\mathbb{C}^{m\times p}} r[\,(A - BX - YC) - (A - BX - YC)^*\,] \\ &= r\begin{bmatrix} A - A^* & B & C^* \\ B^* & 0 & 0 \\ C & 0 & 0 \end{bmatrix} - 2r[\,B,\, C^*\,]. \end{aligned} \tag{4.6}$$

A pair of matrices X and Y satisfying (4.6) *are given by*

$$\begin{aligned} \begin{bmatrix} X \\ Y^* \end{bmatrix} &= [\,B,\, -C^*\,]^\dagger(A - A^*) - \frac{1}{2}[\,B,\, -C^*\,]^\dagger(A - A^*)[\,B,\, -C^*\,][\,B,\, -C^*\,]^\dagger \\ &\quad + U[\,B,\, -C^*\,]^* + (I_{n+p} - [\,B,\, -C^*\,]^\dagger[\,B,\, -C^*\,])V, \end{aligned} \tag{4.7}$$

where both $U = U^* \in \mathbb{C}^{(n+p)\times(n+p)}$ and $V \in \mathbb{C}^{(n+p)\times m}$ are arbitrary. In particular, there exist two matrices X and Y such that $A - BX - YC$ is Hermitian if and only if

$$r\begin{bmatrix} A - A^* & B & C^* \\ B^* & 0 & 0 \\ C & 0 & 0 \end{bmatrix} = 2r[\,B,\, C^*\,],$$

in which case a pair of matrices X and Y such that $A - BX - YC$ is Hermitian is given by (4.7).

Motivated by (2.31) and (2.32), two conjectures on the maximal and minimal ranks of $A - BXC + (BXC)^*$ with $A = -A^*$ are given below:

$$\max_{X\in\mathbb{C}^{p\times q}} r[\,A - BXC + (BXC)^*\,] = \min\left\{ r[\,A,\, B,\, C^*\,],\ \ r\begin{bmatrix} A & B \\ B^* & 0 \end{bmatrix},\ \ r\begin{bmatrix} A & C^* \\ C & 0 \end{bmatrix} \right\},$$

$$\min_{X\in\mathbb{C}^{p\times q}} r[\,A - BXC + (BXC)^*\,] = 2r[\,A,\, B,\, C^*\,] + \max\{s_1,\ \ s_2\},$$

where

$$s_1 = r\begin{bmatrix} A & B \\ B^* & 0 \end{bmatrix} - 2r\begin{bmatrix} A & B & C^* \\ B^* & 0 & 0 \end{bmatrix},\ \ s_2 = r\begin{bmatrix} A & C^* \\ C & 0 \end{bmatrix} - 2r\begin{bmatrix} A & B & C^* \\ C & 0 & 0 \end{bmatrix}.$$

5. Some applications. Linear matrix equations have been a main subject of study in matrix theory and applications. For a given matrix equation, one always wants to know the solvability condition, the general solution, and the uniqueness of solution of the equation. Moreover, if a matrix equation is consistent, one also would like to know properties of solutions of the equation, such as the maximal and minimal ranks of solutions to the equation, the existence of upper-triangular or lower-triangular solutions to the equation, and the existence of Hermitian or skew-Hermitian solutions to the equation. In this section, we give several results on the existence of solutions of matrix equations with special patterns.

From Theorem 3.1, we are able to derive a necessary and sufficient condition for a pair of matrix equations

$$A_1X_1A_1^* = B_1 \quad \text{and} \quad A_2X_2A_2^* = B_2 \tag{5.1}$$

to have a common Hermitian solution, where $A_1 \in \mathbb{C}^{m\times p}$, $B_1 = B_1^* \in \mathbb{C}^{m\times m}$, $A_2 \in \mathbb{C}^{n\times p}$, and $B_2 = B_2^* \in \mathbb{C}^{n\times n}$ are given. Some previous work can be found in [5, 21].

THEOREM 5.1. *Suppose that each of the two linear matrix equations in* (5.1) *is consistent. Then the following hold:*

(a) *The minimal rank of the difference of Hermitian solutions of the two equations in* (5.1) *is*

$$\min_{\substack{X_1=X_1^*,\, X_2=X_2^*\\ A_1X_1A_1^*=B_1\\ A_2X_2A_2^*=B_2}} r(\,X_1 - X_2\,) = r\begin{bmatrix} B_1 & 0 & A_1\\ 0 & -B_2 & A_2\\ A_1^* & A_2^* & 0\end{bmatrix} - 2r\begin{bmatrix} A_1\\ A_2\end{bmatrix}. \tag{5.2}$$

(b) *The pair of matrix equations in* (5.1) *have a common Hermitian solution if and only if*

$$r\begin{bmatrix} B_1 & 0 & A_1\\ 0 & -B_2 & A_2\\ A_1^* & A_2^* & 0\end{bmatrix} = 2r\begin{bmatrix} A_1\\ A_2\end{bmatrix}. \tag{5.3}$$

Proof. If each of the two linear matrix equations in (5.1) is consistent, then by Lemma 1.3 the general Hermitian solutions of the two equations can be written as

$$\begin{aligned} X_1 &= A_1^\dagger B_1(A_1^\dagger)^* + F_{A_1}V_1 + V_1^*F_{A_1},\\ X_2 &= A_2^\dagger B_2(A_2^\dagger)^* - F_{A_2}V_2 - V_2^*F_{A_2}, \end{aligned}$$

where both V_1, $V_2 \in \mathbb{C}^{p\times p}$ are arbitrary. In this case, the difference $X_1 - X_2$ can be written as

$$X_1 - X_2 = A_1^\dagger B_1(A_1^\dagger)^* - A_2^\dagger B_2(A_2^\dagger)^* + [\,F_{A_1},\, F_{A_2}\,]\begin{bmatrix} V_1\\ V_2\end{bmatrix} + [\,V_1^*,\ V_2^*\,]\begin{bmatrix} F_{A_1}\\ F_{A_2}\end{bmatrix}.$$

Applying (3.6) to this expression we find that

$$\begin{aligned} &\min_{\substack{A_1X_1A_1^*=B_1\\ A_2X_2A_2^*=B_2}} r(\,X_1 - X_2\,)\\ &\quad = r\begin{bmatrix} A_1^\dagger B_1(A_1^\dagger)^* - A_2^\dagger B_2(A_2^\dagger)^* & F_{A_1} & F_{A_2}\\ F_{A_1} & 0 & 0\\ F_{A_2} & 0 & 0\end{bmatrix} - 2r\begin{bmatrix} F_{A_1}\\ F_{A_2}\end{bmatrix}. \end{aligned} \tag{5.4}$$

Simplifying the ranks of the above two block matrices by (1.4) and (1.6) and elementary block operations gives

$$
\begin{aligned}
&r\begin{bmatrix} A_1^{\dagger}B_1(A_1^{\dagger})^{*}-A_2^{\dagger}B_2(A_2^{\dagger})^{*} & F_{A_1} & F_{A_2}\\ F_{A_1} & 0 & 0\\ F_{A_2} & 0 & 0 \end{bmatrix}\\
&= r\begin{bmatrix} A_1^{\dagger}B_1(A_1^{\dagger})^{*}-A_2^{\dagger}B_2(A_2^{\dagger})^{*} & I_p & I_p & 0 & 0\\ I_p & 0 & 0 & A_1^{*} & 0\\ I_p & 0 & 0 & 0 & A_2^{*}\\ 0 & A_1 & 0 & 0 & 0\\ 0 & 0 & A_2 & 0 & 0 \end{bmatrix} - 2r(A_1) - 2r(A_2)\\
&= r\begin{bmatrix} 0 & I_p & 0 & 0 & 0\\ I_p & 0 & 0 & A_1^{*} & 0\\ I_p & 0 & 0 & 0 & A_2^{*}\\ -B_1(A_1^{\dagger})^{*} & 0 & -A_1 & 0 & 0\\ B_2(A_2^{\dagger})^{*} & 0 & A_2 & 0 & 0 \end{bmatrix} - 2r(A_1) - 2r(A_2)\\
&= r\begin{bmatrix} 0 & I_p & 0 & 0 & 0\\ I_p & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & -A_1^{*} & A_2^{*}\\ 0 & 0 & -A_1 & B_1 & 0\\ 0 & 0 & A_2 & 0 & -B_2 \end{bmatrix} - 2r(A_1) - 2r(A_2)\\
&= r\begin{bmatrix} B_1 & 0 & A_1\\ 0 & -B_2 & A_2\\ A_1^{*} & A_2^{*} & 0 \end{bmatrix} + 2p - 2r(A_1) - 2r(A_2),
\end{aligned}
$$

$$
r\begin{bmatrix} F_{A_1}\\ F_{A_2} \end{bmatrix} = r\begin{bmatrix} I_p & A_1^{*} & 0\\ I_p & 0 & A_2^{*} \end{bmatrix} - r(A_1) - r(A_2) = r[\,A_1^{*},\,A_2^{*}\,] + p - r(A_1) - r(A_2).
$$

Substituting these two equalities into (5.4) yields (5.2). The result in part (b) is an immediate consequence of (5.2). □

Recall that the general expression of generalized inverses of $A \in \mathbb{C}^{m\times n}$ can be written as

$$A^{-} = A^{\dagger} + F_A V_1 + V_2 E_A,$$

where V_1, $V_2 \in \mathbb{C}^{n\times m}$ are arbitrary. Suppose $A \in \mathbb{C}^{m\times m}$ is Hermitian. Then A^{-} can be written as

$$A^{-} = A^{\dagger} + F_A V_1 + V_2 F_A,$$

where V_1, $V_2 \in \mathbb{C}^{m\times m}$ are arbitrary. In particular, it is easy to verify that the general expression of Hermitian generalized inverses of A can be written as

$$A^{-} = A^{\dagger} + F_A V + V^{*} F_A,$$

where $V \in \mathbb{C}^{m\times m}$ is arbitrary. Applying Theorem 5.1 to a pair of Hermitian matrices gives the following result.

COROLLARY 5.2. *Let A and B be a pair of Hermitian matrices of the same size. Then*

$$\min_{A^{-}=(A^{-})^{*},\,B^{-}=(B^{-})^{*}} r(\,A^{-} - B^{-}\,) = r(\,A - B\,) + r(A) + r(B) - 2r[\,A,\,B\,]. \tag{5.5}$$

Hence A and B have a common Hermitian generalized inverse if and only if

$$r(A-B)=2r[A,\,B]-r(A)-r(B). \tag{5.6}$$

Proof. Since both A and B are Hermitian, we see by Lemma 1.3 that each of $AXA=A$ and $BYB=B$ has an Hermitian solution, i.e., there exist A^- and B^- satisfying $A^-=(A^-)^*$ and $B^-=(B^-)^*$. Thus (5.5) follows from (5.2). ∎

Applying (3.6) to (1.8) yields the following result. The proof is omitted.

COROLLARY 5.3. *Let $A\in\mathbb{C}^{m\times n}$ and $B^*=B\in\mathbb{C}^{m\times m}$ be given and suppose that the matrix equation $AXA^*=B$ has an Hermitian solution. Then*

$$\min_{\substack{AXA^*=B\\ X=X^*}} r(X)=r(B).$$

In addition to Corollary 5.3, we are also able to find minimal ranks of the submatrices in Hermitian solutions of $AXA^*=B$. Suppose that the matrix equation $AXA^*=B$ with $B=B^*$ has an Hermitian solution and write the equation in the partitioned form

$$[A_1,\,A_2]\begin{bmatrix}X_1 & X_2\\ X_3 & X_4\end{bmatrix}\begin{bmatrix}A_1^*\\ A_2^*\end{bmatrix}=B, \tag{5.7}$$

where $A_1\in\mathbb{C}^{m\times n_1}$, $A_2\in\mathbb{C}^{m\times n_2}$, $X_1\in\mathbb{C}^{n_1\times n_1}$, $X_2\in\mathbb{C}^{n_1\times n_2}$, $X_3\in\mathbb{C}^{n_2\times n_1}$, and $X_4\in\mathbb{C}^{n_2\times n_2}$. Also let

$$S_i=\left\{X_i\;\middle|\;[A_1,\,A_2]\begin{bmatrix}X_1 & X_2\\ X_3 & X_4\end{bmatrix}\begin{bmatrix}A_1^*\\ A_2^*\end{bmatrix}=B,\ X_1^*=X_1,\ X_3^*=X_2,\ X_4^*=X_4\right\} \tag{5.8}$$

for $i=1,\dots,4$. By (1.8), the general expressions of $X_i\in S_i$ for $i=1,\dots,4$ can be written as

$$X_1=P_1A^\dagger B(A^\dagger)^*P_1^*+P_1F_AV_1+V_1^*F_AP_1^*, \tag{5.9}$$

$$X_2=X_3^*=P_1A^\dagger B(A^\dagger)^*P_2^*+P_1F_AV_2+V_1^*F_AP_2^*, \tag{5.10}$$

$$X_4=P_2A^\dagger B(A^\dagger)^*P_2^*+P_2F_AV_2+V_2^*F_AP_2^*, \tag{5.11}$$

where $P_1=[I_{n_1},\,0]$ and $P_2=[0,\,I_{n_2}]$; both $V_1\in\mathbb{C}^{n\times n_1}$ and $V_2\in\mathbb{C}^{n\times n_2}$ arbitrary. Applying (3.6) and (3.1) to (5.9), (5.10), and (5.11) yields the following result. The proof is omitted.

COROLLARY 5.4. *Suppose that the matrix equation $AXA^*=B$ has an Hermitian solution and partition the equation as* (5.7). *Then*

$$\min_{X_1\in S_1} r(X_1)=r\begin{bmatrix}B & A_2\\ A_2^* & 0\end{bmatrix}-2r(A_2),$$

$$\min_{X_2\in S_2} r(X_2)=r\begin{bmatrix}B & A_1\\ A_2^* & 0\end{bmatrix}-r(A_1)-r(A_2),$$

$$\min_{X_4\in S_4} r(X_4)=r\begin{bmatrix}B & A_1\\ A_1^* & 0\end{bmatrix}-2r(A_1).$$

Suppose A is a square matrix. An Hermitian matrix X is called an Hermitian $\{i,\dots,j\}$-inverse of A, denoted by $A_h^{(i,\dots,j)}$, if it satisfies the $i,\dots,j$th equations in

the four Penrose equations in section 1. In particular, $A_h^{(1)}$, $A_h^{(1,2)}$, $A_h^{(1,3)}$, and $A_h^{(1,4)}$ are four commonly used Hermitian inverses of A. Hermitian generalized inverses of a general square matrix do not necessarily exist. If, however, A is Hermitian, $A_h^{(1)}$, $A_h^{(1,2)}$, $A_h^{(1,3)}$, and $A_h^{(1,4)}$ exist and their general expressions are given as follows:

(a) $A_h^- = A^\dagger + F_A V + V^* F_A$, where V is arbitrary.

(b) $A_h^{(1,2)} = (\, A^\dagger + F_A V\,)A(\, A^\dagger + V^* F_A\,)$, where V is arbitrary.

(c) $A_h^{(1,3)} = A_h^{(1,4)} = A^\dagger + F_A U F_A$, where $U = U^*$ is arbitrary.

The extremal ranks of these three matrices can easily be derived from the results in sections 2 and 3.

The results in sections 3 and 4 can also be used to characterize the symmetry of various projectors associated with the general linear (Gauss–Markov) model $y = X\beta + \varepsilon$. We shall present the corresponding results in another paper.

Acknowledgments. We are grateful to G.P.H. Styan and Y. Takane for helpful remarks on the work in this paper. We also thank anonymous referees for their suggestions on an earlier version of this paper.

REFERENCES

[1] J.K. Baksalary, *Nonnegative definite and positive definite solutions to the matrix equation* $AXA^* = B$, Linear and Multilinear Algebra, 16 (1984), pp. 133–139.

[2] A. Ben-Israel and T.N.E. Greville, *Generalized Inverses: Theory and Applications*, 2nd ed., Springer-Verlag, New York, 2003.

[3] H.W. Braden, *The equations* $A^T X \pm X^T A = B$, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 295–302.

[4] S.L. Campbell and C.D. Meyer Jr., *Generalized Inverses of Linear Transformations*, corrected reprint of the 1979 original, Dover, New York, 1991.

[5] X.-W. Chang and J. Wang, *The Symmetric solutions of the matrix equations* $AX + YA = C$, $AXA^T + BYB^T = C$ *and* $(A^T XA, B^T XB) = (C, D)$, Linear Algebra Appl., 179 (1993), pp. 171–189.

[6] J. Gross, *Nonnegative-definite and positive-definite solutions to the matrix equation* $AXA^* = B$*–revisited*, Linear Algebra Appl., 321 (2000), pp. 123–129.

[7] C.R. Johnson, *Matrix completion problems: A survey*, in Matrix Theory and Applications, C.R. Johnson, ed., Proc. Sympos. Appl. Math. Amer. Math. Soc. 40, AMS, Providence, RI, 1990, pp. 171–198.

[8] C.G. Khatri and S.K. Mitra, *Hermitian and nonnegative definite solutions of linear matrix equations*, SIAM J. Appl. Math., 31 (1976), pp. 579–585.

[9] G. Marsaglia and G.P.H. Styan, *Equalities and inequalities for ranks of matrices*, Linear and Multilinear Algebra, 2 (1974), pp. 269–292.

[10] S.K. Mitra, *A pair of simultaneous linear matrix equations* $A_1 XB_1 = C_1$ *and* $A_2 XB_2 = C_2$ *and a programming problem*, Linear Algebra Appl., 131 (1990), pp. 107–123.

[11] A. Navarra, P.L. Odell, and D.M. Young, *A representation of the general common solution to the matrix equations* $A_1 XB_1 = C_1$ *and* $A_2 XB_2 = C_2$ *with applications*, Comput. Math. Appl., 41 (2001), pp. 929–935.

[12] C.R. Rao and S.K. Mitra, *Generalized Inverse of Matrices and Its Applications*, Wiley, New York, 1971.

[13] L. Rodman and H.J. Woerdeman, *Perturbations, singular values, and ranks of partial triangular matrices*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 278–288.

[14] W.E. Roth, *The equations* $AX - YB = C$ *and* $AX - XB = C$ *in matrices*, Proc. Amer. Math. Soc., 3 (1952), pp. 392–396.

[15] Y. Tian, *The minimal rank completion of a* 3×3 *partial block matrix*, Linear and Multilinear Algebra, 50 (2002), pp. 125–131.

[16] Y. Tian, *The maximal and minimal ranks of some expressions of generalized inverses of matrices*, Southeast Asian Bull. Math., 25 (2002), pp. 745–755.

[17] Y. Tian, *Upper and lower bounds for ranks of matrix expressions using generalized inverses*, Linear Algebra Appl., 355 (2002), pp. 187–214.

[18] Y. Tian, *The minimal rank of the matrix expression $A - BX - YC$*, Missouri J. Math. Sci., 14 (2002), pp. 40–48.
[19] Y. Tian and S. Cheng, *The maximal and minimal ranks of $A - BXC$ with applications*, New York J. Math., 9 (2003), pp. 345–362.
[20] H.J. Woerdeman, *Minimal rank completions for block matrices*, Linear Algebra Appl., 121 (1989), pp. 105–122.
[21] X. Zhang, *The general common Hermitian nonnegative-definite solution to the matrix equations $AXA^* = BB^*$ and $CXC^* = DD^*$*, J. Multivariate Anal., 93 (2005), pp. 257–266.
[22] X. Zhang and M.Y. Zhang, *The rank-constrained Hermitian nonnegative-definite and positive-definite solutions to the matrix equation $AXA^* = B$*, Linear Algebra Appl., 370 (2003), pp. 163–174.

SPECIAL ISSUE ON ACCURATE SOLUTION OF EIGENVALUE PROBLEMS

The occasion for this special issue is the Fifth International Workshop on Accurate Solution of Eigenvalue Problems, which took place in Hagen, Germany from June 29 to July 1, 2004. This issue provides an outlet for papers from the workshop and recognizes advances in the numerical solution of eigenvalue and related problems.

Refined perturbation theory, careful error analyses, and creative algorithms have led to numerical methods that are more accurate and at the same time more efficient. The fourteen papers in this issue are concerned with ordinary eigenvalue problems, eigenvalue problems for matrix polynomials, and singular value decompositions. A common thread is the judicious exploitation of structure in the matrices.

Thanks go to Henk van der Vorst, Mitch Chernoff, and the SIAM staff for providing a home for the special issue and to the guest editors Jesse Barlow, Beresford Parlett, and Kresimir Veselić, who saw to the careful and timely review of the papers.

Ilse C. F. Ipsen
North Carolina State University

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 4, pp. 907–926

COMPUTING THE BIDIAGONAL SVD USING MULTIPLE RELATIVELY ROBUST REPRESENTATIONS*

PAUL R. WILLEMS†, BRUNO LANG‡, AND CHRISTOF VÖMEL§

Abstract. We describe the design and implementation of a new algorithm for computing the singular value decomposition (SVD) of a real bidiagonal matrix. This algorithm uses ideas developed by Großer and Lang that extend Parlett's and Dhillon's multiple relatively robust representations (MRRR) algorithm for the tridiagonal symmetric eigenproblem. One key feature of our new implementation is that k singular triplets can be computed using only $\mathcal{O}(nk)$ storage units and floating point operations, where n is the dimension of the matrix. The algorithm will be made available as routine `xBDSCR` in the upcoming new release of the LAPACK library.

Key words. bidiagonal singular value decomposition, tridiagonal symmetric eigenproblem, MRRR algorithm, coupling relations, LAPACK library

AMS subject classifications. 15A18, 65-04, 65F15

DOI. 10.1137/050628301

1. Introduction.

Starting in the mid 1990s, Dhillon and Parlett developed the algorithm of multiple relatively robust representations (MRRR) that computes k numerically orthogonal eigenvectors of a symmetric tridiagonal matrix $T \in \mathbb{R}^{n\times n}$ with $\mathcal{O}(nk)$ cost [7, 8, 15, 17, 18]. This algorithm has subsequently been extended by Großer and Lang using so-called *coupling relations* [10, 11, 12] for the stable computation of the bidiagonal singular value decomposition (bSVD). Due to recent improvements in the tridiagonal MRRR algorithm (see, e.g., [9, 21, 20]) as well as in the coupling technique itself, the references [11, 12] no longer describe the most efficient implementation of the bidiagonal MRRR algorithm. This present paper focuses on these recent developments and our resulting new implementation, which is to be incorporated as routine `xBDSCR` into the next release of the widely used LAPACK library [1].

Throughout this article, we have tried to present the MRRR algorithm and its adaptation to the bSVD via coupling relations in such a way that readers without prior expertise in this area should be able to follow the arguments and understand the inner workings of the algorithms in an intuitive way. That is, we will present the topics in enough detail but without too much theory, giving all needed references to update readers on the way.

First we will recall some basic concepts and fix our notation. For a bidiagonal matrix $B \in \mathbb{R}^{n\times n}$, the problem bSVD consist of finding orthogonal matrices U and

*Received by the editors April 1, 2005; accepted for publication (in revised form) by B. Parlett June 27, 2005; published electronically December 18, 2006.
http://www.siam.org/journals/simax/28-4/62830.html

†Central Institute for Applied Mathematics, Research Centre Jülich, 52425 Jülich, Germany and Applied Computer Science and Scientific Computing, University of Wuppertal, 42097 Wuppertal, Germany (willems@math.uni-wuppertal.de).

‡Applied Computer Science and Scientific Computing, University of Wuppertal, 42097 Wuppertal, Germany (lang@math.uni-wuppertal.de).

§Computer Science Division, University of California, Berkeley, CA 94720 (voemel@eecs.berkeley.edu). The work of this author has been supported by a grant from the National Science Foundation (Cooperative Agreement ACI-9619020). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

V and a diagonal matrix Σ such that

$$B = U\Sigma V^T. \tag{1.1}$$

We will follow the convention that Σ contains the singular values $\sigma_1, \ldots, \sigma_n$ of B in descending order. The columns u_i of U and v_i of V are called the left and right singular vectors, respectively, of B. For the scope of this paper we will furthermore assume B to be upper bidiagonal, as it differs from the lower bidiagonal case only concerning the roles of U and V.

Any algorithm solving the bSVD should guarantee small deviations from orthogonality for the matrices U and V,

$$||U^TU - I|| = \mathcal{O}(n\epsilon), \quad ||V^TV - I|| = \mathcal{O}(n\epsilon), \tag{1.2}$$

along with small residuals

$$||BV - U\Sigma|| = \mathcal{O}(n\epsilon||B||), \tag{1.3}$$

where ϵ denotes the machine precision. In addition, for some applications (but not all) it is required that the computed singular values approximate the exact ones to high relative accuracy, so our algorithm should be able to deliver this if requested.

The problem bSVD can be reduced to the tridiagonal symmetric eigenproblem (tSEP) in two ways, using three different matrices. One way works with the *normal equations* of B to compute

$$B^TB = V\Sigma^2V^T, \quad BB^T = U\Sigma^2U^T. \tag{1.4}$$

Alternatively, one can use the Jordan–Wielandt form of B to compute U and V simultaneously via

$$\begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix} = Q \begin{bmatrix} -\Sigma & 0 \\ 0 & \Sigma \end{bmatrix} Q^T, \text{ where } Q = \frac{1}{\sqrt{2}} \begin{bmatrix} U & U \\ -V & V \end{bmatrix}. \tag{1.5}$$

Note that (1.5) can be permuted to be symmetric tridiagonal, resulting in the so-called *Golub–Kahan matrix*.

A very efficient method for tSEP is the MRRR algorithm by Dhillon and Parlett, which we will describe in section 2. In practice, the straightforward use of MRRR to solve either (1.4) or (1.5) does not necessarily imply that both (1.2) and (1.3) hold. This has been observed by Großer and Lang [10, 11, 12] and we will explain their results in section 3. They proposed a remedy to this problem using so-called *coupling relations* to adapt the MRRR algorithm for the bSVD.

We have found that in order to get an efficient and robust computer implementation of this method, the strategy presented in [11, 12] could be improved. This is our main contribution and is described alongside the coupling relations in section 4. Finally, in section 5, we describe additional important issues relevant for high performance, and we will compare our implementation with the Divide & Conquer and QR routines from LAPACK [1].

Some words concerning notation: the symbols already introduced above will remain fixed throughout this article. For example, B will always denote the upper bidiagonal matrix of dimension n for which we want to compute its bSVD. We will refer to the diagonal elements of B as a_i and to its off-diagonal elements as b_i. Besides this, we will deal extensively with diagonal matrices D, R and unit lower or upper bidiagonal matrices L, U. These matrices have only a linear number of nontrivial entries

each, which we will denote using a single index as d_i, r_i, l_i, u_i, respectively. It will be convenient to define, for a given m-vector x and $k \in \mathbb{Z}$, $\mathrm{diag}(x, k)$ as a square matrix of order $m + |k|$ with its kth off-diagonal being x and all other entries set to zero. As an example using this notation, we could write B alternatively as

$$B = \mathrm{diag}([a_1, \ldots, a_n], 0) + \mathrm{diag}([b_1, \ldots, b_{n-1}], +1).$$

One possible point of confusion we are aware of is the fact that we use U for the left singular vectors as well as for some upper unit bidiagonal matrix, and u_i for the columns of U as well as for off-diagonal elements. We did not want to break with one of these uses, as they are part of an evolved and widespread standard, and the meaning will always be clear from the context. Additional notation will be specified where needed.

2. The MRRR algorithm. A quite new method for the tridiagonal symmetric eigenproblem is the MRRR algorithm by Dhillon and Parlett (MRRR stands for multiple relatively robust representations). For brevity, we will sometimes call it simply MRRR.

For a symmetric tridiagonal matrix $T \in \mathbb{R}^{n \times n}$, the MRRR algorithm is able to compute k eigenpairs (λ_i, q_i) in optimal $\mathcal{O}(kn)$ time, at the same time guaranteeing small residual norms

$$||Tq_i - \lambda_i q_i|| = \mathcal{O}(n\epsilon||T||) \tag{2.1}$$

and numerically orthogonal eigenvectors with

$$|q_i^T q_j| = \mathcal{O}(n\epsilon), \quad j \neq i. \tag{2.2}$$

For these reasons, together with the fact the MRRR is well suited for parallelization, the MRRR algorithm is sometimes also dubbed the "holy grail."

In this section, we want to give a short overview of the basic principles underlying MRRR, as this will be necessary for the remainder of the paper. We will omit most of the theory, but give enough information such that readers without prior knowledge of the algorithm should be able to get an intuitive understanding. For a more detailed description see [18, 8, 6].

Since LAPACK 3.0, the algorithm has been included as routine `xSTEGR`. There have been many recent improvements of this implementation, especially for the support of partial spectra and better robustness (see, for example, [9]).

In order to describe the MRRR algorithm, we first need to establish the concept of relative distances between eigenvalues, which is defined in slightly varying ways in the literature. In accordance with [8] we will use

$$\mathrm{reldist}(\lambda, \mu) := \frac{|\lambda - \mu|}{|\lambda|}. \tag{2.3}$$

Then the *relative gap* of an eigenvalue is defined as

$$\mathrm{relgap}(\lambda) := \min \{\mathrm{reldist}(\lambda, \mu) \mid \lambda \neq \mu \in spec(T)\}. \tag{2.4}$$

An eigenvalue is (relatively) *isolated*, or a *singleton*, if its relative gap exceeds some threshold (for example 10^{-3}). A group $\lambda_{c:d}$ of successive nonsingleton eigenvalues $\lambda_c, \lambda_{c+1}, \ldots, \lambda_d$ is called a *cluster*.

From a distant point of view, the MRRR algorithm can be seen as a sophisticated variant of inverse iteration without the need for explicit reorthogonalization. A closer perspective reveals the following two simple but elegant ingredients responsible for its immense success:

1. A method based on so-called *twisted factorizations* to compute, for a *relatively* isolated eigenvalue λ, in $\mathcal{O}(n)$ work an eigenvector $\bar{q}$ satisfying

$$|\sin \angle(q, \bar{q})| = \mathcal{O}\left(n\epsilon/\mathrm{relgap}(\lambda)\right), \tag{2.5}$$

where q denotes the true eigenvector. We will describe this technique in more detail shortly; for now let it suffice that twisted factorizations are a generalization of the standard LDL^T and URU^T bidiagonal factorizations.

2. Eigenvectors are shift invariant, but the relative distances of eigenvalues are not. More precisely, if a shift $\mu \approx \lambda$ close to an eigenvalue is chosen, the relative gap of $\lambda' = \lambda - \mu$ with respect to $T' := T - \mu I$ becomes

$$\mathrm{relgap}_{T'}(\lambda') = \mathrm{relgap}_T(\lambda)\frac{|\lambda|}{|\lambda - \mu|} \gg \mathrm{relgap}_T(\lambda).$$

With these two ideas, the obvious approach is to repeatedly shift the matrix until an eigenvalue is relatively isolated and the corresponding eigenvector can be computed using twisted factorizations.

As we will see shortly, in order to make this strategy work, it is necessary that each encountered shifted matrix defines its eigenvalues and eigenvectors to high relative accuracy. To this end, the MRRR algorithm employs the concept of *relatively robust representations (RRRs)* of a matrix. Any set of numbers defining a matrix is called a representation of the matrix. A representation is relatively robust if small relative changes in these numbers cause only small relative changes in the eigenvalues and eigenvectors. If this holds only for some eigenpairs, the representation is called a *partial RRR*.

It is an interesting fact that most tridiagonal matrices represented directly by their diagonal and off-diagonal elements do not have this property, but a representation based on a bidiagonal factorization of the matrix usually does (see [6] for more details). Therefore the algorithm does not work directly on tridiagonal matrices, but on LDL^T factorizations of these matrices instead, such that the data (L, D) form an RRR.

Armed with these concepts and ideas, the MRRR algorithm can now informally be described as follows. First, an RRR (L, D) is found for the original matrix T, possibly by shifting T. Then the eigenvalues of interest are approximated accurately enough to categorize them into singletons and clusters (for example, using bisection or the dqds algorithm [15, 19]). For each singleton, the eigenvector can be computed directly using twisted factorizations. Because the relative gap of a singleton is by definition large enough, this leads to excellent results according to (2.5).

If there is a cluster of eigenvalues, the algorithm chooses a shift τ close to the cluster s.t. $LDL^T - \tau I =: L^+D^+(L^+)^T$ and (L^+, D^+) again forms a partial RRR for the eigenvalues in the cluster. This factorization is computed using the stationary *differential qds algorithm* (dstqds) [15, 6], as shown in Algorithm 2.1. This transformation is carefully designed to allow a mixed relative error analysis; that is, tiny relative changes to the input (L, D) and the output (L^+, D^+) give an exact relation. Note that this property, together with the fact that (L, D) and (L^+, D^+) are ensured to be (partial) RRRs, is essential to guarantee that the shifting process does not spoil the relation between the eigenpairs of $L^+D^+(L^+)^T$ and LDL^T. This allows us to treat $L^+D^+(L^+)^T$ recursively in the same fashion.

Algorithm 2.1 Factorize $LDL^T - \tau I = L^+D^+(L^+)^T = U^+R^+(U^+)^T$ using the differential stationary (left side) and progressive (right side) qds transformations.

DSTQDS	DPQDS
Input: L, D, τ	**Input:** L, D, τ
Output: L^+, D^+, S^+	**Output:** U^+, R^+, P^+
$s_1^+ = -\tau$	$p_n^+ = d_n - \tau$
for $i = 1 : n-1$ **do**	**for** $i = n-1 : -1 : 1$ **do**
$\quad d_i^+ = d_i + s_i^+$	$\quad r_{i+1}^+ = d_i l_i^2 + p_{i+1}^+$
$\quad l_i^+ = d_i l_i / d_i^+$	$\quad u_i^+ = l_i d_i / r_{i+1}^+$
$\quad s_{i+1} = l_i^+ l_i s_i^+ - \tau$	$\quad p_i^+ = p_{i+1}^+ d_i / r_{i+1}^+ - \tau$
endfor	**endfor**
$d_n^+ = d_n + s_n^+$	$r_1^+ = p_1^+$

A convenient way to describe the resulting flow of computation is as a traversal of a *representation tree*. A node in this tree is given by an index range of eigenvalues, a partial RRR for these eigenvalues, and the accumulated shift from the root. Leaf nodes have only one index; otherwise each index of a node is contained in exactly one of the index ranges of the node's children.

Twisted factorizations. We will finish our description of the MRRR algorithm, giving a more detailed explanation of the method used to compute highly accurate eigenvectors with orthogonality levels inversely proportional to the *relative* gaps of the eigenvalues.

Given an RRR (L, D), we can (for example, using bisection) compute an approximation $\bar{\lambda}$ to an eigenvalue λ of LDL^T satisfying

$$|\lambda - \bar{\lambda}| = \mathcal{O}(\epsilon|\lambda|). \tag{2.6}$$

Then the idea is to find a vector $\bar{q}$ with a small *relative* residual

$$||(LDL^T - \bar{\lambda}I)\bar{q}|| = \mathcal{O}(n\epsilon|\bar{\lambda}|). \tag{2.7}$$

The reward is revealed by the classical gap theorem [2, 14] because, if q denotes the true eigenvector, we get the desired result (2.5).

In order to ensure (2.7), a double factorization

$$LDL^T - \bar{\lambda}I = L^+D^+(L^+)^T = U^+R^+(U^+)^T \tag{2.8}$$

is computed using the stationary and progressive differential qds transformations shown in Algorithm 2.1. If one or both of these factorizations do not exist, the following method can be easily modified; see, for example, [8]. Assuming for now that they do exist, this opens n possible ways to compute an approximation $q^{(k)}, 1 \le k \le n$, to the eigenvector q via

$$\begin{aligned} q_k^{(k)} &= 1, \\ q_i^{(k)} &= -l_i^+ q_{i+1}^{(k)}, \quad i = k-1, \ldots, 1, \\ q_{i+1}^{(k)} &= -u_i^+ q_i^{(k)}, \quad i = k, \ldots, n-1. \end{aligned} \tag{2.9}$$

Formally, applying (2.9) is equivalent to solving the system

$$N_k G_k (N_k)^T q^{(k)} = \gamma_k e_k, \quad q_k^{(k)} = 1, \tag{2.10}$$

where $N_k G_k N_k^T = LDL^T - \bar{\lambda} I$, $G_k = \mathrm{diag}(d_1, \ldots, d_{k-1}, \gamma_k, r_{k+1}, \ldots, r_n)$, and N_k is a tridiagonal matrix with

$$(N_k)_{1:k,1:k} = L^+_{1:k,1:k} \text{ and } (N_k)_{k:n,k:n} = U^+_{k:n,k:n}.$$

The matrix $N_k G_k N_k^T$ is called a *twisted factorization* of $LDL^T - \bar{\lambda} I$ (also called BABE-factorization for "burn at both ends"), and k the *twist index*, because it can be obtained by applying the Gaussian elimination process from the top to row k and then backward from the bottom to row k.

The remaining question now is, which k is best? As, according to (2.10), the residual for $q^{(k)}$ is $\gamma_k e_k$, any twist index k minimizing $|\gamma_k|$ is an obvious candidate. Comparing $(N_k G_k N_k^T)_{k,k}$ with $(LDL^T - \bar{\lambda} I)_{k,k}$ gives

$$\gamma_k = d_k^+ + r_k^+ - \left((LDL^T)_{k,k} - \bar{\lambda} \right), \quad k = 1 : n.$$

A more stable way to compute γ_k is

$$\gamma_k = s_k + p_k + \tau,$$

using the intermediate quantities s_k^+ and p_k^+ of Algorithm 2.1 for the factorizations (2.8).

It is shown in [6] that if k is chosen such that $|\gamma_k|$ is minimized (or small enough), the resulting vector $q^{(k)}$ will indeed fulfill (2.7) and therefore also (2.5).

Another way to understand the above method is as a variant of inverse iteration. A special and interesting property of the twist indices is that if $\bar{\lambda}$ approximates λ to high relative accuracy, i.e., if (2.6) is fulfilled, the index k minimizing $|\gamma_k|$ will correspond to the component of the eigenvector with largest absolute value. (Actually, this is true only in the limit case $\bar{\lambda} \to \lambda$; see [6] for details). Therefore, e_k is guaranteed to be an excellent choice as a starting vector. The simplicity of this right-hand side then allows us to compute the first iterate using only multiplications in (2.9). Recent developments of this technique by Parlett and Vömel in [21] even allow more steps of inverse iteration with twisted factorizations, again using only multiplications to avoid spoiling the relative accuracy of the vectors.

3. The black-box approach fails. As already mentioned in section 1, there are mainly two different ways of reducing the bSVD to the tSEP. The first approach employs the *normal equations* and computes eigendecompositions

$$B^T B = V \Sigma^2 V^T \text{ and } B B^T = U \Sigma^2 U^T,$$

which together give us the desired bSVD $B = U \Sigma V^T$ of B.

Alternatively, one can use the so-called *Golub–Kahan* matrix T_{GK} of B, which is defined as

$$T_{GK} := P_{ps} \cdot \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix} \cdot P_{ps}^T,$$

where P_{ps} is a "perfect shuffle" permutation mapping a vector $x \in \mathbb{R}^{2n}$ to $P_{ps} x = [x_{n+1}, x_1, x_{n+2}, x_2, \ldots, x_{2n}, x_n]^T$. It is easy to see that T_{GK} is symmetric tridiagonal with a zero diagonal and the entries of B interleaved on the off-diagonals, i.e.,

$$T_{GK} = \mathrm{diag}\left([a_1, b_1, a_2, \ldots, b_{n-1}, a_n], \pm 1 \right).$$

Given an eigendecomposition $T_{GK} = Q\Lambda Q^T$, in exact arithmetic Q will have the structure

$$Q = \frac{1}{\sqrt{2}} \cdot P_{ps} \cdot \begin{bmatrix} U & U \\ -V & V \end{bmatrix},$$

so that the singular vectors of B can be easily extracted from the odd- and even-numbered rows of Q. Furthermore, the eigenvalues of T_{GK} are simply the singular values of B and their negations. This has the additional benefit that if an adaptable method like the MRRR algorithm is used for the computation of Q and Λ, only n eigenpairs are needed, reducing the computation time by half.

It is tempting to use the MRRR algorithm as a "black box" on either the normal equations or the Golub–Kahan matrix to obtain the SVD of our bidiagonal matrix B. However, this leads to major numerical instabilities. In the following we will give only a short description of these problems as a motivation for the remainder of the paper. A more detailed analysis can be found in [12, 11].

Using the MRRR algorithm to compute eigendecompositions of B^TB and BB^T separately can result in bad residuals for the singular vectors, in the sense of (1.3). One reason is that MRRR uses different shifts when working on B^TB and BB^T, although in exact arithmetic the spectra of these matrices are identical. But the real source of failure turns out to be more subtle, as it lies in computing the tridiagonal factorizations.

To be more concrete, let us assume that on the first level of the representation trees of B^TB and BB^T, a cluster $\sigma^2_{c:d}$ is encountered. Let us further assume that, in order to proceed to the next level and break up the cluster, the same shift $\mu \approx \sigma_c^2$ is chosen close to the cluster, and that the factorizations

$$\begin{aligned} B^TB - \mu I &= \hat{L}\hat{D}\hat{L}^T, \\ BB^T - \mu I &= \check{L}\check{D}\check{L}^T \end{aligned}$$

are computed with the dstqds transformation from Algorithm 2.1, such that $\hat{L}\hat{D}\hat{L}^T$ and $\check{L}\check{D}\check{L}^T$ form RRRs for their respective eigenpairs $c : d$.

Now, in exact arithmetic, for the local eigenvalues $\hat{\lambda}_i$ of $\hat{L}\hat{D}\hat{L}^T$ and $\check{\lambda}_i$ of $\check{L}\check{D}\check{L}^T$, we would have $\hat{\lambda}_i = \sigma_i^2 - \mu = \check{\lambda}_i$. We already mentioned that the dstqds transformation allows a mixed relative perturbation analysis, that is, an exact relation holds only for small relative perturbations of the inputs and outputs. As all matrices in our small example are RRRs, this implies that the *actual* relationship between $\hat{\lambda}_i$ and $\check{\lambda}_i$ is of the form

$$\begin{aligned} \hat{\lambda}_i(1 + K_1\epsilon) &= \sigma_i^2(1 + K_2\epsilon) - \mu, \\ \check{\lambda}_i(1 + K_3\epsilon) &= \sigma_i^2(1 + K_4\epsilon) - \mu, \end{aligned}$$

where the K_i are small constants. Because $\sigma_c^2 \approx \mu$, and because $\sigma^2_{c:d}$ form a cluster of close eigenvalues, we will in general have $\sigma_i^2 \gg \hat{\lambda}_i$, $\sigma_i^2 \gg \check{\lambda}_i$ for $i = c : d$. Based on this it is easy to show that the *absolute deviation* between the local eigenvalues $\hat{\lambda}_i$ and $\check{\lambda}_i$, $i = c : d$, can be as large as

$$|\hat{\lambda}_i - \check{\lambda}_i| = \mathcal{O}(\sigma_i^2\epsilon). \tag{3.1}$$

In [11] this fact has been demonstrated by a numerical example, where $\hat{\lambda}_i$ and $\check{\lambda}_i$ disagree on nearly all of their digits.

Essentially, independently factorizing $B^TB-\mu I$ and $BB^T-\mu I$ using dstqds causes a slight difference in the implied initial relative perturbations of the σ_i^2 ($K_2 \neq K_4$ in the example above). Then the cancellation which is implicit (and desired) in shifting close to a cluster magnifies this small deviation up to (3.1). The effect can be tolerable for clusters of small singular values σ_i, but for tight clusters of large singular values it can possibly destroy the relationship between the computed vectors, resulting in bad residuals for the bSVD. However, because we apply the MRRR algorithm separately on BB^T and B^TB, the orthogonality requirements (1.2) for the computed matrices U and V will be fulfilled.

For the Golub–Kahan matrix the situation is reversed. The residuals (1.3) of the singular vectors are always excellent, but the deviation from orthogonality (1.2) can become bad for some bidiagonals with tight clusters of very small singular values.

The main reason lies in the fact that every minor of T_{GK} with an odd dimension is singular. This can make it more difficult for the algorithm to find RRRs for the next level. Additionally, we often observe quite differing element growth in the even-numbered and odd-numbered elements of the representations, which can intuitively cause a "displacement of information." This is problematic, as only one twist index is used to compute two singular vectors (via Q).

At this point we want to give an example of the impact of these problems on the orthogonality level of the vectors. As a test matrix we took the upper bidiagonal

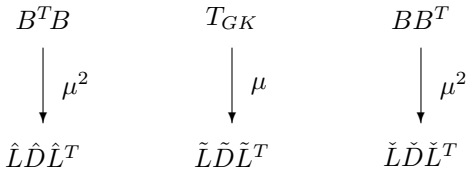
$$B := \operatorname{diag}([1, \alpha, \ldots, \alpha], 0) + \operatorname{diag}([\alpha, \ldots, \alpha], +1), \quad \alpha = 200\epsilon. \tag{3.2}$$

This matrix has one singular value around 1 and the rest clustered at 10^{-14}. We tested for different dimensions of B the newest tridiagonal MRRR implementation `DSTEGR` on the resulting Golub–Kahan matrix T_{GK}. The following table shows in the second column the measured orthogonality $||Q^TQ-I||/(n\epsilon)$ for the tSEP belonging to T_{GK}. The third column shows the orthogonality $\max(||U^TU - I||, ||V^TV - I||)/(n\epsilon)$ of the extracted singular vectors from Q, and in the fourth column the same measure is shown for the results obtained by our software `DBDSCR`.

n	`DSTEGR` on T_{GK}		`DBDSCR` on B
	tSEP	bSVD	bSVD
100	4.45	$> 10^{10}$	3.71
200	5.10	$> 10^{10}$	4.13
400	2.93	$> 10^{10}$	3.80

Note that Q itself fulfills the requirements (2.2) for the tSEP defined nicely by T_{GK}, but the singular vectors extracted from Q are effectively useless. As can be seen in the last column, the coupling techniques which we will present in the next section avoid this problem.

A short note on notation. In the remaining parts of this paper we will have to deal constantly with the three matrices B^TB, BB^T, and T_{GK} and the various bidiagonal factorizations occurring in their respective representation trees for the MRRR algorithm. In order to distinguish between these matrices, we will continue to use superscripts $\wedge, \vee$, and $\sim$, as introduced in this section and presented in the following diagram:



4. The bidiagonal MRRR algorithm. As we have seen in the preceding section, the main problem when trying to solve the bSVD by the application of MRRR to B^TB and BB^T lies in the separate factorizations, which can cause the local eigenvalues to drift apart.

In [12], Großer and Lang proposed a solution to this problem. They devised so-called *coupling relations*, which link the factorizations

$$B^TB - \mu^2 I =: \hat{L}\hat{D}\hat{L}^T, \quad T_{GK} - \mu I =: \tilde{L}\tilde{D}\tilde{L}^T, \text{ and } BB^T - \mu^2 I =: \check{L}\check{D}\check{L}^T$$

in a backward stable way. As a consequence, it is necessary to do only one of the factorizations explicitly; the remaining two factorizations can then be computed implicitly using only multiplications and divisions. This guarantees that the eigenvalues of $\hat{L}\hat{D}\hat{L}^T$ and $\check{L}\check{D}\check{L}^T$ will agree to most of their digits.

This then suggests a new algorithm for the bSVD based on the MRRR algorithm. The idea is to apply MRRR simultaneously on B^TB, T_{GK}, and BB^T (which we sometimes call "the three matrices") with identical shifts μ^2 for B^TB and BB^T and μ for T_{GK}. But the dstqds factorizations needed to proceed from one level to the next are always done for only one of the matrices—in most cases this will be T_{GK}; see below—whereas the above mentioned coupling relations are used to keep track of the other two factorizations *implicitly*. The backward stable nature of the coupling relations will ensure that the eigenvalues of a representation $\hat{L}\hat{D}\hat{L}^T$ in the tree for B^TB and the corresponding representation $\check{L}\check{D}\check{L}^T$ for BB^T always remain relatively close. As a consequence, if upon encountering a singleton the singular vectors are computed using the coupled representations of B^TB and BB^T, we get vectors with small residuals *and* good orthogonality levels.

This section is divided into two parts. First we will present the coupling relations, and then we describe our approach to incorporate them in an efficient and practical way into an MRRR algorithm for the bSVD.

4.1. The coupling relations. At the core of the new algorithm lies the capability to convert between shifted factorizations of the matrices B^TB, BB^T, and T_{GK} in a backward stable way. In the following, we will summarize the main results from [12, 11] needed to understand and implement the algorithm.

We will not give detailed proofs of the coupling relations in this paper, as they can be quite technical (and the following pages are already technical enough). Their main ingredients are that shifted factorizations of the three matrices can be related by

$$\begin{pmatrix} BB^T - \mu^2 I & 0 \\ 0 & B^TB - \mu^2 I \end{pmatrix} = P_{ps}^T (T_{GK} + \mu I)(T_{GK} - \mu I) P_{ps} \tag{4.1}$$

and that $T_{GK} - \mu I = \tilde{L}\tilde{D}\tilde{L}^T$ implies $T_{GK} + \mu I = \bar{L}\bar{D}\bar{L}^T$ with $\bar{D} = -\tilde{D}$ and $\bar{l}_i = -\tilde{l}_i, i = 1 : 2n-1$. A deeper analysis of this simple relation leads to the following result.

LEMMA 4.1. *Let the decompositions*

$$T_{GK} - \mu I = \tilde{L}\tilde{D}\tilde{L}^T = \tilde{U}\tilde{R}\tilde{U}^T = \tilde{N}_r\tilde{G}_r\tilde{N}_r^T, \;\; r = 1:2n,$$

exist and be RRRs. Then the decompositions

$$\begin{aligned} B^TB - \mu^2 I &= \hat{L}\hat{D}\hat{L}^T = \hat{U}\hat{R}\hat{U}^T = \hat{N}_k\hat{G}_k\hat{N}_k^T, \;\; k = 1:n, \\ BB^T - \mu^2 I &= \check{L}\check{D}\check{L}^T = \check{U}\check{R}\check{U}^T = \check{N}_k\check{G}_k\check{N}_k^T, \;\; k = 1:n, \end{aligned}$$

also form RRRs, and for $i = 1:n$ the diagonal pivots and twist elements are given by

$$\begin{aligned} \hat{d}_i &= -\tilde{d}_{2i-1}\tilde{d}_{2i}, & \check{d}_i &= -\tilde{d}_{2i}\tilde{d}_{2i+1}, \\ \hat{r}_i &= -\tilde{r}_{2i-2}\tilde{r}_{2i-1}, & \check{r}_i &= -\tilde{r}_{2i-1}\tilde{d}_{2i}, \\ \hat{\gamma}_i &= \mu\tilde{\gamma}_{2i-1}, & \check{\gamma}_i &= \mu\tilde{\gamma}_{2i}, \end{aligned} \tag{4.2}$$

where we set $\tilde{d}_{2n+1} := \tilde{d}_1$ and $\tilde{r}_0 := \tilde{r}_{2n}$. The elements $\hat{l}_i, \check{l}_i, \hat{u}_i, \check{u}_i$ can then be determined using

$$\hat{l}_i\hat{d}_i = \hat{u}_i\hat{r}_{i+1} = a_ib_i \quad \textit{and} \quad \check{l}_i\check{d}_i = \check{u}_i\check{r}_{i+1} = a_{i+1}b_i. \tag{4.3}$$

for $i = 1:n-1$.

Proof. The couplings (4.2) follow from (4.1), although additional technical argumentation is needed, which is beyond the scope of this paper; see Lemma 3.1 and Corollary 3.2 in [12]. The identities (4.3) result from the fact that the off-diagonal elements a_ib_i of B^TB and $a_{i+1}b_i$ of BB^T are not affected by the shift. ∎

The requirement that each of the $2n$ twisted factorizations of $T_{GK} - \mu I$ has to be an RRR is redundant, as it is shown in [6] that if a twisted factorization is an RRR for some twist index k, then this also holds true for all twist indices.

Concerning the local eigenvalues, it was proved in [11] that if $\hat{L}\hat{D}\hat{L}^T$ and $\check{L}\check{D}\check{L}^T$ are set up from $\tilde{L}\tilde{D}\tilde{L}^T$ using (4.2) and (4.3), the relative distance of the respective eigenvalues $\hat{\lambda}_i$ and $\check{\lambda}_i$ obeys

$$\operatorname{reldist}(\hat{\lambda}_i, \check{\lambda}_i) = \mathcal{O}(\epsilon). \tag{4.4}$$

Thus the local eigenvalues $\hat{\lambda}_i$ and $\check{\lambda}_i$ will agree to most of their digits.

A special point in the above coupling relations is that they are completely oblivious to the way in which the factorization of $T_{GK} - \mu I$ is computed. Therefore they are also valid in the case of successive factorizations, which occur naturally during the MRRR algorithm. As an example, let us assume we apply MRRR to T_{GK}. Let us omit the index ranges of the eigenvalues for now and denote by $(\tilde{L}^{(i)}, \tilde{D}^{(i)}, \mu^{(i)}), i = 1, 2, \ldots,$ a path in the representation tree, i.e.,

$$\begin{aligned} &T_{GK} - \mu^{(1)}I =: \tilde{L}^{(1)}\tilde{D}^{(1)}(\tilde{L}^{(1)})^T \text{ and} \\ &\tilde{L}^{(i)}\tilde{D}^{(i)}(\tilde{L}^{(i)})^T - \mu^{(i+1)}I =: \tilde{L}^{(i+1)}\tilde{D}^{(i+1)}(\tilde{L}^{(i+1)})^T, \;\; i = 1, 2, \ldots \; . \end{aligned}$$

Then we can use Lemma 4.1 to set up the corresponding paths $(\hat{L}^{(i)}, \hat{D}^{(i)}, \nu^{(i)})$ for B^TB and $(\check{L}^{(i)}, \check{D}^{(i)}, \nu^{(i)})$ for BB^T, where the shifts $\nu^{(i)}$ are related to the T_{GK} shifts $\mu^{(i)}$ by

$$\sum_{j=1}^{i} \nu^{(j)} = \left(\sum_{j=1}^{i} \mu^{(j)} \right)^2 .$$

Evaluating this recurrence gives

$$\nu^{(i)} = \mu^{(i)}(2\bar{\mu}^{(i-1)} + \mu^{(i)}) \quad \text{with} \quad \bar{\mu}^{(i-1)} := \sum_{j=1}^{i-1} \mu^{(j)}. \tag{4.5}$$

Together with (4.4) this means that we are able to run MRRR implicitly on B^TB and BB^T in parallel with identical shifts, thereby guaranteeing that the local eigenvalues of the corresponding representations $\hat{L}^{(i)}\hat{D}^{(i)}(\hat{L}^{(i)})^T$ and $\check{L}^{(i)}\check{D}^{(i)}(\check{L}^{(i)})^T$ are always relatively close. This is already one big step toward the solution compared to the separate application of MRRR on B^TB and BB^T.

However, we are still doing much work with the Golub–Kahan matrix and its translates. As was already hinted at in section 3, there are two major problems when working with T_{GK}. First, it is hard to find good shifts μ such that we can prove $T_{GK} - \mu I$ to be an RRR. Second, element growth in the factorizations can lead to problems when computing the vectors. The latter problem is now resolved, as we can use the couplings from Lemma 4.1 to compute the vectors directly with translates of B^TB and BB^T, which results in good orthogonality *and* small residuals because of (4.4). Concerning the first issue, the following lemma summarizes another coupling relation presented in [12], which utilizes the intermediate quantities arising during the differential qds transformations in Algorithm 2.1 in order to avoid factorizing $T_{GK} - \mu^{(1)}I$ and to use $B^TB - \nu^{(1)}$ instead.

LEMMA 4.2. *Let the factorizations*

$$B^TB - \mu^2 I = \hat{L}\hat{D}\hat{L}^T = \hat{U}\hat{R}\hat{U}^T = \hat{N}_k\hat{G}_k\hat{N}_k^T, \ \ k = 1:n,$$

be computed using Algorithm 2.1 *with intermediate quantities* $\hat{S}$ *and* $\hat{P}$. *Then the decompositions*

$$\begin{aligned} T_{GK} - \mu I \ \ &= \tilde{L}\tilde{D}\tilde{L}^T = \tilde{U}\tilde{R}\tilde{U}^T = \tilde{N}_k\tilde{G}_k\tilde{N}_k^T, \quad k = 1:2n, \\ BB^T - \mu^2 I &= \check{L}\check{D}\check{L}^T = \check{U}\check{R}\check{U}^T = \check{N}_k\check{G}_k\check{N}_k^T, \quad k = 1:n, \end{aligned}$$

are given by

$$\check{d}_i = \frac{\hat{s}_{i+1}}{\hat{s}_i}\hat{d}_i, \quad \check{r}_i = \frac{\hat{p}_i}{\hat{p}_{i+1}}\hat{r}_{i+1}, \quad \check{\gamma}_i = -\mu^2\hat{\gamma}_i\frac{\hat{r}_{i+1}}{\hat{s}_i\hat{p}_{i+1}}$$

and

$$\begin{aligned} \tilde{d}_{2i-1} &= \frac{\hat{s}_i}{\mu}, & \tilde{r}_{2i-1} &= \frac{\hat{p}_i}{\mu}, & \tilde{\gamma}_{2i-1} &= \frac{\hat{\gamma}_i}{\mu}, \\ \tilde{d}_{2i} &= -\frac{\hat{d}_i}{\tilde{d}_{2i-1}}, & \tilde{r}_{2i} &= -\frac{\check{r}_i}{\tilde{r}_{2i-1}}, & \tilde{\gamma}_{2i} &= -\frac{\check{\gamma}_i}{\mu} \end{aligned}$$

for $i = 1:n$, *setting* $\hat{s}_{n+1} := -\mu^2, \hat{r}_{n+1} := \hat{p}_{n+1} := 1$.

Proof. See Lemma 2.3 and Corollaries 2.4 and 2.5 in [12]. ∎

The elements $\check{l}_i, \check{u}_i, \tilde{l}_i, \tilde{u}_i$ can be obtained as in (4.3) using $\check{l}_i\check{d}_i = \check{u}_i\check{r}_{i+1} = a_{i+1}b_i$, $\tilde{l}_{2i-1}\tilde{d}_{2i-1} = \tilde{u}_{2i-1}\tilde{r}_{2i} = a_i$ and $\tilde{l}_{2i}\tilde{d}_{2i} = \tilde{u}_{2i}\tilde{r}_{2i+1} = b_i$. Again it holds that if $(\hat{L}, \hat{D})$ is an RRR, then $(\check{L}, \check{D})$ is too and (4.4) is fulfilled; i.e., the eigenvalues of $\hat{L}\hat{D}\hat{L}^T$ and $\check{L}\check{D}\check{L}^T$ are relatively close [11, Theorem 5.4].

4.2. Modified MRRR algorithm with couplings. In this section we will present the structure of our new implementation of the adapted MRRR algorithm for the bSVD with embedded couplings.

First we want to outline the main difference between our approach and the algorithm presented in [12]. There, another coupling transformation was used on deeper levels to couple directly from $(\hat{L}^+, \hat{D}^+)$ to $(\check{L}^+, \check{D}^+)$, similarly to Lemma 4.2 for the first level [12, p. 15]. Unfortunately, this transformation is based on an implicit partial factorization of $\tilde{L}\tilde{D}\tilde{L}^T$. Therefore it cannot guarantee (4.4) and the only way to use this transformation is to compute the eigenvalues of $\hat{L}^+\hat{D}^+(\hat{L}^+)^T$ and $\check{L}^+\check{D}^+(\check{L}^+)^T$ to full accuracy and to compare them [12, p. 18]. The algorithm in [12] was based on the original presentation of the tridiagonal MRRR algorithm in [6], where the eigenvalues were computed to full accuracy on each level of the tree anyway. In this context, the quality of the couplings could be checked easily.

However, as we pointed out in section 2 based on [9], it is sufficient for the MRRR algorithm to refine the eigenvalues on each level only until they can be categorized into singletons and clusters. For example, with the cluster tolerance set to 10^{-3}, this implies essentially that merely the first three decimal digits of the eigenvalues have to be computed on each level. As the computation of the eigenvalues is by far the most expensive part during the MRRR algorithm, this optimization results in a significant speedup of the implementation.

If we wanted to employ the direct coupling strategy from [12], we would have essentially two options. We could ignore the above optimization and still refine the eigenvalues to full precision on each level, which would pose a serious and unnecessary runtime overhead. Alternatively, we could use the optimization but skip checking the quality of the couplings, resulting in a loss of robustness of the method. In our opinion, both options are unacceptable.

Therefore we propose a different strategy for deeper levels. We do the steps from one level to the next with the local translate $\tilde{L}\tilde{D}\tilde{L}^T$ of the Golub–Kahan matrix and use Lemma 4.1 to set up the respective representations in the trees of B^TB and BB^T. As this coupling guarantees (4.4), we do not need to refine the eigenvalues to full accuracy anymore.

As outlined in the previous section, the new algorithm can be described as implicitly running MRRR on the matrices B^TB, T_{GK}, and BB^T simultaneously with equivalent shifts. As a consequence, we are working in some sense on a (synchronized) *three-layered representation tree* with nodes $[\hat{L}\hat{D}\hat{L}^T, \tilde{L}\tilde{D}\tilde{L}^T, \check{L}\check{D}\check{L}^T, \bar{\mu}, jl : ju]$. The local index range $jl : ju$ denotes the subset of desired singular values of B (respectively, eigenvalues of B^TB) and $\bar{\mu}$ is the accumulated shift from T_{GK}, i.e., we have

$$\begin{aligned} B^TB - \bar{\mu}^2 I &= \hat{L}\hat{D}\hat{L}^T, \\ T_{GK} - \bar{\mu} I &= \tilde{L}\tilde{D}\tilde{L}^T, \\ BB^T - \bar{\mu}^2 I &= \check{L}\check{D}\check{L}^T. \end{aligned}$$

Note that in order to be correct, the corresponding index range for the eigenvalues of T_{GK} would be $n + jl : n + ju$. For the sake of shorter indices we will omit this detail; that is, we refer to the $(n + i)$th eigenvalue of T_{GK} and its translates $\tilde{L}\tilde{D}\tilde{L}^T$ as $\tilde{\lambda}_i$, due to the fact that we are interested only in the n positive eigenvalues of T_{GK}.

Recall that in the tridiagonal MRRR algorithm, the main tasks to be done for each node are as follows:

1. Refine the local eigenvalues of the shifted matrix in order to identify clusters and singletons.
2. For singletons, compute the eigenvector using twisted factorizations.
3. For clusters, find a shift close to the cluster resulting in a new partial RRR for the eigenvalues in the cluster. Compute and store the new representation for the

work on the next level.
These steps remain essentially the same in our approach. But as we use Lemma 4.2 to handle the couplings for the root node and Lemma 4.1 for deeper levels, the above steps are applied to different matrices, depending on the level in the trees we are currently on. Therefore we will treat the root node (i.e., level zero) and the deeper levels separately in the following detailed description of our algorithm.

The root node (*level* 0). Given the upper bidiagonal B and a range $il : iu$ for the singular values of interest, in theory the root of our three-layered tree becomes

$$[B^TB, T_{GK}, BB^T, 0, il : iu],$$

but by using Lemma 4.2 we will only need to work with B^TB. A structural overview of the steps performed for the root node is given in Algorithm 4.1.

The computation for the root node starts with refining the eigenvalues $\hat{\lambda}_i = \sigma_i^2, i = il : iu$. For each singleton $\hat{\lambda}_s$, we take advantage of the fact that Lemma 4.2 allows the direct coupling of twisted factorizations of $B^TB - \nu I$ and $BB^T - \nu I$. That is, after refining $\hat{\lambda}_s$ to high relative accuracy, we use Algorithm 2.1 to compute the twisted factorizations $B^TB - \hat{\lambda}_s I = \hat{N}_k \hat{G}_k \hat{N}_k^T, k = 1 : n$. (Note that, as B is upper bidiagonal, B^TB can be written as LDL^T with $d_i = a_i^2, l_i = b_i/a_i$. Thus Algorithm 2.1 can be applied directly.) Then the right singular vector v_s is computed using (2.9) for a suitable twist index $\hat{k}$. For the left singular vector, we invoke Lemma 4.2 to set up the twisted factorizations $BB^T - \hat{\lambda}_s I = \check{N}_k \check{G}_k \check{N}_k^T, k = 1 : n$ directly, and we use them to compute u_s. Note that it is possible to choose a different twist index $\check{k}$ for the left vector. As the couplings are backward stable this results in excellent residuals $||Bv_s - \sigma_s u_s||$.

Now assume that a cluster $\hat{\lambda}_{c:d}$ of eigenvalues of B^TB has been identified. In the same manner as MRRR applied to B^TB alone would proceed, we choose a shift μ^2 s.t. $B^TB - \mu^2 =: \hat{L}\hat{D}\hat{L}^T$ forms a partial RRR for its eigenvalues with indices $c : d$. Then we can again use Lemma 4.2 to set up the remaining data for the child node $[\hat{L}\hat{D}\hat{L}^T, \tilde{L}\tilde{D}\tilde{L}^T, \check{L}\check{D}\check{L}^T, \mu, c : d]$.

Deeper levels. On deeper levels of the tree we work on nodes of the form

$$\left[\hat{L}\hat{D}\hat{L}^T, \tilde{L}\tilde{D}\tilde{L}^T, \check{L}\check{D}\check{L}^T, \bar{\mu}, jl : ju\right],$$

where $jl : ju$ is a subset of the root index range $il : iu$ with $jl < ju$. The computational structure for deeper levels is shown in Algorithm 4.2.

Again we start by refining the eigenvalues $\hat{\lambda}_i$, $i = jl : ju$, of $\hat{L}\hat{D}\hat{L}^T$. For singletons $\hat{\lambda}_s$, we want to compute the singular vectors u_s and v_s as for the root node using twisted factorizations

$$\hat{L}\hat{D}\hat{L}^T - \hat{\lambda}_s I = \hat{N}_k^+ \hat{G}_k^+ (\hat{N}_k^T)^+, \quad k = 1 : n, \tag{4.6}$$

$$\check{L}\check{D}\check{L}^T - \hat{\lambda}_s I = \check{N}_k^+ \check{G}_k^+ (\check{N}_k^T)^+, \quad k = 1 : n. \tag{4.7}$$

Unfortunately, compared to the root node case we have lost the advantage of being able to couple directly from $(\hat{N}_k^+, \hat{G}_k^+)$ to $(\check{N}_k^+, \check{G}_k^+)$. This leaves essentially two options. First we could compute the twisted factorizations $\tilde{L}\tilde{D}\tilde{L}^T - \tilde{\lambda}_s I = \tilde{N}_k^+ \tilde{G}_k^+ (\tilde{N}_k^T)^+, k = 1 : 2n$, and then use Lemma 4.1 to set up the data $\hat{N}_k^+, \hat{G}_k^+, \check{N}_k^+, \check{G}_k^+$, $k = 1 : n$, to compute the vectors. As this coupling obeys (4.4), it does lead to excellent results.

However, there is a drawback to this approach. In practice, the computation of eigenpairs during the MRRR algorithm can be accelerated using a specialized Rayleigh

Algorithm 4.1 Bidiagonal MRRR, root node.

Input: Upper bidiagonal matrix B, range $il : iu$ of desired singular triplets.

1: Refine eigenvalues $\hat{\lambda}_{il:iu}$ of B^TB enough to identify clusters.
2: **for** each singleton $\hat{\lambda}_s$ **do**
3: compute $\hat{\lambda}_s$ to full accuracy
4: compute $B^TB - \hat{\lambda}_s I = \hat{N}_k \hat{G}_k \hat{N}_k^T, k = 1 : n$
5: compute v_s for a suitable twist index $\hat{k}$
6: couple $(\hat{N}_k \hat{G}_k \hat{N}_k^T, k = 1 : n) \to (\check{N}_k \check{G}_k \check{N}_k^T, k = 1 : n)$ *(Lemma 4.2)*
7: compute u_s for a suitable twist index $\check{k}$
8: **endfor**
9: **for** each cluster $\hat{\lambda}_{c:d}$ **do**
10: find close shift μ^2 s.t. $B^TB - \mu^2 I = \hat{L}\hat{D}\hat{L}^T$ and
$(\hat{L}, \hat{D})$ forms a partial RRR for the eigenvalues $c : d$
11: modify eigenvalues for the next level: $\hat{\lambda}_i := \hat{\lambda}_i - \mu^2, i = c : d$
12: store $\hat{L}, \hat{D}, \hat{S}$, μ, and $c : d$
13: **endfor**

quotient iteration (RQI) for twisted factorizations, as described in [6, 10]. Doing this for $\tilde{L}\tilde{D}\tilde{L}^T$ is undesirable, as this matrix is of dimension $2n$, and therefore the loops in the RQI take twice as many operations as for the matrices $\hat{L}\hat{D}\hat{L}^T$ and $\check{L}\check{D}\check{L}^T$. For this reason it was proposed in [10] to forfeit the couplings at this point, i.e., to do RQI on $\hat{L}\hat{D}\hat{L}^T$ for v_s and $\hat{\lambda}_s$, and then to use the resulting approximation $\hat{\lambda}_s$ to do the factorization of $\check{L}\check{D}\check{L}^T - \hat{\lambda}_s I$ explicitly to compute u_s. This does not spoil the residuals, because, as we are on a deeper level of the tree, the local eigenvalues of $\hat{L}\hat{D}\hat{L}^T$ and $\check{L}\check{D}\check{L}^T$ are typically very small compared to the singular values of B. Therefore, the resulting absolute deviation (3.1) does not cause much harm at this point.

After dealing with the singletons we still have to handle possibly upcoming new (sub)clusters $\hat{\lambda}_{c:d}$ on deeper levels, where $jl \le c < d \le ju$. To do the step to the next level, we use the translate $\tilde{L}\tilde{D}\tilde{L}^T$ of the Golub–Kahan matrix with a suitable shift μ to compute a new partial RRR $\tilde{L}\tilde{D}\tilde{L}^T - \mu I = \tilde{L}^+\tilde{D}^+(\tilde{L}^+)^T$. Then we apply Lemma 4.1 to set up the representations

$$\hat{L}\hat{D}\hat{L}^T - \nu I = \hat{L}^+\hat{D}^+(\hat{L}^+)^T \quad \text{and} \quad \check{L}\check{D}\check{L}^T - \nu I = \check{L}^+\check{D}^+(\check{L}^+)^T$$

for the child node

$$\left[\hat{L}^+\hat{D}^+(\hat{L}^+)^T, \tilde{L}^+\tilde{D}^+(\tilde{L}^+)^T, \check{L}^+\check{D}^+(\check{L}^+)^T, \bar{\mu} + \mu, c : d\right].$$

Its local eigenvalues are related via $\tilde{\lambda}_i = \sigma_i - \bar{\mu}$ and

$$\hat{\lambda}_i = \sigma_i^2 - \bar{\mu}^2 = (\sigma_i - \bar{\mu})(\sigma_i + \bar{\mu}) = \tilde{\lambda}_i(2\bar{\mu} + \tilde{\lambda}_i).$$

Remember that $\tilde{\lambda}_i$ refers actually to the $(n+i)$th eigenvalue of $\tilde{L}\tilde{D}\tilde{L}^T$. Together with (4.5) we can therefore express the relations between the local eigenvalues $\hat{\lambda}_i$ and $\tilde{\lambda}_i$, and between the local shifts μ and ν as

$$\hat{\lambda}_s = \mathrm{conv}(\tilde{\lambda}_s, \bar{\mu}), \quad \nu = \mathrm{conv}(\mu, \bar{\mu}), \quad \text{where} \quad \mathrm{conv}(x, y) := x(2y + x). \tag{4.8}$$

Algorithm 4.2 Bidiagonal MRRR, deeper levels ($level \geq 1$).

if $level = 1$ **then**
  Retrieve $\hat{L}, \hat{D}, \hat{S}$, GK-shift from root $\bar{\mu}$ and cluster bounds $jl : ju$
else
  Retrieve $\tilde{L}, \tilde{D}$, GK-shift from root $\bar{\mu}$ and cluster bounds $jl : ju$
  couple $(\tilde{L}, \tilde{D}) \rightarrow (\hat{L}, \hat{D})$ *(Lemma 4.1)*
endif
Refine eigenvalues $jl : ju$ of $\hat{L}\hat{D}\hat{L}^T$ enough to identify clusters
if singletons found **then**
  if $level = 1$ **then**
    couple $(\hat{L}, \hat{D}, \hat{S}) \rightarrow (\check{L}, \check{D})$ *(Lemma 4.2)*
  else
    couple $(\tilde{L}, \tilde{D}) \rightarrow (\check{L}, \check{D})$ *(Lemma 4.1)*
  endif
  for each singleton $\hat{\lambda}_s$ **do**
    compute $\hat{\lambda}_s$ to full accuracy
    compute $\hat{L}\hat{D}\hat{L}^T - \hat{\lambda}_s I = \hat{N}_k \hat{G}_k \hat{N}_k^T$, $k = 1 : n$, to get v_s
    compute $\check{L}\check{D}\check{L}^T - \hat{\lambda}_s I = \check{N}_k \check{G}_k \check{N}_k^T$, $k = 1 : n$, to get u_s
  endfor
endif
if new clusters found **then**
  if $level = 1$ **then**
    couple $(\hat{L}, \hat{D}, \hat{S}) \rightarrow (\tilde{L}, \tilde{D})$ *(Lemma 4.2)*
  endif
  for each cluster $\hat{\lambda}_{c:d}$ **do**
    approximate eigenvalues $\tilde{\lambda}_{c:d}$ based on $\hat{\lambda}_{c:d}$
    find close shift μ s.t. $\tilde{L}\tilde{D}\tilde{L}^T - \mu I =: \tilde{L}^+\tilde{D}^+(\tilde{L}^+)^T$ is an RRR
    transform shift: $\nu := \mu(2\bar{\mu} + \mu)$
    modify eigenvalues for the next level: $\hat{\lambda}_i := \hat{\lambda}_i - \nu, i = c : d$
    store $\tilde{L}, \tilde{D}$, $\bar{\mu} + \mu$ and $c : d$
  endfor
endif

This relation is needed for two reasons. First, after choosing μ, we need ν to get initial guesses $\hat{\lambda}_i - \nu$ for the local eigenvalues of $\hat{L}^+\hat{D}^+(\hat{L}^+)^T$. Additionally, in order to choose μ sensibly, some approximation to the eigenvalues $\tilde{\lambda}_{c:d}$ of $\tilde{L}\tilde{D}\tilde{L}^T$ is needed. We want to avoid performing any direct eigenvalue computations for $\tilde{L}\tilde{D}\tilde{L}^T$, as this would be expensive, so we approximate $\tilde{\lambda}_i$ from $\hat{\lambda}_i$ instead. The relation (4.8) implies that $\tilde{\lambda}_i$ is defined as the larger root of the quadratic equation $x^2 + 2\bar{\mu}x - \hat{\lambda}_i$. Care has to be taken to compute this root in a stable way; see [13], for example.

5. The software `xBDSCR`. The bidiagonal MRRR algorithm as described in the last section has been realized as software `xBDSCR` in FORTRAN 77 and is to be incorporated into the upcoming release of the LAPACK library. In this section we want to discuss several practical issues concerning the implementation.

Refining the eigenvalues. Internally the approximations to the eigenvalues

$\lambda_i := \hat{\lambda}_i$ of $\hat{L}\hat{D}\hat{L}^T$ are handled as half-open intervals $[\underline{\lambda}_i, \overline{\lambda}_i)$, with $\underline{\lambda}_i \leq \lambda_i < \overline{\lambda}_i$. In order to identify singletons, neighboring intervals are repeatedly refined using bisection until $\overline{\lambda}_i \leq \underline{\lambda}_{i+1}$ and reldist$(\overline{\lambda}_i, \underline{\lambda}_{i+1})$, as defined in (2.3), is larger than the cluster tolerance, or until reldist$(\underline{\lambda}_i, \overline{\lambda}_{i+1})$ is smaller than the cluster tolerance or the relative width of the intervals becomes smaller than 4ϵ. In the first case, the eigenvalues can safely be regarded as separated, whereas in the second case they cannot.

Computing the eigenpairs. As already mentioned, the final computation of an eigenpair is actually done using RQI with twisted factorizations (see Algorithm 4.1, lines 3–5 and Algorithm 4.2, lines 15–16). For more details on this technique see [6, pp. 136ff.]. Note that for each singular triplet, we need to do this iteration only for $\hat{L}\hat{D}\hat{L}^T$ (or B^TB). The coupling relations guarantee that the resulting refined eigenvalue $\hat{\lambda}$ approximates the corresponding eigenvalue of $\check{L}\check{D}\check{L}^T$ to high relative accuracy, therefore it can directly be used to compute the right singular vector (see Algorithm 4.2, line 17).

Data storage. Algorithms 4.1 and 4.2 describe only the computations for each node in the tree, but not the order in which the nodes are to be visited. In theory, this order has no effect on the algorithm at all. In practice, however, the data for each new child node has to be stored somewhere until it is visited.

It suffices to store enough information to rebuild the three representations belonging to a node using the coupling relations. For level one, we can employ Lemma 4.2 and therefore need only the elements of $\hat{L}, \hat{D}$, and $\hat{S}$ from the dstqds factorization of $B^TB - \mu^2 I$. For deeper levels, the elements of $\tilde{L}, \tilde{D}$ are enough to set up the other two representations $\hat{L}\hat{D}\hat{L}^T$ and $\check{L}\check{D}\check{L}^T$ via Lemma 4.1. As a result, we need to store $4n + \mathcal{O}(1)$ numbers for any node in the three-layered tree. As each node represents at least two singular triplets, we can use, for example, the storage for the first two left and right singular vectors belonging to the node temporarily for this purpose.

With this approach, a breadth-first traversal of the tree is sensible, as this avoids unnecessary swapping of the node data. A similar technique is used in [9] for the implementation of the tridiagonal MRRR algorithm.

IEEE arithmetic. The MRRR algorithm has to deal with possible breakdowns in the factorizations. This is easy to accomplish if support of the IEEE-754 standard for floating point arithmetic is present, or at least an equivalent handling of NaN's (see [10, p. 47]). If this is not the case, special care is necessary to avoid divisions by zero and overflows. The new version `xSTEGR` of the tridiagonal MRRR algorithm works with or without IEEE support [9], and we adapted the employed techniques for the factorizations within `xBDSCR`.

However, we also have to take care of possible division by zero when using the couplings in Lemma 4.2. It was shown in [10, p. 80] how to fix this in the case when IEEE arithmetic is present. In a similar manner to that with the factorizations, these modifications were extended for the case when IEEE arithmetic is not supported.

So, as with the new `xSTEGR`, our code does not need IEEE arithmetic, but is able to exploit it. Preliminary tests indicated approximately a 10% performance improvement with IEEE support, due to the fact that the innermost loop can be formulated with fewer conditionals.

Preprocessing. In [10] it was noticed that for some kinds of matrices, the algorithm can benefit from preceding it by a few sweeps of the bidiagonal QR method, which is described, for example, in [4]. This sort of preprocessing for the original matrix B was integrated in the code, although only some QR sweeps are done per default, as updating the vectors afterward with the employed orthogonal rotations is

not cheap.

Splitting. The MRRR algorithm and consequently its bidiagonal adaptation work only on unreduced matrices; that is, no off-diagonal of the original tridiagonal T, respectively, no element of the bidiagonal B, should be zero.

If some off-diagonal element b_i of the original upper bidiagonal matrix B is zero, the matrix can be split into two submatrices $B_{1:i,1:i}$ and $B_{i+1:n,i+1:n}$, which then can be treated independently. If a diagonal element a_i is zero, an elegant way to "deflate" this zero out is to apply one sweep of the implicit zero-shift QR method, described in [4]. This results in a matrix B' with $b'_{i-1} = b'_{n-1} = a'_n = 0$ [4, p. 21], i.e., one zero singular value has been rotated out nicely and we can split the matrix into three blocks $B_{1:i-1,1:i-1}$, $B_{i:n-1,i:n-1}$ and $B_{n,n}$, the latter one being trivial.

Extensive splitting of the matrix should be exploited wherever possible, as it has a beneficial effect on both orthogonality and runtime. Therefore it is sensible to replace very small elements of B by zero if this affects the SVD only slightly. Standard absolute perturbation theory for the bSVD [3, Cor. 5.1] shows that setting an element c_i of B to zero can cause an absolute change of $|c_i|$ in the singular values and consequently also in the residual (1.3). This suggests the absolute splitting criterion

$$|c_i| \leq \kappa n \epsilon ||B|| \quad \Rightarrow \quad c_i := 0, \tag{5.1}$$

with some small constant κ. However, doing this implies that the singular values will not be computed to high relative accuracy.

Our implementation employs a 2-phase splitting. In the first phase we split the matrix as much as possible *without* spoiling the relative accuracy of the singular values. This can be achieved using powerful criteria developed by Demmel and Kahan in the context of the zero-shift QR algorithm [4, p. 18], or similar but slightly stronger criteria developed by Li, which are based on the dqds transformation (see [19] for details).

Based on the resulting *relative split*, we then apply the absolute splitting criterion (5.1) on each of the blocks and, if necessary, again do a zero-shift QR sweep to deflate zeros on the block diagonals. This results in the *absolute split*, where the blocks are unreduced and subblocks of the relative split.

Then the core bidiagonal MRRR algorithm is applied to each block in the absolute split. Should relative accuracy be desired (indicated by a flag when calling `xBDSCR`), the singular values are afterward refined to high relative accuracy for the respective "father" block in the relative split. Note that there is no need to refine the computed singular vectors in order to get good orthogonality and small residuals.

This splitting approach has two advantages. First, we can always apply the absolute splitting criterion, even if relative accuracy is desired, and if so, we exploit the smaller blocks in the relative split to save runtime when refining the singular values.

6. Comparison with other methods. In this section, we compare our MRRR-based bidiagonal SVD code with other algorithms available in LAPACK. In its current release 3.0, LAPACK provides two driver routines, `xGESVD` and `xGESDD`, for computing the SVD of a general rectangular matrix. In both cases, the matrix is first transformed to bidiagonal form; afterward the singular values are computed from the bidiagonal matrix using the QR algorithm `xBDSQR` or Divide & Conquer `xBDSDC`, respectively. As part of the new release of LAPACK, we will provide a similar driver for our algorithm `xBDSCR`. In the following, we compare the performance of the three computational kernels `xBDSCR`, `xBDSQR`, and `xBDSDC` for the bidiagonal SVD.

As a testbed we used a Pentium 4, 2.8GHz processor with 512kb cache. All

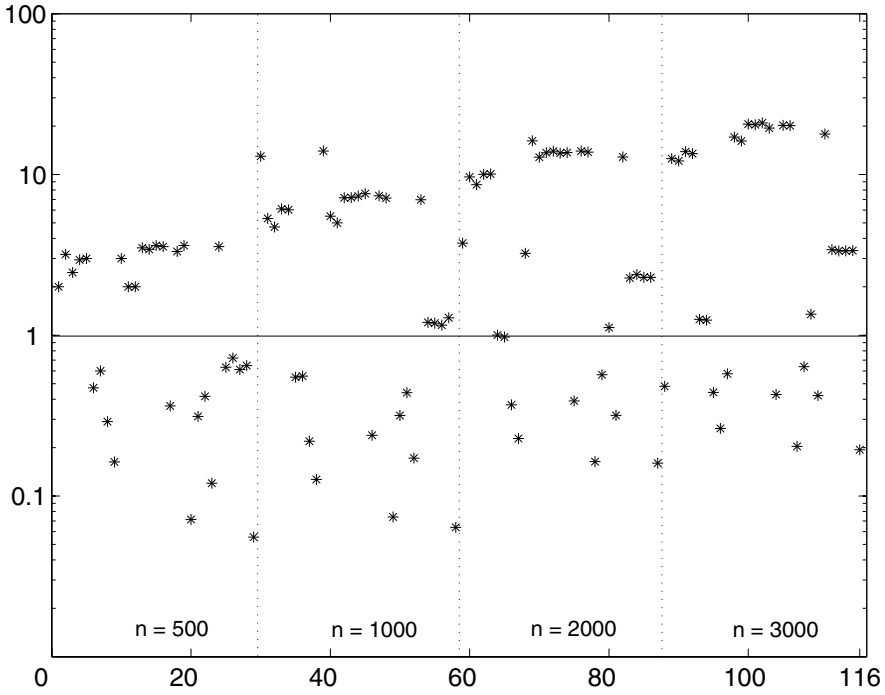


FIG. 6.1. *Speedup of the bidiagonal MRRR algorithm over the Divide & Conquer algorithm for matrices of varying dimension and distribution of the singular values.*

routines and the LAPACK library were compiled using the Intel Fortran Compiler, version 8.1, with compiler options -O3, -tpp7, and -mp.

Figure 6.1 shows the speedup of the bidiagonal MRRR algorithm over the Divide & Conquer algorithm for computing the *full* SVD. For all matrices considered, the QR algorithm was at least five times (in some cases several hundred times) slower than these two algorithms; therefore the QR data is not shown in the figure.

The matrices underlying the figure were designed for testing the robustness of the algorithms. In particular, many of them have very tight, and sometimes large, clusters of singular values. This situation can be favorable for the Divide & Conquer algorithm, which benefits from heavy deflation. By contrast, tight clusters may force the bidiagonal MRRR algorithm to descend several levels in the representation tree, thus increasing its operation count. For these reasons, neither of the two algorithms is consistently superior, in particular for the small matrices. As can be seen in the figure, with increasing matrix dimension the optimal $O(n^2)$ complexity of the bidiagonal MRRR becomes decisive, such that this algorithm is faster in most cases.

Our bidiagonal MRRR routine `xBDSCR` provides the option to compute only selected singular vectors. Table 6.1 shows that this feature can indeed reduce the computation time significantly. For algorithmic reasons, neither the QR nor the Divide & Conquer routine can provide partial SVDs.

Concerning accuracy, each of the three bSVD routines yielded deviations from orthogonality and residuals within the bounds (1.2) and (1.3), respectively. The errors of the bidiagonal MRRR algorithm tend to be larger than those of the remaining two methods, but only by a moderate factor between 10 and 20. As our routine is

TABLE 6.1
Average execution times in seconds for computing a random subset of consecutive singular triplets for the test matrix defined in (3.2) *with dimension* 2000. *Each test was repeated* 10 *times.*

%	DBDSCR	DBDSDC	DBDSQR
100%	2.84	6.63	368
50%	1.56	—	—
25%	1.38	—	—
10%	0.63	—	—

strongly based on the newest implementation `xSTEGR`, the behavior is very similar to that algorithm concerning the comparision of orthogonality, residuals, and runtime with the Divide & Conquer and QR routines. Therefore we refer readers interested in a more detailed discussion of test cases to [5].

7. Conclusions. We have described improvements to the bidiagonal MRRR algorithm and its realization in our new software implementation, which allows for the computation of subsets of k singular values and vectors at $\mathcal{O}(nk)$ cost. Due to the nature of both the QR and the Divide & Conquer algorithms, this functionality was not available; the whole set of singular values and vectors had to be computed at full cost with respect to operations and storage.

As the bidiagonal MRRR algorithm is structurally very similar to the tridiagonal MRRR algorithm for tSEP, it inherits the superior features of the latter. The theoretical complexity of the (bidiagonal) MRRR algorithm is $\mathcal{O}(n^2)$, versus $\mathcal{O}(n^3)$ for the QR algorithm, and Divide & Conquer lies in between $\mathcal{O}(n^2)$ and $\mathcal{O}(n^3)$, depending on the matrix and the amount of deflation. Additionally, the (bidiagonal) MRRR algorithm is naturally parallelizable since the computation for the (singular) vectors within a cluster does not depend on the computation for any other cluster. Furthermore, most of the computation time is spent refining the eigenvalues, and this part is perfectly parallelizable. We plan to develop a parallel version of our algorithm for ScaLAPACK in the future.

As a major challenge for future research we consider the task of devising a stable coupling scheme between successively shifted factorizations of B^TB and BB^T. This would significantly improve and simplify the bidiagonal MRRR algorithm, as then there would be no need to work with the Golub–Kahan matrix anymore. To this end it would be sufficient to develop reliable and cheap criteria in order to test if the couplings proposed in [12] are stable.

An alternative solution to this problem would be to eliminate the need for deeper level couplings at all, that is, to improve the tridiagonal MRRR algorithm in a way that the depth of the representation tree remains limited to one. This would be a major achievement indeed, but at this stage of research it appears to be a very distant goal. As one possible plan of attack in this direction we see a combination of multistep inverse iteration as presented in [21] with some variant of the submatrix method for tightly clustered eigenvalues [16].

Acknowledgments. The authors thank Beresford Parlett and James Demmel for many enlightening discussions. Their contributions have greatly influenced the work leading to this paper. We also thank Osni Marques for letting us build upon his test suite for `xSTEGR` in order to develop a test environment for our own code. In addition, Paul Willems thanks the Research Centre Jülich for financial support during a research visit in Berkeley at the end of 2004 and Beresford Parlett and James Demmel for their hospitality during this time.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, 3rd ed., SIAM, Philadelphia, 1999.

[2] C. Davis and W. M. Kahan, *The rotation of eigenvectors by a perturbation.* III, SIAM J. Numer. Anal., 7 (1970), pp. 1–46.

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

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

[5] J. W. Demmel, O. A. Marques, B. N. Parlett, and C. Vömel, *Performance and Accuracy of the Symmetric Eigensolvers in LAPACK*, University of California, Berkeley, 2005, in preparation.

[6] I. S. Dhillon, *A New $O(n^2)$ Algorithm for the Symmetric Tridiagonal Eigenvalue/Eigenvector Problem*, Ph.D. thesis, University of California, Berkeley, CA, 1997.

[7] 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.

[8] I. S. Dhillon and B. N. Parlett, *Orthogonal eigenvectors and relative gaps*, SIAM J. Matrix Anal. Appl., 25 (2004), pp. 858–899.

[9] I. S. Dhillon, B. N. Parlett, and C. Vömel, *The Design and Implementation of the MRRR Algorithm*, Tech. Report UCBCSD-04-1346, University of California, Berkeley, 2004 (also LAPACK Working Note 162).

[10] B. Großer, *Ein Paralleler und Hochgenauer $O(n^2)$ Algorithmus für die Bidiagonale Singulärwertzerlegung*, Ph.D. thesis, Bergische Universität Wuppertal, Fachbereich Mathematik, Wuppertal, Germany, 2001.

[11] B. Großer and B. Lang, *On symmetric eigenproblems induced by the bidiagonal SVD*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 599–620.

[12] B. Großer and B. Lang, *An $O(n^2)$ algorithm for the bidiagonal SVD*, Linear Algebra Appl., 358 (2003), pp. 45–70.

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

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

[15] B. N. Parlett, *The new gd algorithms*, Cambridge University Press, (1995), pp. 459–491.

[16] B. N. Parlett, *Invariant subspaces for tightly clustered eigenvalues of tridiagonals*, BIT, 36 (1996), pp. 542–562.

[17] B. N. Parlett and I. S. Dhillon, *Fernando's solution to Wilkinson's problem: An application of double factorization*, Linear Algebra Appl., 267 (1997), pp. 247–279.

[18] B. N. Parlett and I. S. Dhillon, *Relatively robust representations of symmetric tridiagonals*, Linear Algebra Appl., 309 (2000), pp. 121–151.

[19] B. N. Parlett and O. Marques, *An implementation of the DQDS algorithm (positive case)*, Linear Algebra Appl., 309 (2000), pp. 217–259.

[20] B. N. Parlett and C. Vömel, *How the MRRR Algorithm Can Fail on Tight Eigenvalue Clusters*, Tech. Report UCB-CSD-04-1367, University of California, Berkeley, 2004 (also LAPACK Working Note 163).

[21] B. N. Parlett and C. Vömel, *How to Improve FP Vectors in the MRRR Algorithm by Twisted Inverse Iteration*, University of California, Berkeley, 2005, in preparation.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 4, pp. 927–948

UNIFORM ACCURACY OF EIGENPAIRS FROM A SHIFT-INVERT LANCZOS METHOD*

U. L. HETMANIUK† AND R. B. LEHOUCQ†

Abstract. This paper analyzes the accuracy of the shift-invert Lanczos iteration for computing eigenpairs of the symmetric definite generalized eigenvalue problem. We provide bounds for the accuracy of the eigenpairs produced by shift-invert Lanczos given a residual reduction. We discuss the implications of our analysis for practical shift-invert Lanczos iterations. When the generalized eigenvalue problem arises from a conforming finite element method, we also comment on the uniform accuracy of bounds (independent of the mesh size h).

Key words. shift-invert Lanczos decomposition, symmetric generalized eigenvalue problem, inner product

AMS subject classifications. 65F15, 65N25

DOI. 10.1137/050629288

1. Introduction.

A popular approach for the solution of the generalized symmetric eigenvalue problem

$$\mathbf{A}\mathbf{u} = \mathbf{M}\mathbf{u}\lambda, \qquad \left(\mathbf{A}, \mathbf{M} \in \mathbb{R}^{n\times n}\right), \tag{1.1}$$

is the shift-invert Lanczos method [3, 4, 5]. The pencil $(\mathbf{A}, \mathbf{M})$ is symmetric definite, i.e., the matrices $\mathbf{A}$ and $\mathbf{M}$ are, respectively, symmetric and symmetric positive semidefinite and some combination $\alpha\mathbf{A} + \mu\mathbf{M}$ is positive definite. This technique is used to approximate the eigenvalues in a given interval, for instance, the smallest eigenvalues.

Once approximations to the eigenvalues (and eigenvectors) are computed, a posteriori error bounds can assess the accuracy of the results. Ericsson and Ruhe [3] presented the first bounds for (1.1) when the matrix $\mathbf{M}$ is positive definite and the Lanczos basis is $\mathbf{M}$-orthonormal. In this paper, we review their results and extend them to an $(\alpha\mathbf{A} + \mu\mathbf{M})$-orthonormal Lanczos basis (where the matrix $\alpha\mathbf{A} + \mu\mathbf{M}$ is positive definite). We also derive the explicit constants for one- and two-sided bounds.

When the matrix $\mathbf{M}$ is symmetric positive semidefinite and some combination $\alpha\mathbf{A}+\mu\mathbf{M}$ is positive definite, an approach is to apply the shift-invert Lanczos method to the pencil $(\mathbf{A}, \alpha\mathbf{A}+\mu\mathbf{M})$ because the eigenvectors are unchanged and the eigenvalues are easily recovered. Another approach was introduced in [6] where Nour-Omid et al. explained how to implement shift-invert Lanczos for the pencil $(\mathbf{A}, \mathbf{M})$ in the range of $\mathbf{M}$ and build an $\mathbf{M}$-orthonormal basis. Our motivation for using the $(\alpha\mathbf{A} + \mu\mathbf{M})$-inner product with the pencil $(\mathbf{A}, \mathbf{M})$ arises in the definition of the stopping criterion with a more general norm while still building the same Krylov space as Nour-Omid et al. [6]. The norm involved in the stopping criterion is important as an inappropriate choice can result in more Lanczos iterations than necessary.

*Received by the editors April 15, 2005; accepted for publication (in revised form) by B. Parlett September 9, 2005; published electronically December 18, 2006.
http://www.siam.org/journals/simax/28-4/62928.html

†Sandia National Laboratories, P.O. Box 5800, MS 1110, Albuquerque, NM 87185-1110 (ulhetma@sandia.gov, rblehou@sandia.gov). Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000.

Our analysis also has implications on practical shift-invert Lanczos iterations. In particular, when the eigenvalue problem (1.1) arises from the finite element discretization of elliptic self-adjoint differential eigenvalue problem, invariance of the bounds with respect to the mesh size is an important property. We study the bounds with respect to the mesh size.

Our paper is organized as follows. Section 2 reviews the shift-invert Lanczos decomposition and introduces our notation. Section 3 reviews useful general accuracy results. Sections 4, 5, and 6 apply these accuracy results to approximations generated by shift-invert Lanczos iterations. Finally, section 7 comments on the results of our analysis for practical Lanczos iterations. For the sake of clarity, we gathered most of the proofs in the appendix.

2. Shift-invert Lanczos decomposition. The eigenvalue problem (1.1) has a set of $\mathbf{M}$-orthogonal eigenvectors $\mathbf{u}_j$ and corresponding eigenvalues λ_j. When $\mathbf{M}$ is positive semidefinite, the symmetric pencil $(\mathbf{A}, \mathbf{M})$ has infinite eigenvalues. In most practical cases, these infinite eigenvalues are not of interest. Hence we focus in this report on the real finite eigenvalues and we assume that these finite eigenvalues (and associated eigenvectors) are ordered in ascending order.

In order to use the shift-invert Lanczos iteration to compute eigenpairs of (1.1), a spectral transformation is employed. If $\sigma \in \mathbb{R}$, then the standard eigenvalue problem

$$(\mathbf{A} - \sigma\mathbf{M})^{-1}\mathbf{M}\mathbf{u} = \mathbf{u}\nu, \qquad \left(\nu = \frac{1}{\lambda - \sigma}\right) \tag{2.1}$$

results by subtracting $\sigma\mathbf{M}$ from both sides of (1.1) followed by "cross-multiplication." This standard eigenvalue problem is no longer symmetric. However, a careful choice of inner product renders the operator $(\mathbf{A} - \sigma\mathbf{M})^{-1}\mathbf{M}$ symmetric. For instance, when the inner product is induced by the matrix $\mathbf{H}$, selecting $\mathbf{H}$ equal to $\alpha\mathbf{A} + \mu\mathbf{M}$ results in an $\mathbf{H}$-symmetric matrix $(\mathbf{A} - \sigma\mathbf{M})^{-1}\mathbf{M}$ (where $\alpha\mathbf{A} + \mu\mathbf{M}$ is symmetric positive definite). To see this symmetry, note that

$$\alpha\mathbf{A} + \mu\mathbf{M} = \alpha(\mathbf{A} - \sigma\mathbf{M}) + (\mu + \sigma\alpha)\mathbf{M}$$

and so

$$\begin{aligned}(\alpha\mathbf{A} + \mu\mathbf{M})(\mathbf{A} - \sigma\mathbf{M})^{-1}\mathbf{M} &= \alpha\mathbf{M} + (\mu + \sigma\alpha)\mathbf{M}(\mathbf{A} - \sigma\mathbf{M})^{-1}\mathbf{M} \\ &= \mathbf{M}(\mathbf{A} - \sigma\mathbf{M})^{-1}(\alpha\mathbf{A} + \mu\mathbf{M}).\end{aligned}$$

Suppose that

$$\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{V}_j = \mathbf{V}_j\mathbf{T}_j + \mathbf{f}_j\mathbf{e}_j^T, \qquad (\mathbf{A}_\sigma = \mathbf{A} - \sigma\mathbf{M}) \tag{2.2}$$

is a Lanczos reduction of length j where $\mathbf{e}_j$ is the jth canonical basis vector, we have

$$\mathbf{V}_j^T\mathbf{H}\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{V}_j = \mathbf{T}_j, \tag{2.3a}$$

$$\mathbf{V}_j^T\mathbf{H}\mathbf{V}_j = \mathbf{I}_j, \tag{2.3b}$$

$$\mathbf{V}_j^T\mathbf{H}\mathbf{f}_j = \mathbf{0}, \tag{2.3c}$$

where $\mathbf{T}_j$ is a symmetric tridiagonal matrix. The j columns of $\mathbf{V}_j$ form a basis $\mathbf{H}$-orthonormal for the Krylov subspace

$$\mathcal{K}_j(\mathbf{A}_\sigma^{-1}\mathbf{M}, \mathbf{v}_1) = \mathrm{Span}\{\mathbf{v}_1, \mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{v}_1, \ldots, (\mathbf{A}_\sigma^{-1}\mathbf{M})^{j-1}\mathbf{v}_1\}. \tag{2.4}$$

If we denote

$$\mathbf{T}_j = \begin{pmatrix} \alpha_1 & \beta_1 & \cdots & 0 \\ \beta_1 & \alpha_2 & \cdots & 0 \\ \vdots & & \ddots & \beta_{j-1} \\ 0 & \cdots & \beta_{j-1} & \alpha_j \end{pmatrix},$$

then the familiar Lanczos three-term recurrence is recovered by equating the jth column of (2.2) to obtain

$$\mathbf{f}_j = \mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{v}_j - \mathbf{v}_j\alpha_j - \mathbf{v}_{j-1}\beta_{j-1}.$$

Using $\mathbf{H}$-orthonormality, we have

$$\alpha_j = \mathbf{v}_j^T\mathbf{H}\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{v}_j, \tag{2.5a}$$

$$\beta_j = \|\mathbf{f}_j\|_{\mathbf{H}} = \sqrt{\mathbf{f}_j^T\mathbf{H}\mathbf{f}_j}, \tag{2.5b}$$

and the new direction $\mathbf{v}_{j+1}$ is equal to $\mathbf{f}_j/\beta_j$, where we assume that β_j is not zero. This case is eliminated when every tridiagonal matrix $\mathbf{T}_j$ is unreduced or, equivalently, has simple eigenpairs.

The Lanczos reduction (2.2) provides two choices of approximating eigenvectors. If $(\mathbf{s}, \theta)$ is an eigenpair for $\mathbf{T}_j$ such that

$$\mathbf{T}_j\mathbf{s} = \mathbf{s}\theta \quad \text{and} \quad \|\mathbf{s}\| = \sqrt{\mathbf{s}^T\mathbf{s}} = 1,$$

then we can postmultiply (2.2) by $\mathbf{s}$ to obtain

$$\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{V}_j\mathbf{s} = \mathbf{V}_j\mathbf{s}\theta + \mathbf{f}_j\omega_j \tag{2.6a}$$

$$= (\mathbf{V}_j\mathbf{s} + \mathbf{f}_j\frac{\omega_j}{\theta})\theta, \tag{2.6b}$$

where $\omega_j = \mathbf{e}_j^T\mathbf{s}$. To approximate an eigenvector, we can use the Ritz vector, defined as

$$\mathbf{x} = \mathbf{V}_j\mathbf{s}, \tag{2.7}$$

and the purified vector, defined as

$$\mathbf{p} = \mathbf{V}_j\mathbf{s} + \mathbf{f}_j\frac{\omega_j}{\theta} = \mathbf{x} + \mathbf{f}_j\frac{\omega_j}{\theta}. \tag{2.8}$$

The purified vector was introduced in [3, 6] as a simple postprocessing step to recover the vector that results from a step of inverse iteration on the Ritz vector.

For a desired tolerance ε, a convergence criterion often used in practice [3, 5] is

$$\|\mathbf{f}_j\|_{\mathbf{H}}\,|\omega_j| \le \varepsilon\,|\theta|\,. \tag{2.9}$$

Note that by (2.5b), the convergence criterion is available as a by-product of the Lanczos reduction.

In the remainder of this report, we omit the index j when the context is clear and we always assume that σ is an arbitrary real number such that $\mathbf{A}_\sigma$ is invertible. We also emphasize that the matrix

$$\mathbf{H} = \alpha\mathbf{A} + \mu\mathbf{M} \tag{2.10}$$

is symmetric positive definite. Finally, we caution the reader that a distinction is drawn between the inner product used for orthogonality of the Lanczos vectors and the inner product used for the error bounds.

3. General results. This section recalls several standard accuracy results that will be useful for our analysis. These results provide bounds on the errors when approximating an eigenpair in terms of its residual. First, we review a general result for a simple eigenvalue problem.

THEOREM 3.1. *Let $\hat{\mathbf{A}} \in \mathbb{R}^{n\times n}$ be a symmetric matrix, $\hat{\mathbf{y}}$ a nonzero vector in $\mathbb{R}^n$, $\hat{\theta}$ a real number, and $\hat{\mathbf{r}}$ the residual vector*

$$\hat{\mathbf{r}} = \hat{\mathbf{A}}\hat{\mathbf{y}} - \hat{\mathbf{y}}\hat{\theta}.$$

If $\hat{\alpha}$ is the eigenvalue of $\hat{\mathbf{A}}$ closest to $\hat{\theta}$, where $\hat{\mathbf{A}}\hat{\mathbf{z}} = \hat{\mathbf{z}}\hat{\alpha}$ and $\|\hat{\mathbf{z}}\| = 1$, then

$$0 \leq \left|\hat{\theta} - \hat{\alpha}\right| \leq \frac{\|\hat{\mathbf{r}}\|}{\|\hat{\mathbf{y}}\|}, \tag{3.1a}$$

$$0 \leq |\sin\angle(\hat{\mathbf{y}}, \hat{\mathbf{z}})| \leq \frac{1}{\min\limits_{\hat{\lambda}_i \neq \hat{\alpha}} \left|\hat{\lambda}_i - \hat{\theta}\right|} \frac{\|\hat{\mathbf{r}}\|}{\|\hat{\mathbf{y}}\|}. \tag{3.1b}$$

Furthermore, if $\hat{\theta}$ is the Rayleigh quotient of $\hat{\mathbf{y}}$,

$$\hat{\theta} = \frac{\hat{\mathbf{y}}^T\hat{\mathbf{A}}\hat{\mathbf{y}}}{\hat{\mathbf{y}}^T\hat{\mathbf{y}}},$$

then

$$0 \leq \left|\hat{\theta} - \hat{\alpha}\right| \leq \min\left(\frac{\|\hat{\mathbf{r}}\|}{\|\hat{\mathbf{y}}\|}, \frac{1}{\min\limits_{\hat{\lambda}_i \neq \hat{\alpha}} \left|\hat{\lambda}_i - \hat{\theta}\right|} \frac{\|\hat{\mathbf{r}}\|^2}{\|\hat{\mathbf{y}}\|^2}\right), \tag{3.2a}$$

$$\frac{1}{\hat{\lambda}_n - \hat{\lambda}_1} \frac{\|\hat{\mathbf{r}}\|}{\|\hat{\mathbf{y}}\|} \leq |\sin\angle(\hat{\mathbf{y}}, \hat{\mathbf{z}})| \leq \frac{1}{\min\limits_{\hat{\lambda}_i \neq \hat{\alpha}} \left|\hat{\lambda}_i - \hat{\theta}\right|} \frac{\|\hat{\mathbf{r}}\|}{\|\hat{\mathbf{y}}\|}. \tag{3.2b}$$

$\hat{\lambda}_n$ and $\hat{\lambda}_1$ are, respectively, the largest and smallest eigenvalue of $\hat{\mathbf{A}}$.

Proof. See Parlett [7, section 11.7]. ∎

Next, we derive several accuracy bounds to assess the approximation of an eigenpair $(\mathbf{u}, \lambda)$ for the pencil $(\mathbf{A}, \mathbf{M})$. The difference between the bounds arises from the various norms used to measure the residual.

PROPOSITION 3.2. *Let $(\mathbf{A}, \mathbf{M})$ be a symmetric definite pencil, $\mathbf{H} = \alpha\mathbf{A} + \mu\mathbf{M}$ a symmetric positive definite matrix, and σ a real number such that $\mathbf{A}_\sigma$ is invertible. We define also $\mathbf{y}$ a nonzero vector in $\mathbb{R}^n$, $\tilde{\lambda}$ a real number such that $\tilde{\lambda} \neq \sigma$, and the residual vector*

$$\mathbf{r} = \mathbf{A}\mathbf{y} - \mathbf{M}\mathbf{y}\tilde{\lambda}.$$

If λ satisfies

$$\left|\frac{1}{\lambda - \sigma} - \frac{1}{\tilde{\lambda} - \sigma}\right| = \min_{\lambda_i} \left|\frac{1}{\lambda_i - \sigma} - \frac{1}{\tilde{\lambda} - \sigma}\right| \tag{3.3}$$

and $\mathbf{A}\mathbf{u} = \mathbf{M}\mathbf{u}\lambda$, where $\|\mathbf{u}\|_{\mathbf{H}} = 1$, then

$$\left|\frac{\lambda - \tilde{\lambda}}{\lambda - \sigma}\right| \leq \frac{\|\mathbf{r}\|_{\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}}}{\|\mathbf{y}\|_{\mathbf{H}}}, \tag{3.4a}$$

$$0 \leq |\sin\angle_{\mathbf{H}}(\mathbf{y}, \mathbf{u})| \leq \left|\frac{\lambda_\gamma - \sigma}{\lambda_\gamma - \tilde{\lambda}}\right| \frac{\|\mathbf{r}\|_{\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}}}{\|\mathbf{y}\|_{\mathbf{H}}}, \tag{3.4b}$$

where

$$\left|\frac{1}{\lambda_\gamma-\sigma}-\frac{1}{\tilde\lambda-\sigma}\right| = \min_{\lambda_i\neq\lambda}\left|\frac{1}{\lambda_i-\sigma}-\frac{1}{\tilde\lambda-\sigma}\right|.$$

In addition, when $\tilde\lambda$ *satisfies*

$$\tilde\lambda = \sigma + \frac{\mathbf{y}^T\mathbf{H}\mathbf{y}}{\mathbf{y}^T\mathbf{H}\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{y}}, \tag{3.5}$$

we have

$$\left|\frac{\lambda-\tilde\lambda}{\lambda-\sigma}\right| \leq \frac{\|\mathbf{r}\|_{\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}}}{\|\mathbf{y}\|_{\mathbf{H}}} \min\left(1, \left|\frac{\lambda_\gamma-\sigma}{\lambda_\gamma-\tilde\lambda}\right| \frac{\|\mathbf{r}\|_{\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}}}{\|\mathbf{y}\|_{\mathbf{H}}}\right), \tag{3.6a}$$

$$\left|\frac{(\lambda_\sigma^+-\sigma)(\lambda_\sigma^--\sigma)}{(\lambda_\sigma^+-\lambda_\sigma^-)(\tilde\lambda-\sigma)}\right| \frac{\|\mathbf{r}\|_{\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}}}{\|\mathbf{y}\|_{\mathbf{H}}} \leq |\sin\angle_{\mathbf{H}}(\mathbf{y},\mathbf{u})| \leq \left|\frac{\lambda_\gamma-\sigma}{\lambda_\gamma-\tilde\lambda}\right| \frac{\|\mathbf{r}\|_{\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}}}{\|\mathbf{y}\|_{\mathbf{H}}}, \tag{3.6b}$$

where

$$(\lambda_\sigma^-,\lambda_\sigma^+) = \begin{cases} (\arg\min_{\lambda_i<\sigma}|\lambda_i-\sigma|, \arg\min_{\sigma<\lambda_i}|\lambda_i-\sigma|) & \text{when } \lambda_1<\sigma<\lambda_n, \\ (\lambda_1,\lambda_n) & \text{otherwise.} \end{cases}$$

Proof. This result is a reformulation of Theorem 3.1 when $\hat{\mathbf{A}}$, $\hat{\mathbf{y}}$, $\hat\theta$, and $\hat{\mathbf{r}}$ satisfy

$$\begin{cases} \hat{\mathbf{A}} &= \mathbf{H}^{-1/2}\mathbf{H}\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{H}^{-1/2}, \\ \hat{\mathbf{y}} &= \mathbf{H}^{1/2}\mathbf{y}, \\ \hat\theta &= (\tilde\lambda-\sigma)^{-1}, \\ \hat{\mathbf{r}} &= \hat{\mathbf{A}}\hat{\mathbf{y}}-\hat{\mathbf{y}}\hat\theta = \mathbf{H}^{1/2}\mathbf{A}_\sigma^{-1}(\mathbf{M}\mathbf{x}\tilde\lambda-\mathbf{A}\mathbf{x})(\tilde\lambda-\sigma)^{-1}. \end{cases}$$

The Rayleigh quotient of $\hat{\mathbf{y}}$ satisfies

$$\frac{\hat{\mathbf{y}}^T\hat{\mathbf{A}}\hat{\mathbf{y}}}{\hat{\mathbf{y}}^T\hat{\mathbf{y}}} = \frac{\mathbf{y}^T\mathbf{H}\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{y}}{\mathbf{y}^T\mathbf{H}\mathbf{y}}. \qquad \square$$

When $\tilde\lambda$ satisfies (3.5), quadratic convergence of $\tilde\lambda$ towards the eigenvalue λ is triggered when

$$\left|\frac{\lambda_\gamma-\sigma}{\lambda_\gamma-\tilde\lambda}\right| \frac{\|\mathbf{r}\|_{\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}}}{\|\mathbf{y}\|_{\mathbf{H}}} < 1.$$

The eigenvalue bounds (3.4a) and (3.6a) guarantee a relative error on a nonzero eigenvalue of the same level as the "normalized" residual norm because

$$\left|\frac{\lambda-\tilde\lambda}{\lambda}\right| \leq \left|\frac{\lambda-\sigma}{\lambda}\right| \frac{\|\mathbf{r}\|_{\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}}}{\|\mathbf{y}\|_{\mathbf{H}}}.$$

When $\left|\tilde\lambda-\sigma\right|$ increases, so does $|\lambda_\gamma-\sigma|$. Then the bounds (3.6b) for the ratio between the sine of the angle and the residual norm

$$\left|\frac{(\lambda_\sigma^+-\sigma)(\lambda_\sigma^--\sigma)}{(\lambda_\sigma^+-\lambda_\sigma^-)(\tilde\lambda-\sigma)}\right| \leq \frac{|\sin\angle_{\mathbf{H}}(\mathbf{y},\mathbf{u})|\,\|\mathbf{y}\|_{\mathbf{H}}}{\|\mathbf{A}\mathbf{y}-\mathbf{M}\mathbf{y}\tilde\lambda\|_{\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}}} \leq \left|\frac{\lambda_\gamma-\sigma}{\lambda_\gamma-\tilde\lambda}\right|$$

define a wider interval. When σ is close to an eigenvalue and $\tilde{\lambda}$ approximates a different eigenvalue, the constants in all the bounds are bounded.

When σ is close to an eigenvalue and $\tilde{\lambda}$ approximates the same eigenvalue, relation (3.3) constrains $\tilde{\lambda}$. For instance, when the shift σ satisfies

$$\frac{\lambda_{j-1}+\lambda_j}{2} < \sigma < \frac{\lambda_j+\lambda_{j+1}}{2},$$

then $\tilde{\lambda}$ approximates λ_j if and only if the constraints

$$\min\left(\sigma, \max_{\lambda_i\neq\lambda_j} \frac{2\lambda_j\lambda_i-\sigma(\lambda_j+\lambda_i)}{\lambda_j+\lambda_i-2\sigma}\right) < \tilde{\lambda} < \max\left(\sigma, \min_{\lambda_i\neq\lambda_j} \frac{2\lambda_j\lambda_i-\sigma(\lambda_j+\lambda_i)}{\lambda_j+\lambda_i-2\sigma}\right)$$

are satisfied. Hence, the constant

$$\left|\frac{\lambda_\gamma-\sigma}{\lambda_\gamma-\tilde{\lambda}}\right|$$

approaches 1 when σ and $\tilde{\lambda}$ are near λ.

For the next two results, we need to make further assumptions on $\mathbf{M}$ or $\mathbf{A}$ to define, with one of these matrices, an inner product.

PROPOSITION 3.3. *Let $(\mathbf{A},\mathbf{M})$ be a symmetric definite pencil where $\mathbf{M}$ is symmetric positive definite. We define also $\mathbf{y}$ a nonzero vector, $\tilde{\lambda}$ a real number, and $\mathbf{r}$ the residual vector*

$$\mathbf{r} = \mathbf{A}\mathbf{y} - \mathbf{M}\mathbf{y}\tilde{\lambda}.$$

If λ is the closest eigenvalue to $\tilde{\lambda}$ and $\mathbf{A}\mathbf{u} = \mathbf{M}\mathbf{u}\lambda$, where $\|\mathbf{u}\|_{\mathbf{M}} = 1$, then

$$\left|\lambda-\tilde{\lambda}\right| \leq \frac{\|\mathbf{r}\|_{\mathbf{M}^{-1}}}{\|\mathbf{y}\|_{\mathbf{M}}}, \tag{3.7a}$$

$$0 \leq |\sin\angle_{\mathbf{M}}(\mathbf{y},\mathbf{u})| \leq \frac{1}{\min\limits_{\lambda_i\neq\lambda}\left|\lambda_i-\tilde{\lambda}\right|}\frac{\|\mathbf{r}\|_{\mathbf{M}^{-1}}}{\|\mathbf{y}\|_{\mathbf{M}}}. \tag{3.7b}$$

In addition, when $\tilde{\lambda}$ satisfies

$$\tilde{\lambda} = \frac{\mathbf{y}^T\mathbf{A}\mathbf{y}}{\mathbf{y}^T\mathbf{M}\mathbf{y}},$$

we have

$$\left|\lambda-\tilde{\lambda}\right| \leq \frac{\|\mathbf{r}\|_{\mathbf{M}^{-1}}}{\|\mathbf{y}\|_{\mathbf{M}}} \min\left(1, \frac{1}{\min\limits_{\lambda_i\neq\lambda}\left|\lambda_i-\tilde{\lambda}\right|}\frac{\|\mathbf{r}\|_{\mathbf{M}^{-1}}}{\|\mathbf{y}\|_{\mathbf{M}}}\right), \tag{3.8a}$$

$$\frac{1}{\lambda_n-\lambda_1}\frac{\|\mathbf{r}\|_{\mathbf{M}^{-1}}}{\|\mathbf{y}\|_{\mathbf{M}}} \leq |\sin\angle_{\mathbf{M}}(\mathbf{y},\mathbf{u})| \leq \frac{1}{\min\limits_{\lambda_i\neq\lambda}\left|\lambda_i-\tilde{\lambda}\right|}\frac{\|\mathbf{r}\|_{\mathbf{M}^{-1}}}{\|\mathbf{y}\|_{\mathbf{M}}}. \tag{3.8b}$$

Proof. This result is a reformulation of Theorem 3.1 when $\hat{\mathbf{A}}$, $\hat{\mathbf{y}}$, $\hat{\theta}$, and $\hat{\mathbf{r}}$ satisfy

$$\begin{cases} \hat{\mathbf{A}} &= \mathbf{M}^{-1/2}\mathbf{A}\mathbf{M}^{-1/2}, \\ \hat{\mathbf{y}} &= \mathbf{M}^{1/2}\mathbf{y}, \\ \hat{\theta} &= \tilde{\lambda}, \\ \hat{\mathbf{r}} &= \hat{\mathbf{A}}\hat{\mathbf{y}} - \hat{\mathbf{y}}\hat{\theta} = \mathbf{M}^{-1/2}(\mathbf{A}\mathbf{y}-\mathbf{M}\mathbf{y}\tilde{\lambda}). \end{cases}$$

□

Quadratic convergence towards the eigenvalue λ is triggered as soon as

$$\frac{1}{\min\limits_{\lambda_i \neq \lambda} \left|\lambda_i - \tilde{\lambda}\right|} \frac{\|\mathbf{r}\|_{\mathbf{M}^{-1}}}{\|\mathbf{y}\|_{\mathbf{M}}} < 1.$$

When λ belongs to a cluster of eigenvalues, this quadratic convergence will require a small residual norm and the constant in the upper bound on the angle will become large. Furthermore, when the spread $\lambda_n - \lambda_1$ is large, the lower bound in (3.8b) becomes crude.

PROPOSITION 3.4. *Let* $(\mathbf{A}, \mathbf{M})$ *be a symmetric definite pencil where* $\mathbf{A}$ *is symmetric positive definite. We define also* $\mathbf{y}$ *a nonzero vector,* $\tilde{\lambda}$ *a nonzero real number, and* $\mathbf{r}$ *the residual vector*

$$\mathbf{r} = \mathbf{A}\mathbf{y} - \mathbf{M}\mathbf{y}\tilde{\lambda}.$$

If $1/\lambda$ *is the closest reciprocal eigenvalue to* $1/\tilde{\lambda}$ *and* $\mathbf{A}\mathbf{u} = \mathbf{M}\mathbf{u}\lambda$*, then*

$$\left|\frac{\lambda - \tilde{\lambda}}{\lambda}\right| \leq \frac{\|\mathbf{r}\|_{\mathbf{A}^{-1}}}{\|\mathbf{y}\|_{\mathbf{A}}}, \tag{3.9a}$$

$$0 \leq |\sin \angle_{\mathbf{A}}(\mathbf{y}, \mathbf{u})| \leq \left|\frac{\lambda_\delta}{\lambda_\delta - \tilde{\lambda}}\right| \frac{\|\mathbf{r}\|_{\mathbf{A}^{-1}}}{\|\mathbf{y}\|_{\mathbf{A}}}, \tag{3.9b}$$

where

$$\left|\frac{1}{\lambda_\delta} - \frac{1}{\tilde{\lambda}}\right| = \min_{\lambda_i \neq \lambda} \left|\frac{1}{\lambda_i} - \frac{1}{\tilde{\lambda}}\right|.$$

In addition, when $\tilde{\lambda}$ *satisfies*

$$\tilde{\lambda} = \frac{\mathbf{y}^T \mathbf{A}\mathbf{y}}{\mathbf{y}^T \mathbf{M}\mathbf{y}},$$

we have

$$\left|\frac{\lambda - \tilde{\lambda}}{\lambda}\right| \leq \frac{\|\mathbf{r}\|_{\mathbf{A}^{-1}}}{\|\mathbf{y}\|_{\mathbf{A}}} \min\left(1, \left|\frac{\lambda_\delta}{\lambda_\delta - \tilde{\lambda}}\right| \frac{\|\mathbf{r}\|_{\mathbf{A}^{-1}}}{\|\mathbf{y}\|_{\mathbf{A}}}\right), \tag{3.10a}$$

$$\frac{\lambda_1 \lambda_n}{(\lambda_n - \lambda_1)\tilde{\lambda}} \frac{\|\mathbf{r}\|_{\mathbf{A}^{-1}}}{\|\mathbf{y}\|_{\mathbf{A}}} \leq |\sin \angle_{\mathbf{A}}(\mathbf{y}, \mathbf{u})| \leq \left|\frac{\lambda_\delta}{\lambda_\delta - \tilde{\lambda}}\right| \frac{\|\mathbf{r}\|_{\mathbf{A}^{-1}}}{\|\mathbf{y}\|_{\mathbf{A}}}. \tag{3.10b}$$

Proof. This result is a reformulation of Theorem 3.1 when $\hat{\mathbf{A}}$, $\hat{\mathbf{y}}$, $\hat{\theta}$, and $\hat{\mathbf{r}}$ satisfy

$$\begin{cases} \hat{\mathbf{A}} &= \mathbf{A}^{-1/2}\mathbf{M}\mathbf{A}^{-1/2}, \\ \hat{\mathbf{y}} &= \mathbf{A}^{1/2}\mathbf{y}, \\ \hat{\theta} &= \tilde{\lambda}^{-1}, \\ \hat{\mathbf{r}} &= \hat{\mathbf{A}}\hat{\mathbf{y}} - \hat{\mathbf{y}}\hat{\theta} = \mathbf{A}^{-1/2}(\mathbf{M}\mathbf{y}\tilde{\lambda} - \mathbf{A}\mathbf{y})\tilde{\lambda}^{-1}. \end{cases} \qquad \square$$

The bounds (3.9a) and (3.10a) guarantee a relative error on the eigenvalue of the same level as the "normalized" residual norm. Quadratic convergence towards the eigenvalue λ is triggered as soon as

$$\left|\frac{\lambda_\delta}{\lambda_\delta - \tilde{\lambda}}\right| \frac{\|\mathbf{r}\|_{\mathbf{A}^{-1}}}{\|\mathbf{y}\|_{\mathbf{A}}} < 1.$$

We remark that when λ_n is large, the lower bound of (3.10b) is not modified. On the other hand, when the Rayleigh quotient of $\mathbf{y}$ gets larger, the bounds (3.10b) for the ratio between the sine of the angle and the residual norm

$$\frac{\lambda_1\lambda_n}{(\lambda_n-\lambda_1)\tilde{\lambda}} \leq \frac{|\sin\angle_{\mathbf{A}}(\mathbf{y},\mathbf{u})|\,\|\mathbf{y}\|_{\mathbf{A}}}{\|\mathbf{A}\mathbf{y}-\mathbf{M}\mathbf{y}\tilde{\lambda}\|_{\mathbf{A}^{-1}}} \leq \left|\frac{\lambda_\delta}{\lambda_\delta-\tilde{\lambda}}\right|$$

define a wider interval.

In sections 4–6, we will consider different approximations, generated by shift-invert Lanczos iterations, for an eigenpair $(\mathbf{u},\lambda)$. In order to apply the previous propositions, we will evaluate the different norms of the corresponding residuals. When possible, we will give explicit expressions of the norms. Else, we will give asymptotic expansions.

4. Study of the Ritz vector and the Ritz value. In this section, an eigenpair $(\mathbf{u},\lambda)$ is approximated by the Ritz vector $\mathbf{x}$ (2.7) and the value

$$\sigma+\frac{1}{\theta}=\sigma+\frac{\mathbf{x}^T\mathbf{H}\mathbf{x}}{\mathbf{x}^T\mathbf{H}\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{x}},$$

where θ is the eigenvalue of $\mathbf{T}$ associated with the Ritz vector $\mathbf{x}$. We also emphasize that the matrix $\mathbf{H}=\alpha\mathbf{A}+\mu\mathbf{M}$ is symmetric positive definite. The next result relates the Lanczos vector $\mathbf{f}$ with different norms of the residual.

PROPOSITION 4.1. *Let $\mathbf{x}$ be a Ritz vector* (2.7) *produced by a shift-invert Lanczos reduction where the Lanczos vectors are $\mathbf{H}$-orthonormal. Let $\sigma+1/\theta$ approximate an eigenvalue, where*

$$\sigma+\frac{1}{\theta}=\sigma+\frac{\mathbf{x}^T\mathbf{H}\mathbf{x}}{\mathbf{x}^T\mathbf{H}\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{x}}. \tag{4.1}$$

If $\mathbf{r}=\mathbf{A}\mathbf{x}-\mathbf{M}\mathbf{x}(\sigma+1/\theta)$, then we have

$$\frac{\|\mathbf{r}\|_{\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}}}{\|\mathbf{x}\|_{\mathbf{H}}}=\left|\frac{\omega}{\theta}\right|\|\mathbf{f}\|_{\mathbf{H}}. \tag{4.2}$$

When $\mathbf{M}$ is symmetric positive definite, we have

$$\frac{\|\mathbf{r}\|_{\mathbf{M}^{-1}}}{\|\mathbf{x}\|_{\mathbf{M}}}=\|\mathbf{f}\|_{\mathbf{A}_\sigma\mathbf{M}^{-1}\mathbf{A}_\sigma}\left|\frac{\omega}{\theta}\right|\sqrt{\mu+\alpha\sigma+\frac{\alpha}{\theta}}+\mathcal{O}\left(\|\mathbf{f}\|_{\mathbf{H}}^2\frac{\omega^2}{\theta^2}\right) \tag{4.3}$$

for small values of $\|\mathbf{f}\|_{\mathbf{H}}\,|\omega/\theta|$.

When $\mathbf{A}$ is symmetric positive definite, we have

$$\frac{\|\mathbf{r}\|_{\mathbf{A}^{-1}}}{\|\mathbf{x}\|_{\mathbf{A}}}=\|\mathbf{f}\|_{\mathbf{A}_\sigma\mathbf{A}^{-1}\mathbf{A}_\sigma}\left|\frac{\omega}{\theta}\right|\sqrt{\frac{\mu+\alpha\sigma+\frac{\alpha}{\theta}}{\sigma+\frac{1}{\theta}}}+\mathcal{O}\left(\|\mathbf{f}\|_{\mathbf{H}}^2\frac{\omega^2}{\theta^2}\right) \tag{4.4}$$

for small values of $\|\mathbf{f}\|_{\mathbf{H}}\,|\omega/\theta|$.

Proof. See section 8.1 in the appendix. ∎

In general, $\sigma+1/\theta$, defined by (4.1), is not equal to the Rayleigh quotient of $\mathbf{x}$,

$$\frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{M}\mathbf{x}}. \tag{4.5}$$

Indeed, we have

$$\sigma + \frac{1}{\theta} = \sigma + \frac{\alpha \mathbf{x}^T \mathbf{A}\mathbf{x} + \mu \mathbf{x}^T \mathbf{M}\mathbf{x}}{\alpha \mathbf{x}^T \mathbf{M}\mathbf{x} + (\mu + \alpha\sigma)\mathbf{x}^T \mathbf{M}\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{x}}. \tag{4.6}$$

However, we remark that when μ is equal to $-\alpha\sigma$ and σ is smaller than λ_1 (i.e., $\mathbf{H}$ is equal to $\mathbf{A}_\sigma$), the Rayleigh quotient (4.5) is equal to $\sigma + 1/\theta$. Section 5 studies the quality of the Rayleigh quotient (4.5). In the next subsections, we comment on accuracy bounds when measuring the residual $\mathbf{r}$ with different norms.

4.1. Accuracy bounds with H-inner product. With the relation (4.2), we can use the bounds (3.6). The general comments on bounds (3.6) still apply. In particular, quadratic convergence towards the eigenvalue is triggered as soon as

$$\left| \frac{\lambda_\gamma - \sigma}{\lambda_\gamma - \sigma - \frac{1}{\theta}} \right| \|\mathbf{f}\|_{\mathbf{H}} \left| \frac{\omega}{\theta} \right| < 1.$$

We remark that as soon as the convergence criterion (2.9) is satisfied, we can introduce the tolerance ε in these upper bounds of (3.6).

When approximating many eigenpairs with one shift σ, $|\theta|$ becomes smaller and $|\lambda_\gamma - \sigma|$ larger. Therefore, the bounds (3.6b) for the ratio between the sine of the angle and the residual norm (4.2)

$$\left| \frac{(\lambda_\sigma^+ - \sigma)(\lambda_\sigma^- - \sigma)}{\lambda_\sigma^+ - \lambda_\sigma^-} \right| |\theta| \leq \frac{|\sin \angle_{\mathbf{H}}(\mathbf{x}, \mathbf{u})|}{\|\mathbf{A}\mathbf{x} - \mathbf{M}\mathbf{x}(\sigma + \frac{1}{\theta})\|_{\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}}} \leq \left| \frac{\lambda_\gamma - \sigma}{\lambda_\gamma - \sigma - \frac{1}{\theta}} \right|$$

define a wider interval. Consequently, the bounds (3.6b) do not guarantee the same level of accuracy for the approximations close to the shift σ and the approximations far from the shift.

4.2. Accuracy bounds when M is symmetric positive definite. With the relation (4.3), we can use the bounds (3.7). We remark that $\|\mathbf{f}\|_{\mathbf{A}_\sigma \mathbf{M}^{-1} \mathbf{A}_\sigma}$ is not available as a by-product of the Lanczos reduction (2.2) and that the remainder term in the expansion comes from the computation of $\|\mathbf{x}\|_{\mathbf{M}}$.

4.3. Accuracy bounds when A is symmetric positive definite. With the relation (4.4), we can use the bounds (3.9). We remark that $\|\mathbf{f}\|_{\mathbf{A}_\sigma \mathbf{A}^{-1} \mathbf{A}_\sigma}$ is not available as a by-product of the Lanczos reduction (2.2) and that the remainder term in the expansion comes from $\|\mathbf{x}\|_{\mathbf{A}}$.

5. Study of the Ritz vector and its Rayleigh quotient. We approximate now an eigenpair $(\mathbf{u}, \lambda)$ by the Ritz vector $\mathbf{x}$ (2.7) and its Rayleigh quotient

$$\rho(\mathbf{x}) = \frac{\mathbf{x}^T \mathbf{A}\mathbf{x}}{\mathbf{x}^T \mathbf{M}\mathbf{x}}.$$

PROPOSITION 5.1. *Let* $\mathbf{x}$ *be a Ritz vector* (2.7) *produced by a shift-invert Lanczos reduction where the Lanczos vectors are* $\mathbf{H}$*-orthonormal. Then we have*

$$\rho(\mathbf{x}) = \frac{\mathbf{x}^T \mathbf{A}\mathbf{x}}{\mathbf{x}^T \mathbf{M}\mathbf{x}} = \sigma + \frac{1}{\theta} + \frac{(\mu + \sigma\alpha)\dfrac{\omega^2}{\theta^2}\mathbf{f}^T \mathbf{A}_\sigma \mathbf{f}}{1 - \alpha \dfrac{\mu + \sigma\alpha}{\mu + \sigma\alpha + \frac{\alpha}{\theta}} \dfrac{\omega^2}{\theta^2} \mathbf{f}^T \mathbf{A}_\sigma \mathbf{f}}. \tag{5.1}$$

If $\mathbf{r} = \mathbf{A}\mathbf{x} - \mathbf{M}\mathbf{x}\rho(\mathbf{x})$ *denotes the residual, then we have*

$$\frac{\|\mathbf{r}\|_{\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}}}{\|\mathbf{x}\|_{\mathbf{H}}} = \sqrt{\theta^2\left(\sigma + \frac{1}{\theta} - \rho\right)^2 + \|\mathbf{f}\|_{\mathbf{H}}^2 \frac{\omega^2}{\theta^2}\left(1 - \left(\sigma + \frac{1}{\theta} - \rho\right)\theta\right)^2}. \tag{5.2}$$

When $\mathbf{M}$ *is symmetric positive definite, we have*

$$\frac{\|\mathbf{r}\|_{\mathbf{M}^{-1}}}{\|\mathbf{x}\|_{\mathbf{M}}} = \|\mathbf{f}\|_{\mathbf{A}_\sigma\mathbf{M}^{-1}\mathbf{A}_\sigma}\left|\frac{\omega}{\theta}\right|\sqrt{\mu + \alpha\sigma + \frac{\alpha}{\theta}} + \mathcal{O}\left(\|\mathbf{f}\|_{\mathbf{H}}^2\frac{\omega^2}{\theta^2}\right) \tag{5.3}$$

for small values of $\|\mathbf{f}\|_{\mathbf{H}}\,|\omega/\theta|$.

When $\mathbf{A}$ *is symmetric positive definite, we have*

$$\frac{\|\mathbf{r}\|_{\mathbf{A}^{-1}}}{\|\mathbf{x}\|_{\mathbf{A}}} = \|\mathbf{f}\|_{\mathbf{A}_\sigma\mathbf{A}^{-1}\mathbf{A}_\sigma}\left|\frac{\omega}{\theta}\right|\sqrt{\frac{\mu + \alpha\sigma + \frac{\alpha}{\theta}}{\rho}} + \mathcal{O}\left(\|\mathbf{f}\|_{\mathbf{H}}^2\frac{\omega^2}{\theta^2}\right) \tag{5.4}$$

for small values of $\|\mathbf{f}\|_{\mathbf{H}}\,|\omega/\theta|$.

Proof. See section 8.2 in the appendix. ∎

We note that relation (5.1) does not indicate whether $\rho(\mathbf{x})$ is greater or smaller than $\sigma + 1/\theta$ without further assumptions on σ, $\mathbf{A}$, or $\mathbf{M}$. Ericsson and Ruhe [3] noticed that when the matrix $\mathbf{M}$ is symmetric positive definite and $\mathbf{H}$ is equal to $\mathbf{M}$, the Rayleigh quotient $\rho(\mathbf{x})$ is not close to $\sigma + 1/\theta$. The relation (5.1) explains their comment. Indeed, when $\mathbf{H}$ is equal to $\mathbf{M}$ ($\alpha = 0$, $\mu = 1$), we have

$$\rho(\mathbf{x}) - \left(\sigma + \frac{1}{\theta}\right) = \mathbf{f}^T\mathbf{A}_\sigma\mathbf{f}\frac{\omega^2}{\theta^2}.$$

Hence $\mathbf{f}^T\mathbf{A}_\sigma\mathbf{f}$ controls the difference between $\rho(\mathbf{x})$ and $\sigma + 1/\theta$. The eigenvalue decomposition of $(\mathbf{A}, \mathbf{M})$ implies that

$$(\lambda_1 - \sigma)\mathbf{f}^T\mathbf{M}\mathbf{f} \le \mathbf{f}^T\mathbf{A}_\sigma\mathbf{f} \le (\lambda_n - \sigma)\mathbf{f}^T\mathbf{M}\mathbf{f}. \tag{5.5}$$

Even if $\|\mathbf{f}\|_{\mathbf{M}}\,|\omega/\theta| < \varepsilon$, a tight bound on $\mathbf{f}^T\mathbf{A}_\sigma\mathbf{f}$ is, in general, not guaranteed. Therefore, the Rayleigh quotient $\rho(\mathbf{x})$ can be far from $\sigma + 1/\theta$. For the general case where $\mathbf{H} = \alpha\mathbf{A} + \mu\mathbf{M}$, relation (5.1) indicates that $\mathbf{f}^T\mathbf{A}_\sigma\mathbf{f}$ still controls the difference between $\rho(\mathbf{x})$ and $\sigma + 1/\theta$. Here we have

$$\left(\min_i \frac{\lambda_i - \sigma}{\alpha\lambda_i + \mu}\right)\mathbf{f}^T\mathbf{H}\mathbf{f} \le \mathbf{f}^T\mathbf{A}_\sigma\mathbf{f} \le \left(\max_i \frac{\lambda_i - \sigma}{\alpha\lambda_i + \mu}\right)\mathbf{f}^T\mathbf{H}\mathbf{f} \tag{5.6}$$

and, for any eigenvalue λ_i,

$$\left|\frac{\lambda_i - \sigma}{\alpha\lambda_i + \mu}\right| \le \frac{1}{|\alpha|}.$$

Consequently, the convergence criterion $\|\mathbf{f}\|_{\mathbf{H}}\,|\omega/\theta| < \varepsilon$ should result in a tighter control of $\mathbf{f}^T\mathbf{A}_\sigma\mathbf{f}$ when α is nonzero. Such a control would justify an approximation with a Ritz vector (2.7) produced by a shift-invert Lanczos reduction where the Lanczos vectors are $\mathbf{H}$-orthonormal when α is nonzero.

These expansions are valid when $\|\mathbf{f}\|_{\mathbf{H}}\,|\omega/\theta|$ is small. However, the proof does not indicate how small is small. The ratios (4.3, 4.4) and (5.3, 5.4) have similar expansions. But, for the pair $(\mathbf{x}, \rho(\mathbf{x}))$, the remainder terms come from both the residual norm and the norm of the Ritz vector. So we can expect that the expansions (5.3, 5.4) will not be valid when $\rho(\mathbf{x})$ is far from $\sigma + 1/\theta$.

5.1. Accuracy bounds with H-inner product. With the relation (5.2), we can apply the bounds (3.4) and the corresponding comments. We remark that the residual norm (5.2) is available as a by-product of the Lanczos reduction (2.2).

5.2. Accuracy bounds when M is symmetric positive definite. With the relation (5.3), we can apply the bounds (3.8) and the corresponding comments. We remark that $\|\mathbf{f}\|_{\mathbf{A}_\sigma \mathbf{M}^{-1} \mathbf{A}_\sigma}$ is not available as a by-product of the Lanczos reduction (2.2).

5.3. Accuracy bounds when A is symmetric positive definite. With the relation (5.4), we can use the bounds (3.10). We remark that $\|\mathbf{f}\|_{\mathbf{A}_\sigma \mathbf{A}^{-1} \mathbf{A}_\sigma}$ is not available as a by-product of the Lanczos reduction (2.2).

6. Study of the purified vector and its Rayleigh quotient. This section studies the case when the eigenpair is approximated by the purified vector $\mathbf{p}$ (2.8) and its Rayleigh quotient

$$\rho(\mathbf{p}) = \frac{\mathbf{p}^T \mathbf{A} \mathbf{p}}{\mathbf{p}^T \mathbf{M} \mathbf{p}}.$$

We recall that the purified vector was introduced in [3, 6].

PROPOSITION 6.1. *If* $\mathbf{p}$ *is the purified vector* (2.8) *given by the shift-invert Lanczos reduction* (2.2) *where the Lanczos vectors are* $\mathbf{H}$*-orthonormal, then*

$$\rho(\mathbf{p}) = \frac{\mathbf{p}^T \mathbf{A} \mathbf{p}}{\mathbf{p}^T \mathbf{M} \mathbf{p}} = \sigma + \frac{1}{\theta} \left(1 + \frac{\alpha}{\theta} \frac{\omega^2}{\theta^2} \frac{\mathbf{f}^T \mathbf{H} \mathbf{f} - (\mu + \alpha\sigma) \mathbf{f}^T \mathbf{M} \mathbf{f}}{\mu + \alpha\sigma + \frac{\alpha}{\theta}} \right) \eta, \tag{6.1}$$

where η *defines a denominator such that*

$$\frac{1}{\eta} = 1 + \frac{\omega^2}{\theta^2} \left(\frac{\alpha}{\theta} \frac{1}{\mu + \alpha\sigma + \frac{\alpha}{\theta}} + 1 \right) \mathbf{f}^T \mathbf{H} \mathbf{f} + \frac{\alpha^2 \omega^2}{\theta^4} \frac{\mathbf{f}^T \mathbf{M} \mathbf{f}}{\mu + \alpha\sigma + \frac{\alpha}{\theta}}.$$

If $\mathbf{r} = \mathbf{A}\mathbf{p} - \mathbf{M}\mathbf{p}\rho(\mathbf{p})$ *denotes the residual, then we have*

$$\frac{\|\mathbf{r}\|_{\mathbf{A}_\sigma^{-1} \mathbf{H} \mathbf{A}_\sigma^{-1}}}{\|\mathbf{p}\|_{\mathbf{H}}} = \|\mathbf{M}\mathbf{f}\|_{\mathbf{A}_\sigma^{-1} \mathbf{H} \mathbf{A}_\sigma^{-1}} \frac{|\omega|}{\theta^2} + \mathcal{O}\left(\|\mathbf{f}\|_{\mathbf{H}}^2 \frac{\omega^2}{\theta^2} \right) \tag{6.2}$$

for small values of $\|\mathbf{f}\|_{\mathbf{H}} |\omega/\theta|$.

When $\mathbf{M}$ *is symmetric positive definite, we have*

$$\frac{\|\mathbf{r}\|_{\mathbf{M}^{-1}}}{\|\mathbf{p}\|_{\mathbf{M}}} = \|\mathbf{f}\|_{\mathbf{M}} \frac{|\omega|}{\theta^2} \sqrt{\mu + \alpha\sigma + \frac{\alpha}{\theta}} + \mathcal{O}\left(\|\mathbf{f}\|_{\mathbf{H}}^2 \frac{\omega^2}{\theta^2} \right) \tag{6.3}$$

for small values of $\|\mathbf{f}\|_{\mathbf{H}} |\omega/\theta|$.

When $\mathbf{A}$ *is symmetric positive definite, we have*

$$\frac{\|\mathbf{r}\|_{\mathbf{A}^{-1}}}{\|\mathbf{p}\|_{\mathbf{A}}} = \|\mathbf{M}\mathbf{f}\|_{\mathbf{A}^{-1}} \frac{|\omega|}{\theta^2} \sqrt{\frac{\mu + \alpha\sigma + \frac{\alpha}{\theta}}{\rho(\mathbf{p})}} + \mathcal{O}\left(\|\mathbf{f}\|_{\mathbf{H}}^2 \frac{\omega^2}{\theta^2} \right) \tag{6.4}$$

for small values of $\|\mathbf{f}\|_{\mathbf{H}} |\omega/\theta|$.

Proof. See section 8.3 in the appendix. ∎

We note that relation (6.1) does not indicate whether $\rho(\mathbf{p})$ is greater or smaller than $\sigma + 1/\theta$ without further assumptions on σ, $\mathbf{A}$, or $\mathbf{M}$.

6.1. Accuracy bounds with H-inner product. With the relation (6.2), we can use the bounds (3.4).

We remark that the norm

$$\|\mathbf{Mf}\|_{\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}} = \|\mathbf{A}_\sigma^{-1}\mathbf{Mf}\|_{\mathbf{H}}$$

is not directly available as a by-product of the Lanczos reduction (2.2). But performing the next Lanczos iteration requires computing $\mathbf{A}_\sigma^{-1}\mathbf{Mf}$, of which we could compute the $\mathbf{H}$-norm.

6.2. Accuracy bounds when M is symmetric positive definite. Ericsson and Ruhe [3] proved

$$\frac{\mathbf{p}^T\mathbf{Ap}}{\mathbf{p}^T\mathbf{Mp}} = \sigma + \frac{1}{\theta}\frac{1}{1+\|\mathbf{f}\|_{\mathbf{M}}^2\frac{\omega^2}{\theta^2}}, \tag{6.5a}$$

$$\frac{\|\mathbf{r}\|_{\mathbf{M}^{-1}}}{\|\mathbf{p}\|_{\mathbf{M}}} = \|\mathbf{f}\|_{\mathbf{M}}\frac{|\omega|}{\theta^2}\frac{1}{1+\|\mathbf{f}\|_{\mathbf{M}}^2\frac{\omega^2}{\theta^2}} \tag{6.5b}$$

when the Lanczos vectors are $\mathbf{M}$-orthonormal. We recover their results when α is 0 and μ is 1.

We remark that $\|\mathbf{f}\|_{\mathbf{M}}$ is available when $\mathbf{H} = \mathbf{M}$. For the general case where α is nonzero, the norm $\|\mathbf{f}\|_{\mathbf{M}}$ is not available as a by-product of the Lanczos reduction (2.2). With relation (6.3), we can apply the bounds (3.8).

6.3. Accuracy bounds when A is symmetric positive definite. With the relation (6.4), we can apply the bounds (3.10) and the corresponding comments. In general, the norm $\|\mathbf{Mf}\|_{\mathbf{A}^{-1}}$ is not available as a by-product of the Lanczos reduction (2.2).

7. Conclusions. This paper analyzes the accuracy of the shift-invert Lanczos iteration, given a stopping criterion, for computing eigenpairs of the symmetric definite generalized eigenvalue problem. We provide bounds for the accuracy of the eigenpairs produced by shift-invert Lanczos for a "normalized" residual norm induced by different inner products. We point out, however, that we do not discuss the number of Lanczos iterations required to achieve the stopping criterion (see [7] for a discussion).

During the remainder of this final section, we discuss some implications of our analysis and draw some practical recommendations for shift-invert Lanczos iterations. Finally, when the generalized eigenvalue problem arises from a conforming finite element method, we also comment on the uniform accuracy of bounds (independent of the mesh size h).

7.1. Practical shift-invert Lanczos iterations. To evaluate the accuracy of a pair $(\mathbf{y}, \tilde{\lambda})$, we have measured the associated residual in different norms. We now provide some comments on selecting the appropriate residual norm. For the sake of simplicity, we assume that $\mathbf{A}$ and $\mathbf{M}$ are symmetric positive definite and that the set of eigenvectors $\{\mathbf{u}_i\}_{i=1}^n$ is $\mathbf{M}$-orthonormal. When the vector $\mathbf{y}$ satisfies

$$\mathbf{y} = \sum_i \mathbf{u}_i\psi_i \qquad (\psi_i = \mathbf{y}^T\mathbf{Mu}_i)$$

and the residual is $\mathbf{r} = \mathbf{A}\mathbf{y} - \mathbf{M}\mathbf{y}\tilde{\lambda}$, we have

$$\frac{\|\mathbf{r}\|^2_{\mathbf{M}^{-1}}}{\|\mathbf{y}\|^2_{\mathbf{M}}} = \sum_i \frac{\psi_i^2}{\sum_j \psi_j^2}(\lambda_i - \tilde{\lambda})^2, \tag{7.1a}$$

$$\frac{\|\mathbf{r}\|^2_{\mathbf{A}^{-1}}}{\|\mathbf{y}\|^2_{\mathbf{A}}} = \sum_i \frac{\lambda_i \psi_i^2}{\sum_j \lambda_j \psi_j^2}\left(\frac{\lambda_i - \tilde{\lambda}}{\lambda_i}\right)^2, \tag{7.1b}$$

$$\frac{\|\mathbf{r}\|^2_{\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}}}{\|\mathbf{y}\|^2_{\mathbf{H}}} = \sum_i \frac{(\alpha\lambda_i + \mu)\psi_i^2}{\sum_j (\alpha\lambda_j + \mu)\psi_j^2}\left(\frac{\lambda_i - \tilde{\lambda}}{\lambda_i - \sigma}\right)^2. \tag{7.1c}$$

When λ_n is large and is not an eigenvalue of interest, we note that the $\mathbf{M}^{-1}$-norm has the largest weights, while the weights associated with the two other norms are bounded. Moreover, when the spread $\lambda_n - \lambda_1$ is large, the lower bound in (3.8b) becomes crude. Consequently, we recommend to measure the residual with a norm involving $\mathbf{A}_\sigma^{-1}$ or $\mathbf{A}^{-1}$.

Our analysis has indicated that a Ritz vector (2.7) and its Rayleigh quotient can produce a good approximation to an eigenpair of $(\mathbf{A}, \mathbf{M})$. When the Lanczos vectors are $\mathbf{H}$-orthonormal, the size of $\mathbf{f}^T\mathbf{A}_\sigma\mathbf{f}$ controls the quality of such an approximation. When α is nonzero, relation (5.6) combined with the convergence criterion (2.9) results in a tight control of $\mathbf{f}^T\mathbf{A}_\sigma\mathbf{f}$ and, hence, in a good approximation to an eigenpair.

In practice, the purified vector is often used to approximate an eigenvector. When $\mathbf{M}$ is positive definite and $\mathbf{H}$ is equal to $\mathbf{M}$, Ericsson and Ruhe [3] justified this choice. For the general case where the Lanczos vectors are $\mathbf{H}$-orthonormal (α is nonzero), our analysis does not indicate whether the purified vector is a better choice than the Ritz vector. However, we notice that the purified vector uses all the information available from the Krylov subspace spanned by $\mathbf{V}$ and the vector $\mathbf{f}$ while the Ritz vector belongs only to the Krylov subspace.

We now compare formally the right-hand sides of (4.2), (5.2), and (6.2). For (4.2), the right-hand side is

$$\|\mathbf{f}\|_{\mathbf{H}} \left|\frac{\omega}{\theta}\right|.$$

For small values of $\|\mathbf{f}\|_{\mathbf{H}} |\omega/\theta|$, the right-hand side of (5.2) is equal to

$$\|\mathbf{f}\|_{\mathbf{H}} \left|\frac{\omega}{\theta}\right| + \mathcal{O}\left(\|\mathbf{f}\|^2_{\mathbf{H}} \frac{\omega^2}{\theta^2}\right).$$

Hence, asymptotically, (4.2) and (5.2) are equivalent. The right-hand side of (6.2) employs a different norm and contains an additional factor of $1/\theta$. From (6.1), we have that

$$\frac{1}{\theta} = \rho(\mathbf{p}) - \sigma + \mathcal{O}\left(\|\mathbf{f}\|^2_{\mathbf{H}} \frac{\omega^2}{\theta^2}\right).$$

The equation $\mathbf{A}_\sigma\mathbf{u}_i = \mathbf{M}\mathbf{u}_i(\lambda_i - \sigma)$ implies that

$$\min_{\lambda_i} \frac{1}{|\lambda_i - \sigma|}\|\mathbf{y}\|_{\mathbf{H}} \le \|\mathbf{M}\mathbf{y}\|_{\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}} \le \max_{\lambda_i} \frac{1}{|\lambda_i - \sigma|}\|\mathbf{y}\|_{\mathbf{H}}$$

for all $\mathbf{y} \in \mathbb{R}^n$ and in particular for $\mathbf{f}$. Therefore, up to high order terms $\mathcal{O}\left(\|\mathbf{f}\|^2_{\mathbf{H}} \frac{\omega^2}{\theta^2}\right)$, we obtain

$$\min_{\lambda_i} \left|\frac{\rho(\mathbf{p}) - \sigma}{\lambda_i - \sigma}\right| \|\mathbf{f}\|_{\mathbf{H}} \left|\frac{\omega}{\theta}\right| \le \|\mathbf{M}\mathbf{f}\|_{\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}} \frac{|\omega|}{\theta^2} \le \max_{\lambda_i} \left|\frac{\rho(\mathbf{p}) - \sigma}{\lambda_i - \sigma}\right| \|\mathbf{f}\|_{\mathbf{H}} \left|\frac{\omega}{\theta}\right|.$$

Consequently, the right-hand sides of (4.2), (5.2), and (6.2) are equivalent. A similar analysis can be carried out when measuring the residuals with the $\mathbf{M}^{-1}$-norm or the $\mathbf{A}^{-1}$-norm.

We also comment that a common practice among many structural analysts is to set the shift σ to the left of the smallest eigenvalue λ_1 and use a preconditioned iteration to approximate the application of $\mathbf{A}_\sigma^{-1}$. If we neglect the approximation error due to the preconditioned iteration, then when approximating many eigenpairs from the low end of the spectrum, the bound (3.4a) guarantees the same level of relative error on the eigenvalue. On the other hand, relation (3.4b) cannot provide the same level of accuracy on the angle, because the constant

$$\left| \frac{\lambda_\gamma - \sigma}{\lambda_\gamma - \tilde{\lambda}} \right|$$

increases with $|\tilde{\lambda} - \sigma|$ as described in the discussion following Proposition 3.2.

When using purified vectors as approximation of eigenvectors, the practitioner should be aware of possible departure from orthogonality. The paper [1] shows that the tolerance for the eigensolver can be set at large values, say 10^{-3}–10^{-5}. When using large values for ε, two purified vectors $\mathbf{p}$ and $\mathbf{q}$,

$$\mathbf{p} = \mathbf{V}\mathbf{s}_\mathbf{p} + \mathbf{f}\frac{\mathbf{e}^T\mathbf{s}_\mathbf{p}}{\theta_\mathbf{p}} \quad \text{and} \quad \mathbf{q} = \mathbf{V}\mathbf{s}_\mathbf{q} + \mathbf{f}\frac{\mathbf{e}^T\mathbf{s}_\mathbf{q}}{\theta_\mathbf{q}}$$

such that

$$\|\mathbf{f}\|_\mathbf{H} \left| \frac{\mathbf{e}^T\mathbf{s}_\mathbf{p}}{\theta_\mathbf{p}} \right| \leq \varepsilon \quad \text{and} \quad \|\mathbf{f}\|_\mathbf{H} \left| \frac{\mathbf{e}^T\mathbf{s}_\mathbf{q}}{\theta_\mathbf{q}} \right| \leq \varepsilon,$$

may not be $\mathbf{H}$-orthogonal to working precision because

$$\left|\mathbf{p}^T\mathbf{H}\mathbf{q}\right| = \|\mathbf{f}\|_\mathbf{H}^2 \left| \frac{\mathbf{e}^T\mathbf{s}_\mathbf{p}}{\theta_\mathbf{p}} \right| \left| \frac{\mathbf{e}^T\mathbf{s}_\mathbf{q}}{\theta_\mathbf{q}} \right| \leq \varepsilon^2.$$

One easy solution is to perform a Rayleigh–Ritz analysis for the pencil $(\mathbf{A}, \mathbf{M})$ and the space spanned by $\mathbf{V}$ and $\mathbf{f}$. This extra Rayleigh–Ritz step will restore the $\mathbf{M}$-orthogonality and improve the approximation of eigenpairs. Such a postprocessing step would require the construction of the projected matrices

$$\begin{pmatrix} \mathbf{V}^T\mathbf{A}\mathbf{V} & \mathbf{V}^T\mathbf{A}\mathbf{f} \\ \mathbf{f}^T\mathbf{A}\mathbf{V} & \mathbf{f}^T\mathbf{A}\mathbf{f} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \mathbf{V}^T\mathbf{M}\mathbf{V} & \mathbf{V}^T\mathbf{M}\mathbf{f} \\ \mathbf{f}^T\mathbf{M}\mathbf{V} & \mathbf{f}^T\mathbf{M}\mathbf{f} \end{pmatrix}.$$

When $\mathbf{M}$ is symmetric definite positive and $\mathbf{H}$ is equal to $\mathbf{M}$, the construction of these projected matrices is described in [6] and does not require extra operations with $\mathbf{A}$ nor $\mathbf{M}$. When the shift σ is smaller than λ_1 and $\mathbf{H}$ is equal to $\mathbf{A}_\sigma$, the projected matrices are available as by-products of the Lanczos reduction. For a general matrix $\mathbf{H}$ and an arbitrary shift σ, we can project the matrix $\mathbf{A}$ (or $\mathbf{M}$) onto the space spanned by $\mathbf{V}$ and $\mathbf{f}$ and then use the $\mathbf{H}$-orthogonality, i.e.,

$$\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{f}^T\mathbf{H}\mathbf{f} \end{pmatrix} = \alpha \begin{pmatrix} \mathbf{V}^T\mathbf{A}\mathbf{V} & \mathbf{V}^T\mathbf{A}\mathbf{f} \\ \mathbf{f}^T\mathbf{A}\mathbf{V} & \mathbf{f}^T\mathbf{A}\mathbf{f} \end{pmatrix} + \mu \begin{pmatrix} \mathbf{V}^T\mathbf{M}\mathbf{V} & \mathbf{V}^T\mathbf{M}\mathbf{f} \\ \mathbf{f}^T\mathbf{M}\mathbf{V} & \mathbf{f}^T\mathbf{M}\mathbf{f} \end{pmatrix}.$$

The projection step would incur extra matrix-vector products. This additional cost will restore the $\mathbf{M}$-orthogonality of the vectors and it could be cheaper than performing further shift-invert Lanczos iterations (i.e., further linear solves with $\mathbf{A}_\sigma$).

7.2. Example of a finite-element based eigenproblem. Here we assume that the problem (1.1) arises from the discretization of an elliptic partial differential equation with a conforming finite element method; we could write (1.1) as

$$\mathbf{A}^h \mathbf{u}^h = \mathbf{M}^h \mathbf{u}^h \lambda^h, \tag{7.2}$$

where h is the characteristic mesh size. Our error bounds need to be uniform with respect to this mesh size. Theorem 3.1 uses the largest and smallest eigenvalues of $\hat{\mathbf{A}}$. So, for the sake of completeness, we recall a standard result from finite element theory [2]:

$$\lim_{h\to 0} \lambda_1^h = \lambda_1^\star, \qquad \lim_{h\to 0} \lambda_n^h = +\infty, \tag{7.3}$$

where $\lambda_1^\star$ is the smallest eigenvalue of the differential eigenvalue problem. Note that $n \to +\infty$ as $h \to 0$.

The discussion in section 3 implies that the lower bound of (3.8b) loses its sharpness when the mesh size is refined. Consequently, when measuring the residual with the $\mathbf{M}^{-1}$-norm, the error on the angle can be much smaller than the upper bound in (3.7b, 3.8b), which could result in more iterations than necessary. On the other hand, the constants in the bounds (3.6b) and (3.10b) are invariant with respect to the mesh size.

The previous subsection demonstrated, via a formal analysis, that the right-hand sides of (4.2), (5.2), and (6.2) are equivalent. We also remarked that the left- and right-hand sides of (4.2) use different norms in contrast to (6.2). If we assume that (7.2) arises from a continuous eigenvalue problem posed in the function space $H_0^1(\Omega)$, then we can identify the $\mathbf{H}$-norm as a discrete $H_0^1(\Omega)$-norm. In a similar fashion, the $(\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1})$-norm may be identified as a discrete norm on the dual space $H^{-1}(\Omega)$. Scaled by a factor $1/|\theta|$, the quantity

$$\frac{1}{|\theta|}\|\mathbf{M}\mathbf{f}\|_{\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}}$$

is now equivalent to the discrete $H_0^1(\Omega)$-norm of $\mathbf{f}$. Hence, the residual $\mathbf{r}$ is measured in a discrete $H^{-1}(\Omega)$-norm while the right-hand side amounts to a discrete $H_0^1(\Omega)$ norm of $\mathbf{f}$.

As we explained in section 5, Ericsson and Ruhe [3] noticed that when $\mathbf{M}$ is symmetric positive definite and $\mathbf{H}$ is equal to $\mathbf{M}$, the Rayleigh quotient of $\mathbf{x}$ is not close to $\sigma + 1/\theta$. The following example demonstrates that the Rayleigh quotient can differ significantly from $\sigma + 1/\theta$ and that this difference can grow when the mesh is refined. Consider the $(2m-1)\times(2m-1)$ tridiagonal matrices

$$\mathbf{A}^h = 2m \begin{pmatrix} 2 & -1 & \cdots & 0 \\ -1 & 2 & \cdots & 0 \\ \vdots & & \ddots & -1 \\ 0 & \cdots & -1 & 2 \end{pmatrix}, \qquad \mathbf{M}^h = \frac{1}{12m} \begin{pmatrix} 4 & 1 & \cdots & 0 \\ 1 & 4 & \cdots & 0 \\ \vdots & & \ddots & 1 \\ 0 & \cdots & 1 & 4 \end{pmatrix}$$

that arise from a uniform finite element discretization of the Laplace equation with homogeneous Dirichlet boundary conditions on the unit interval. A shift-invert Lanczos iteration with an $\mathbf{M}^h$-inner product is used to approximate the smallest eigenpair. The starting vector is $\mathbf{M}^h$-normalized and proportional to $\mathbf{e}_m$. The shift σ is set to 0. The Ritz vector $\mathbf{x}$ is obtained as soon as the stopping criterion (2.9) is satisfied

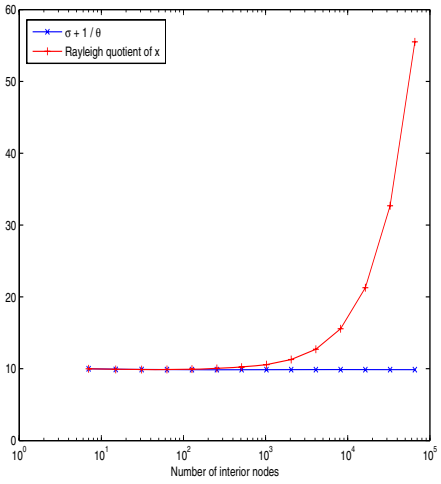


FIG. 7.1. *Size of the Rayleigh quotient with mesh refinement* ($\varepsilon = 10^{-2}$) *using an* $\mathbf{M}^h$*-orthonormal shift-invert Lanczos method.*

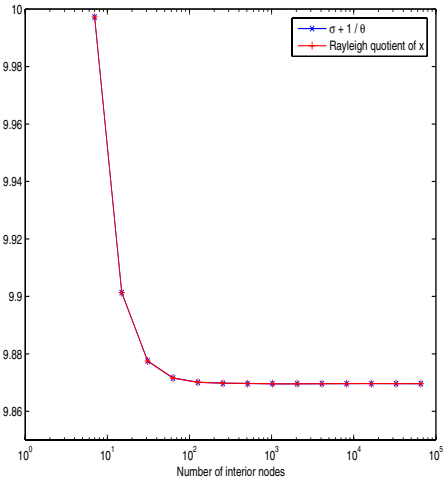


FIG. 7.2. *Size of the Rayleigh quotient with mesh refinement* ($\varepsilon = 10^{-2}$) *using an* $(\mathbf{A}^h + \mathbf{M}^h)$*-orthonormal shift-invert Lanczos method.*

with $\varepsilon = 10^{-2}$. Figure 7.1 demonstrates that the Rayleigh quotient of $\mathbf{x}$ for the pencil $(\mathbf{A}^h, \mathbf{M}^h)$ differs from $\sigma + 1/\theta$ and that the difference grows as we refine the mesh. This difference decreases with ε but disappears only when $\varepsilon = 0$. This behavior is well explained by relation (5.5). On the other hand, when building an $(\mathbf{A}^h + \mathbf{M}^h)$-orthonormal Lanczos basis for the same test problem, Figure 7.2 demonstrates that the Rayleigh quotient of $\mathbf{x}$ for the pencil $(\mathbf{A}^h, \mathbf{M}^h)$ does not differ significantly from $\sigma+1/\theta$ and independently of the mesh size. This behavior is well explained by relation (5.6).

8. Appendix.

8.1. Proof for Proposition 4.1. The Ritz vector $\mathbf{x}$ is $\mathbf{H}$-normalized. Equation (2.6) implies $\mathbf{A}_\sigma^{-1}(\mathbf{M}\mathbf{x} - \mathbf{A}_\sigma \mathbf{x}\theta) = \mathbf{f}\omega$ so that

$$\mathbf{A}_\sigma^{-1}\left(\mathbf{A}\mathbf{x} - \mathbf{M}\mathbf{x}\left(\sigma + \frac{1}{\theta}\right)\right) = -\mathbf{f}\frac{\omega}{\theta} \tag{8.1}$$

which proves the relation (4.2).

To compute the norms of the residual, we use (8.1) and obtain easily the leading term with the norm of $\mathbf{f}$. To finish the proof, we need to compute $\mathbf{x}^T\mathbf{M}\mathbf{x}$ and $\mathbf{x}^T\mathbf{A}\mathbf{x}$.

Using the definition of $\mathbf{H}$, we have

$$\alpha \mathbf{x}^T\mathbf{M}\mathbf{x} = \mathbf{x}^T\mathbf{H}\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{x} - (\mu + \alpha\sigma)\mathbf{x}^T\mathbf{M}\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{x}.$$

From (2.6), we get

$$\mathbf{x}^T\mathbf{M}\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{x} = \theta\mathbf{x}^T\mathbf{M}\mathbf{x} + \mathbf{x}^T\mathbf{M}\mathbf{f}\omega.$$

Combining the last two relations with the definition of θ, $\theta = \mathbf{x}^T\mathbf{H}\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{x}$, we obtain

$$\mathbf{x}^T\mathbf{M}\mathbf{x} = \frac{1}{\mu + \alpha\sigma + \frac{\alpha}{\theta}}\left(1 - (\mu + \alpha\sigma)\mathbf{x}^T\mathbf{M}\mathbf{f}\frac{\omega}{\theta}\right). \tag{8.2}$$

To compute $\mathbf{x}^T\mathbf{M}\mathbf{f}$, we follow the same steps

$$\alpha \mathbf{f}^T\mathbf{M}\mathbf{x} = \mathbf{f}^T\mathbf{H}\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{x} - (\mu + \alpha\sigma)\mathbf{f}^T\mathbf{M}\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{x}.$$

From (2.6), we have

$$\left\{\begin{array}{rcl} \mathbf{f}^T\mathbf{H}\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{x} &=& \mathbf{f}^T\mathbf{H}\mathbf{f}\omega, \\ \mathbf{f}^T\mathbf{M}\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{x} &=& \mathbf{f}^T\mathbf{M}\mathbf{x}\theta + \mathbf{f}^T\mathbf{M}\mathbf{f}\omega. \end{array}\right.$$

Finally, we obtain

$$\mathbf{x}^T\mathbf{M}\mathbf{f} = \frac{\omega}{\theta}\frac{\mathbf{f}^T\mathbf{H}\mathbf{f} - (\mu + \alpha\sigma)\mathbf{f}^T\mathbf{M}\mathbf{f}}{\mu + \alpha\sigma + \frac{\alpha}{\theta}} = \frac{\omega}{\theta}\frac{\alpha\mathbf{f}^T\mathbf{A}_\sigma\mathbf{f}}{\mu + \alpha\sigma + \frac{\alpha}{\theta}} \tag{8.3}$$

which proves

$$\mathbf{x}^T\mathbf{M}\mathbf{x} = \frac{1}{\mu + \alpha\sigma + \frac{\alpha}{\theta}} + \mathcal{O}\left(\|\mathbf{f}\|_{\mathbf{H}}^2\frac{\omega^2}{\theta^2}\right).$$

To compute $\mathbf{x}^T\mathbf{A}\mathbf{x}$, we start by using the $\mathbf{H}$-norm of $\mathbf{x}$,

$$\begin{aligned} \alpha\mathbf{x}^T\mathbf{A}\mathbf{x} + \mu\mathbf{x}^T\mathbf{M}\mathbf{x} &= 1 \\ \alpha\mathbf{x}^T\mathbf{A}\mathbf{x} &= \frac{\alpha\sigma + \frac{\alpha}{\theta}}{\mu + \alpha\sigma + \frac{\alpha}{\theta}} + \mathcal{O}\left(\|\mathbf{f}\|_{\mathbf{H}}^2\frac{\omega^2}{\theta^2}\right). \end{aligned}$$

By continuity, we can extend the formula to the case where α is 0 (which would require that $\mathbf{M}$ is positive definite).

8.2. Proof for Proposition 5.1. We start by writing the **H**-norm of **x**.

$$\begin{aligned}
\alpha \mathbf{x}^T\mathbf{A}\mathbf{x} + \mu \mathbf{x}^T\mathbf{M}\mathbf{x} &= 1,\\
\frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{M}\mathbf{x}} &= \frac{1}{\alpha \mathbf{x}^T\mathbf{M}\mathbf{x}} - \frac{\mu}{\alpha},\\
\frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{M}\mathbf{x}} &= \frac{\mu + \alpha\sigma + \frac{\alpha}{\theta}}{\alpha\left(1 - \alpha\frac{\mu+\alpha\sigma}{\mu+\alpha\sigma+\frac{\alpha}{\theta}}\frac{\omega^2}{\theta^2}\mathbf{f}^T\mathbf{A}_\sigma\mathbf{f}\right)} - \frac{\mu}{\alpha},\\
\frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{M}\mathbf{x}} &= \frac{\sigma + \frac{1}{\theta} + \mu\frac{\mu+\alpha\sigma}{\mu+\alpha\sigma+\frac{\alpha}{\theta}}\frac{\omega^2}{\theta^2}\mathbf{f}^T\mathbf{A}_\sigma\mathbf{f}}{1 - \alpha\frac{\mu+\alpha\sigma}{\mu+\alpha\sigma+\frac{\alpha}{\theta}}\frac{\omega^2}{\theta^2}\mathbf{f}^T\mathbf{A}_\sigma\mathbf{f}},\\
\frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{M}\mathbf{x}} &= \sigma + \frac{1}{\theta} + \frac{(\mu+\sigma\alpha)\frac{\omega^2}{\theta^2}\mathbf{f}^T\mathbf{A}_\sigma\mathbf{f}}{1 - \alpha\frac{\mu+\sigma\alpha}{\mu+\sigma\alpha+\frac{\alpha}{\theta}}\frac{\omega^2}{\theta^2}\mathbf{f}^T\mathbf{A}_\sigma\mathbf{f}}.
\end{aligned}$$

We compute now the norm of the residual $\mathbf{r} = \mathbf{A}\mathbf{x} - \mathbf{M}\mathbf{x}\rho(\mathbf{x})$. We have

$$\begin{aligned}
\mathbf{r} &= \mathbf{A}\mathbf{x} - \mathbf{M}\mathbf{x}\left(\sigma + \frac{1}{\theta}\right) + \mathbf{M}\mathbf{x}\left(\sigma + \frac{1}{\theta} - \rho(\mathbf{x})\right),\\
\mathbf{r} &= -\mathbf{A}_\sigma\mathbf{f}\frac{\omega}{\theta} + \mathbf{M}\mathbf{x}\left(\sigma + \frac{1}{\theta} - \rho(\mathbf{x})\right),\\
\mathbf{A}_\sigma^{-1}\mathbf{r} &= -\mathbf{f}\frac{\omega}{\theta} + \mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{x}\left(\sigma + \frac{1}{\theta} - \rho(\mathbf{x})\right),\\
\mathbf{A}_\sigma^{-1}\mathbf{r} &= \mathbf{f}\frac{\omega}{\theta}\left(\left(\sigma + \frac{1}{\theta} - \rho(\mathbf{x})\right)\theta - 1\right) + \mathbf{x}\left(\sigma + \frac{1}{\theta} - \rho(\mathbf{x})\right)\theta.
\end{aligned}$$

We conclude by using the **H**-normalization of **x** and the **H**-orthogonality between **f** and **x**.

Next, we consider the case where **M** is symmetric positive definite. We have

$$\mathbf{r}^T\mathbf{M}^{-1}\mathbf{r} = \mathbf{f}^T\mathbf{A}_\sigma\mathbf{M}^{-1}\mathbf{A}_\sigma\mathbf{f}\frac{\omega^2}{\theta^2} - 2\mathbf{f}^T\mathbf{A}_\sigma\mathbf{x}\left(\sigma + \frac{1}{\theta} - \rho(\mathbf{x})\right)\frac{\omega}{\theta} + \mathbf{x}^T\mathbf{M}\mathbf{x}\left(\sigma + \frac{1}{\theta} - \rho(\mathbf{x})\right)^2.$$

The relations

$$\left\{\begin{array}{l}
\left(\min\limits_i \dfrac{(\lambda_i - \sigma)^2}{\alpha\lambda_i + \mu}\right)\|\mathbf{f}\|_{\mathbf{H}}^2 \leq \|\mathbf{f}\|_{\mathbf{A}_\sigma\mathbf{M}^{-1}\mathbf{A}_\sigma}^2 \leq \left(\max\limits_i \dfrac{(\lambda_i - \sigma)^2}{\alpha\lambda_i + \mu}\right)\|\mathbf{f}\|_{\mathbf{H}}^2,\\
\mathbf{x}^T\mathbf{A}_\sigma\mathbf{f}\dfrac{\omega}{\theta} = \left(\sigma + \dfrac{1}{\theta} - \rho(\mathbf{x})\right)\mathbf{x}^T\mathbf{M}\mathbf{x} = \mathcal{O}\left(\|\mathbf{f}\|_{\mathbf{H}}^2\dfrac{\omega^2}{\theta^2}\right),\\
\sigma + \dfrac{1}{\theta} - \rho(\mathbf{x}) = \mathcal{O}\left(\|\mathbf{f}\|_{\mathbf{H}}^2\dfrac{\omega^2}{\theta^2}\right),\\
\mathbf{x}^T\mathbf{M}\mathbf{x} = \dfrac{1}{\mu + \alpha\sigma + \frac{\alpha}{\theta}} + \mathcal{O}\left(\|\mathbf{f}\|_{\mathbf{H}}^2\dfrac{\omega^2}{\theta^2}\right)
\end{array}\right.$$

establish

$$\left\{\begin{array}{rcl}
\|\mathbf{r}\|_{\mathbf{M}^{-1}} &=& \|\mathbf{f}\|_{\mathbf{A}_\sigma\mathbf{M}^{-1}\mathbf{A}_\sigma}\left|\frac{\omega}{\theta}\right| + \mathcal{O}\left(\|\mathbf{f}\|_{\mathbf{H}}^2\frac{\omega^2}{\theta^2}\right),\\
\|\mathbf{x}\|_{\mathbf{M}} &=& \frac{1}{\sqrt{\mu + \alpha\sigma + \frac{\alpha}{\theta}}} + \mathcal{O}(\|\mathbf{f}\|_{\mathbf{H}}^2\frac{\omega^2}{\theta^2}),
\end{array}\right.$$

which proves (5.3).

Finally, let us assume $\mathbf{A}$ is symmetric positive definite. We have

$$\mathbf{r}^T\mathbf{A}^{-1}\mathbf{r} = \mathbf{f}^T\mathbf{A}_\sigma\mathbf{A}^{-1}\mathbf{A}_\sigma\mathbf{f}\frac{\omega^2}{\theta^2} - 2\mathbf{f}^T\mathbf{A}_\sigma\mathbf{A}^{-1}\mathbf{M}\mathbf{x}\left(\sigma + \frac{1}{\theta} - \rho(\mathbf{x})\right)\frac{\omega}{\theta} + \mathbf{x}^T\mathbf{M}\mathbf{A}^{-1}\mathbf{M}\mathbf{x}\left(\sigma + \frac{1}{\theta} - \rho(\mathbf{x})\right)^2.$$

The relations

$$\begin{cases} \left(\min_i \dfrac{(\lambda_i-\sigma)^2}{\lambda_i(\alpha\lambda_i+\mu)}\right)\|\mathbf{f}\|_{\mathbf{H}}^2 \le \|\mathbf{f}\|_{\mathbf{A}_\sigma\mathbf{A}^{-1}\mathbf{A}_\sigma}^2 \le \left(\max_i \dfrac{(\lambda_i-\sigma)^2}{\lambda_i(\alpha\lambda_i+\mu)}\right)\|\mathbf{f}\|_{\mathbf{H}}^2, \\ \mathbf{x}^T\mathbf{M}\mathbf{A}^{-1}\mathbf{A}_\sigma\mathbf{f} = \mathbf{x}^T\mathbf{M}\mathbf{f} - \sigma\mathbf{x}^T\mathbf{M}\mathbf{A}^{-1}\mathbf{M}\mathbf{f}, \\ \mathbf{x}^T\mathbf{M}\mathbf{A}^{-1}\mathbf{M}\mathbf{f} = \mathbf{x}^T\mathbf{M}\mathbf{f}\theta - \mathbf{x}\mathbf{M}\mathbf{A}^{-1}\mathbf{M}\mathbf{f}\sigma\theta + \mathbf{f}^T\mathbf{M}\mathbf{f}\omega - \mathbf{f}^T\mathbf{M}\mathbf{A}^{-1}\mathbf{M}\mathbf{f}\sigma\omega, \\ \sigma + \dfrac{1}{\theta} - \rho(\mathbf{x}) = \mathcal{O}\left(\|\mathbf{f}\|_{\mathbf{H}}^2\dfrac{\omega^2}{\theta^2}\right) \end{cases}$$

establish

$$\begin{cases} \|\mathbf{r}\|_{\mathbf{A}^{-1}} &= \|\mathbf{f}\|_{\mathbf{A}_\sigma\mathbf{A}^{-1}\mathbf{A}_\sigma}\left|\frac{\omega}{\theta}\right| + \mathcal{O}\left(\|\mathbf{f}\|_{\mathbf{H}}^2\frac{\omega^2}{\theta^2}\right), \\ \|\mathbf{x}\|_{\mathbf{A}} &= \sqrt{\frac{\rho(\mathbf{x})}{\mu+\alpha\sigma+\frac{\alpha}{\theta}}} + \mathcal{O}(\|\mathbf{f}\|_{\mathbf{H}}^2\frac{\omega^2}{\theta^2}), \end{cases}$$

which proves (5.4).

8.3. Proof for Proposition 6.1. We start by using the definitions of $\rho(\mathbf{p})$ and $\mathbf{A}_\sigma$,

$$\rho(\mathbf{p}) = \sigma + \frac{\mathbf{p}^T\mathbf{A}_\sigma\mathbf{p}}{\mathbf{p}^T\mathbf{M}\mathbf{p}}.$$

From (2.6), we get

$$\mathbf{p}^T\mathbf{M}\mathbf{x} = \mathbf{p}^T\mathbf{A}_\sigma\mathbf{p}\theta.$$

Consequently, $\rho(\mathbf{p})$ satisfies

$$\rho(\mathbf{x}) = \sigma + \frac{1}{\theta}\frac{1}{1+\frac{\omega}{\theta}\frac{\mathbf{p}^T\mathbf{M}\mathbf{f}}{\mathbf{p}^T\mathbf{M}\mathbf{x}}}. \tag{8.4}$$

By definition of $\mathbf{p}$ (2.8), we have

$$\begin{cases} \mathbf{p}^T\mathbf{M}\mathbf{f} &= \mathbf{x}^T\mathbf{M}\mathbf{f} + \mathbf{f}^T\mathbf{M}\mathbf{f}\dfrac{\omega}{\theta}, \\ \mathbf{p}^T\mathbf{M}\mathbf{x} &= \mathbf{x}^T\mathbf{M}\mathbf{x} + \mathbf{x}^T\mathbf{M}\mathbf{f}\dfrac{\omega}{\theta}. \end{cases}$$

Now, we introduce (8.2) and (8.3)

$$\mathbf{p}^T\mathbf{M}\mathbf{f} = \frac{\omega}{\theta}\frac{\mathbf{f}^T\mathbf{H}\mathbf{f} + \frac{\alpha}{\theta}\mathbf{f}^T\mathbf{M}\mathbf{f}}{\mu+\alpha\sigma+\frac{\alpha}{\theta}}, \tag{8.5a}$$

$$\mathbf{p}^T\mathbf{M}\mathbf{x} = \frac{1}{\mu+\alpha\sigma+\frac{\alpha}{\theta}}\left(1 + \frac{\alpha}{\theta}\frac{\omega^2}{\theta^2}\frac{\mathbf{f}^T\mathbf{H}\mathbf{f} - (\mu+\alpha\sigma)\mathbf{f}^T\mathbf{M}\mathbf{f}}{\mu+\alpha\sigma+\frac{\alpha}{\theta}}\right). \tag{8.5b}$$

By combining (8.5b) with (8.5a), we obtain the formula for $\rho(\mathbf{p})$.

The Lanczos reduction (2.2) implies

(8.6)
$$\mathbf{r} = \mathbf{A}\mathbf{p} - \left(\sigma + \frac{1}{\theta}\right)\mathbf{M}\mathbf{p} + \left(\sigma + \frac{1}{\theta} - \rho(\mathbf{p})\right)\mathbf{M}\mathbf{p} = -\mathbf{M}\mathbf{f}\frac{\omega}{\theta^2} + \left(\sigma + \frac{1}{\theta} - \rho(\mathbf{p})\right)\mathbf{M}\mathbf{p}$$

so that

$$\mathbf{r}^T\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}\mathbf{r} = \mathbf{f}^T\mathbf{M}\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{f}\left(\frac{\omega}{\theta^2}\right)^2 - \mathbf{f}^T\mathbf{M}\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{p}\frac{2\omega}{\theta^2}\left(\sigma + \frac{1}{\theta} - \rho(\mathbf{p})\right) + \mathbf{p}^T\mathbf{M}\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{p}\left(\sigma + \frac{1}{\theta} - \rho(\mathbf{p})\right)^2.$$

The relations

$$\begin{cases} \left(\min_i \dfrac{1}{(\lambda_i - \sigma)^2}\right)\|\mathbf{f}\|_{\mathbf{H}}^2 \leq \|\mathbf{M}\mathbf{f}\|_{\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}}^2 \leq \left(\max_i \dfrac{1}{(\lambda_i - \sigma)^2}\right)\|\mathbf{f}\|_{\mathbf{H}}^2, \\ \mathbf{p}^T\mathbf{M}\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{f} = \mathcal{O}\left(\|\mathbf{f}\|_{\mathbf{H}}^2\dfrac{\omega^2}{\theta^2}\right), \\ \sigma + \dfrac{1}{\theta} - \rho(\mathbf{p}) = \mathcal{O}\left(\|\mathbf{f}\|_{\mathbf{H}}^2\dfrac{\omega^2}{\theta^2}\right), \\ \mathbf{p}^T\mathbf{M}\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{p} = \mathbf{x}^T\mathbf{M}\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{x} + \mathcal{O}\left(\|\mathbf{f}\|_{\mathbf{H}}^2\dfrac{\omega^2}{\theta^2}\right), \\ \mathbf{x}^T\mathbf{M}\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}\mathbf{M}\mathbf{x} = (\mathbf{x}\theta + \mathbf{f}\omega)^T\mathbf{H}(\mathbf{x}\theta + \mathbf{f}\omega) = \theta^2 + \|\mathbf{f}\|_{\mathbf{H}}^2\dfrac{\omega^2}{\theta^2} \end{cases}$$

establish

$$\begin{cases} \|\mathbf{r}\|_{\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}} &= \|\mathbf{M}\mathbf{f}\|_{\mathbf{A}_\sigma^{-1}\mathbf{H}\mathbf{A}_\sigma^{-1}}\frac{|\omega|}{\theta^2} + \mathcal{O}\left(\|\mathbf{f}\|_{\mathbf{H}}^2\frac{\omega^2}{\theta^2}\right), \\ \|\mathbf{p}\|_{\mathbf{H}} &= \sqrt{1 + \|\mathbf{f}\|_{\mathbf{H}}^2\frac{\omega^2}{\theta^2}}, \end{cases}$$

which proves (6.2).

Next, we consider the case where $\mathbf{M}$ is symmetric positive definite. Starting from (8.6), we have

$$\mathbf{r}^T\mathbf{M}^{-1}\mathbf{r} = \mathbf{f}^T\mathbf{M}\mathbf{f}\left(\frac{\omega}{\theta^2}\right)^2 - \mathbf{f}^T\mathbf{M}\mathbf{p}\frac{2\omega}{\theta^2}\left(\sigma + \frac{1}{\theta} - \rho(\mathbf{p})\right) + \mathbf{p}^T\mathbf{M}\mathbf{p}\left(\sigma + \frac{1}{\theta} - \rho(\mathbf{p})\right)^2.$$

The relations

$$\begin{cases} \left(\min_i \dfrac{1}{\alpha\lambda_i + \mu}\right)\|\mathbf{f}\|_{\mathbf{H}}^2 \leq \|\mathbf{f}\|_{\mathbf{M}}^2 \leq \left(\max_i \dfrac{1}{\alpha\lambda_i + \mu}\right)\|\mathbf{f}\|_{\mathbf{H}}^2, \\ \mathbf{p}^T\mathbf{M}\mathbf{p} = \dfrac{1}{\mu + \alpha\sigma + \frac{\alpha}{\theta}} + \mathcal{O}\left(\|\mathbf{f}\|_{\mathbf{H}}^2\dfrac{\omega^2}{\theta^2}\right), \\ \sigma + \dfrac{1}{\theta} - \rho(\mathbf{p}) = \mathcal{O}\left(\|\mathbf{f}\|_{\mathbf{H}}^2\dfrac{\omega^2}{\theta^2}\right), \\ \mathbf{p}^T\mathbf{M}\mathbf{f}\dfrac{\omega}{\theta} = \mathcal{O}\left(\|\mathbf{f}\|_{\mathbf{H}}^2\dfrac{\omega^2}{\theta^2}\right) \end{cases}$$

establish

$$\left\{\begin{array}{rcl} \|\mathbf{r}\|_{\mathbf{M}^{-1}} & = & \|\mathbf{f}\|_{\mathbf{M}} \frac{|\omega|}{\theta^2} + \mathcal{O}\left(\|\mathbf{f}\|^2_{\mathbf{H}} \frac{\omega^2}{\theta^2}\right), \\ \|\mathbf{p}\|_{\mathbf{M}} & = & \sqrt{\frac{1}{\mu+\alpha\sigma+\frac{\alpha}{\theta}}} + \mathcal{O}\left(\|\mathbf{f}\|^2_{\mathbf{H}} \frac{\omega^2}{\theta^2}\right), \end{array}\right.$$

which proves (6.3).

We assume here that $\mathbf{A}$ is symmetric positive definite. Starting from (8.6), we have

$$\mathbf{r}^T\mathbf{A}^{-1}\mathbf{r} = \mathbf{f}^T\mathbf{M}\mathbf{A}^{-1}\mathbf{M}\mathbf{f}\left(\frac{\omega}{\theta^2}\right)^2 - \mathbf{f}^T\mathbf{M}\mathbf{A}^{-1}\mathbf{M}\mathbf{p}\frac{2\omega}{\theta^2}\left(\sigma+\frac{1}{\theta}-\rho(\mathbf{p})\right) + \mathbf{p}^T\mathbf{M}\mathbf{A}^{-1}\mathbf{M}\mathbf{p}\left(\sigma+\frac{1}{\theta}-\rho(\mathbf{p})\right)^2.$$

The relations

$$\left\{\begin{array}{l} \left(\min_i \frac{1}{\lambda_i(\alpha\lambda_i+\mu)}\right)\|\mathbf{f}\|^2_{\mathbf{H}} \le \|\mathbf{M}\mathbf{f}\|^2_{\mathbf{A}^{-1}} \le \left(\max_i \frac{1}{\lambda_i(\alpha\lambda_i+\mu)}\right)\|\mathbf{f}\|^2_{\mathbf{H}}, \\ \sigma+\frac{1}{\theta}-\rho(\mathbf{p}) = \mathcal{O}\left(\|\mathbf{f}\|^2_{\mathbf{H}}\frac{\omega^2}{\theta^2}\right), \\ \mathbf{p}^T\mathbf{M}\mathbf{A}^{-1}\mathbf{M}\mathbf{f}\frac{\omega}{\theta} = \mathbf{x}^T\mathbf{M}\mathbf{A}^{-1}\mathbf{M}\mathbf{f}\frac{\omega}{\theta} + \mathbf{f}^T\mathbf{M}\mathbf{A}^{-1}\mathbf{M}\mathbf{f}\frac{\omega^2}{\theta^2}, \\ \mathbf{x}^T\mathbf{M}\mathbf{A}^{-1}\mathbf{M}\mathbf{f} = \mathcal{O}\left(\|\mathbf{f}\|^2_{\mathbf{H}}\frac{\omega^2}{\theta^2}\right), \\ \mathbf{p}^T\mathbf{M}\mathbf{A}^{-1}\mathbf{M}\mathbf{p} = \mathbf{x}^T\mathbf{M}\mathbf{A}^{-1}\mathbf{M}\mathbf{x} + \mathcal{O}\left(\|\mathbf{f}\|^2_{\mathbf{H}}\frac{\omega^2}{\theta^2}\right) \end{array}\right.$$

establish

$$\left\{\begin{array}{rcl} \|\mathbf{r}\|_{\mathbf{A}^{-1}} & = & \|\mathbf{M}\mathbf{f}\|_{\mathbf{A}^{-1}} \frac{|\omega|}{\theta^2} + \mathcal{O}\left(\|\mathbf{f}\|^2_{\mathbf{H}} \frac{\omega^2}{\theta^2}\right), \\ \|\mathbf{p}\|_{\mathbf{A}} & = & \sqrt{\frac{\rho(\mathbf{p})}{\mu+\alpha\sigma+\frac{\alpha}{\theta}}} + \mathcal{O}\left(\|\mathbf{f}\|^2_{\mathbf{H}} \frac{\omega^2}{\theta^2}\right), \end{array}\right.$$

which proves (6.4).

Acknowledgments. The authors thank Prof. Beresford Parlett (UC Berkeley) and the anonymous referees for comments that led to improvements of the manuscript.

REFERENCES

[1] P. Arbenz, U. Hetmaniuk, R. Lehoucq, and R. Tuminaro, *A comparison of eigensolvers for large-scale 3D modal analysis using AMG-preconditioned iterative methods*, Internat. J. Numer. Methods Engrg., 64 (2005), pp. 204–236.

[2] I. Babuška and J. Osborn, *Eigenvalue problems*, in Handbook of Numerical Analysis, Vol. II, P. G. Ciarlet and J.-L. Lions, eds., Elsevier, Amsterdam, 1991, pp. 641–788.

[3] T. Ericsson and A. Ruhe, *The spectral transformation Lanczos method for the numerical solution of large sparse generalized symmetric eigenvalue problems*, Math. Comp., 35 (1980), pp. 1251–1268.
[4] R. Grimes, J. Lewis, and H. Simon, *A shifted block Lanczos algorithm for solving sparse symmetric generalized eigenproblems*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 228–272.
[5] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems by Implicitly Restarted Arnoldi Methods*, Society for Industrial and Applied Mathematics, Philadelphia, 1998.
[6] B. Nour-Omid, B. N. Parlett, T. Ericsson, and P. Jensen, *How to implement the spectral transformation*, Math. Comp., 48 (1987), pp. 663–673.
[7] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Society for Industrial and Applied Mathematics, Philadelphia, 1998.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 4, pp. 949–960

RELATIVE RESIDUAL BOUNDS FOR EIGENVALUES OF HERMITIAN MATRICES*

NINOSLAV TRUHAR†

Abstract. This paper presents a linear and quadratic residual bound for eigenvalues of an indefinite possible singular Hermitian matrix. These bounds are a generalization of results on a semidefinite Hermitian matrix [Z. Drmač and V. Hari, *SIAM J. Matrix Anal. Appl.*, 18 (1997), pp. 21–29]. The bounds here contain an extra factor which depends on the norm of a J-unitary matrix, where J is diagonal matrix with ± 1 on its diagonal.

Key words. residual bounds, quadratic residual bounds, indefinite Hermitian matrix, eigenvalues, perturbation theory, relative perturbations

AMS subject classifications. 15A03, 15A18, 47A55

DOI. 10.1137/050625059

1. Introduction.

Let $H \in \mathbf{C}^{n\times n}$ be a Hermitian matrix, let $X \in \mathbf{C}^{n\times m}$ have orthonormal columns (such X we call orthonormal), and

$$M = X^*HX, \qquad R = HX - XM, \qquad \mathcal{X} = \mathcal{R}(X)\,, \tag{1.1}$$

where $\mathcal{X} = \mathcal{R}(X)$ denotes the column space of X. Furthermore, let

$$\lambda_1 \geq \cdots \geq \lambda_n \quad \text{and} \quad \mu_1 \geq \cdots \geq \mu_m \tag{1.2}$$

be the eigenvalues of H and M, respectively. Throughout the paper $\|\cdot\|$ denotes the 2-norm.

The main purpose of this paper is to derive a linear and quadratic residual bound for eigenvalues of an indefinite possible singular Hermitian matrix and to obtain a geometric interpretation of these bounds.

In [9] the following linear residual bound for nonsingular indefinite Hermitian matrices has been presented.

THEOREM 1.1. *Let $H = GJG^*$, where G and J are nonsingular and J is diagonal with ± 1 on its diagonal and let*

$$\delta H = RX^* + XR^*, \tag{1.3}$$

where X is an $n \times m$ orthonormal matrix. Then there are at least m eigenvalues λ_{i_k}, $k = 1, \ldots, m$, of H for which

$$\frac{|\lambda_{i_k} - \mu_k|}{|\lambda_{i_k}|} \leq \kappa(V)\|L^{-1}\delta H L^{-*}\|, \qquad k = 1, \ldots, m\,. \tag{1.4}$$

Here, μ_k are eigenvalues of the matrix M defined by (1.1) *and* (1.2) *and V is a J-unitary matrix which diagonalizes the pair (G^*G, J), that is, $V^*G^*GV = |\Lambda|$ and $V^*JV = J$.*

*Received by the editors February 23, 2005; accepted for publication (in revised form) by J. Barlow October 11, 2005; published electronically December 18, 2006. This work was supported by grants 0235001 and 0023002 from the Croatian Ministry of Science and Technology.
http://www.siam.org/journals/simax/28-4/62505.html
†Department of Mathematics, University J.J. Strossmayer, Trg Ljudevita Gaja 6, 31000 Osijek, Croatia (ntruhar@mathos.hr).

We show that we can obtain a bound on $|\lambda_{i_k} - \mu_k|/|\mu_k|$ that allows H to be singular. Our bounds are a generalization of linear residual bounds for positive semidefinite matrices presented in [1, Theorems 1.1. and 2.1].

On the other hand, all existing quadratic residual bounds for general Hermitian matrices belong to the classical perturbation theory.

Let $\sigma(H)$ denote the spectra of H. The first result is due to Sun [6].

THEOREM 1.2 (Sun). *Let $\mathcal{Y} = \mathcal{R}(Y)$ be an invariant subspace of H with orthonormal basis $Y \in \mathbf{C}^{n\times m}$. Let $\lambda_{j_1} \geq \cdots \geq \lambda_{j_m}$ be the eigenvalues of Y^*HY, and $\Lambda_{\mathcal{Y}} = \operatorname{diag}(\lambda_{j_1}, \ldots, \lambda_{j_m})$, $\Lambda_{\mathcal{X}} = \operatorname{diag}(\mu_1, \ldots, \mu_m)$. If for some $\alpha, \beta \in \mathbf{R}$ and $\delta_0 > 0$, $\sigma(M) \subset [\alpha, \beta]$, $\sigma(H) \setminus \sigma(Y^*HY) \subset (-\infty, \alpha - \delta_0] \cup [\beta + \delta_0, +\infty)$ (or vice versa), and if $\rho \equiv \|R\|/\delta_0 < 1$, where R is defined by (1.1), then for any unitary invariant norm $|||\cdot|||$,*

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

The second result is due to Mathias [5], and this result is a generalization of the result obtained by Theorem 1.2.

Let

$$H = \begin{bmatrix} A & R_0 \\ R_0^* & B \end{bmatrix} \quad \text{and} \quad \widetilde{H} = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} \tag{1.5}$$

be Hermitian matrices. For the measure of separation between eigenvalues λ_k, $k = 1, \ldots, n$ of the matrix H from eigenvalues $\mu_i(B)$ of the matrix B we define

$$\delta_k \equiv \min_{i=1,\ldots,m} |\lambda_k - \mu_i(B)|.$$

For the measure of separation between eigenvalues $\widetilde{\lambda}_k$, $k = 1, \ldots, n$ of the matrix $\widetilde{H}$ from eigenvalues $\mu_i(B)$ of the matrix B we use

$$\widetilde{\delta}_k \equiv \min_{i=1,\ldots,m} |\widetilde{\lambda}_k - \mu_i(B)|.$$

THEOREM 1.3 (see [5, Theorem 1]). *If $\lambda_k \notin \sigma(B)$, then*

$$|\lambda_k - \widetilde{\lambda}_k| \leq \delta_k^{-1} \|R_0\|^2,$$

while if $\widetilde{\lambda}_k \notin \sigma(B)$, then

$$|\lambda_k - \widetilde{\lambda}_k| \leq \widetilde{\delta}_k^{-1} \|R_0\|^2.$$

One of the latest results which considers similar problems belongs to Chi-Kwong Li and Ren-Cang Li. Let H and $\widetilde{H}$ be defined as in 1.5, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ and $\widetilde{\lambda}_1 \geq \cdots \geq \widetilde{\lambda}_n$, respectively. In [3, Theorem 2] they have shown that the following bound holds:

$$|\lambda_j - \widetilde{\lambda}_j| \leq \frac{2\|R_0\|^2}{\eta + \sqrt{\eta^2 + 4\|R_0\|^2}}, \tag{1.6}$$

where $\eta = \min_{i,j} |\mu_i(A) - \mu_j(B)|$ denotes the spectral gap between the spectra of A and B in (1.5) (here $\mu_i(A)$ denotes the ith eigenvalue of the matrix A).

Similar to the linear case our quadratic bound is a generalization of the quadratic residual bound for positive semidefinite matrices presented in [1].

2. Linear residual bound. In this section, we present relative residual bounds for indefinite possible singular Hermitian matrices. Let $\mathcal{X}$ be an m-dimensional subspace of $\mathbf{C}^n$ and let X and $X_\perp$ be any orthonormal basis for $\mathcal{X}$ and $\mathcal{X}_\perp$, respectively. Let $H = GJG^*$ be any factorization of H. For simplicity, assume that G has a full column rank (this assumption is not crucial for the obtained results; all results will remain the same if G has one or more zero columns).

First we will separate the null subspace of the matrix H from the rest of the subspaces.

Let $X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$ and $X_\perp = \begin{bmatrix} X_{\perp,1} & X_{\perp,2} \end{bmatrix}$ be orthonormal bases of $\mathcal{X}$, $\mathcal{X}_\perp$, respectively, such that $G^*X_1 = 0$ and $G^*X_{\perp,2} = 0$ and

$$M = \begin{bmatrix} X_1 & X_2 \end{bmatrix}^* H \begin{bmatrix} X_1 & X_2 \end{bmatrix} = \begin{bmatrix} 0 & \\ & \Lambda_1 \end{bmatrix} \begin{matrix} m - r_M \\ r_M \end{matrix}, \tag{2.1}$$

$$N = \begin{bmatrix} X_{\perp,1} & X_{\perp,2} \end{bmatrix}^* H \begin{bmatrix} X_{\perp,1} & X_{\perp,2} \end{bmatrix} = \begin{bmatrix} \Lambda_2 & \\ & 0 \end{bmatrix} \begin{matrix} r_N \\ n - m - r_M \end{matrix}. \tag{2.2}$$

If we define $r_M = \operatorname{rank}(M)$ and $r_N = \operatorname{rank}(N)$, then we can write

$$\begin{bmatrix} X & X_\perp \end{bmatrix}^* H \begin{bmatrix} X & X_\perp \end{bmatrix} = \begin{bmatrix} 0 & & \\ & \widehat{H} & \\ & & 0 \end{bmatrix} \begin{matrix} m - r_M \\ r_M + r_N \\ n - m - r_N \end{matrix},$$

where

$$\widehat{H} = \begin{bmatrix} \Lambda_1 & K^* \\ K & \Lambda_2 \end{bmatrix} \begin{matrix} r_M \\ r_N \end{matrix}. \tag{2.3}$$

Note that $\widehat{H}$ is nonsingular.

The following theorem contains a relative perturbation bound for eigenvalues of H and Rayleigh–Ritz approximations of eigenvalues of H, that is, eigenvalues of M, where M is given by (2.1) .

THEOREM 2.1. *Let $H = GJG^*$ be an indefinite Hermitian matrix (possibly singular), where J is a diagonal matrix with ± 1 on its diagonal. Let X and $X_\perp$ be orthonormal matrices as in* (2.1) *and* (2.2)*. If we define $K_S = |\Lambda_2|^{-\frac{1}{2}} K |\Lambda_1|^{-\frac{1}{2}}$, then*

$$\frac{|\lambda_{i_k} - \mu_{m-r_M+k}|}{|\mu_{m-r_M+k}|} \le \|K_S\|, \qquad k = 1, \ldots, r_M, \tag{2.4}$$

$$\mu_k = \lambda_{j_k} = 0, \qquad k = 1, \ldots, m - r_M,$$

$$\frac{|\lambda_{i_k} - \mu_{m+k}|}{|\mu_{m+k}|} \le \|K_S\|, \qquad k = 1, \ldots, r_N, \tag{2.5}$$

$$\mu_k = \lambda_{j_k} = 0, \qquad k = m + r_N + 1, \ldots, n.$$

Proof. Let $\widehat{H}_0$ be a diagonal matrix

$$\widehat{H}_0 = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} \begin{matrix} r_M \\ r_N \end{matrix}, \tag{2.6}$$

where Λ_1 and Λ_2 are defined by (2.1) and (2.2). Then we can write $\widehat{H}_0 = D^*AD$, where

$$D = \begin{bmatrix} |\Lambda_1|^{1/2} & 0 \\ 0 & |\Lambda_2|^{1/2} \end{bmatrix}, \qquad A = \begin{bmatrix} J_1 & 0 \\ 0 & J_2 \end{bmatrix}. \tag{2.7}$$

Here J_1 and J_2 are diagonal matrices with signs of eigenvalues of Λ_1 and Λ_2, respectively.

Note that $\widehat{H}$ from (2.3) can be considered as a perturbation of $\widehat{H}_0$. Indeed, since

$$\widehat{H} = D^* D^{-*} \widehat{H} D^{-1} D,$$

we have $\widehat{H} = D^*(A + \delta A)D$, where

$$\delta A = \begin{bmatrix} 0 & K_S^* \\ K_S & 0 \end{bmatrix}. \tag{2.8}$$

By applying a result of Veselić and Slapničar [10, Theorem 2.1], we know that if

$$|x^* \delta H x| \leq \eta x^* |H|_S x \qquad \forall\ x, \qquad \eta < 1,$$

where $|H|_S = U|\Lambda|U^*$ is the spectral absolute value, and $H = U|\Lambda|U^*$ is a corresponding eigenvalue decomposition of H, then the following bound holds:

$$1 - \eta \leq \frac{\widetilde{\lambda}_i}{\lambda_i} \leq 1 + \eta.$$

Note that (2.6)–(2.8) imply $\eta = \|\delta A\|$. Indeed

$$\begin{aligned} |x^* \delta H x| = |x^* D^* \delta A D x| &= \left| x^* D^* |A|_S^{1/2} \, |A|_S^{-1/2} \, \delta A \, |A|_S^{-1/2} \, |A|_S^{1/2} D x \right| \\ &\leq \left\| \, |A|_S^{-1/2} \delta A |A|_S^{-1/2} \right\| \ \left| \, |A|_S^{1/2} D x \right|^2 = \eta \, x^* |H|_S x \, , \end{aligned}$$

where $\eta = \| \, |A|_S^{-1/2} \delta A |A|_S^{-1/2} \|$, and $|H|_S = x^* D^* |A|_S D x$. From (2.7) it follows that $|A|_S = I$; thus we can write $\eta = \|\delta A\|$.

Now all above mentioned results imply that we can write

$$\frac{|\lambda_{i_k} - \mu_{m - r_M + k}|}{|\mu_{m - r_M + k}|} \leq \eta, \qquad k = 1, \ldots, r_M.$$

Since $\eta = \|\delta A\| = \|K_S\|$, we obtained the first part of (2.4). Similar holds for (2.5). □

In the case of a positive semidefinite H the following holds (properties of a positive semidefinite H are considered in [1] in detail).

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

$$y^* G G^* x = 0 \, , \qquad x \in \mathcal{X} \, , \qquad y \in \mathcal{X}_\perp \, .$$

Hence for the subspaces $\mathcal{Y}_G = G^* \mathcal{X}$ and $\mathcal{U}_G = G^* \mathcal{X}_\perp$ we have $\mathcal{Y}_G \subseteq {\mathcal{U}_G}^\perp$ and $\mathcal{U}_G \subseteq {\mathcal{Y}_G}^\perp$. This indicates that, if G is square and nonsingular, the maximal canonical angle between $\mathcal{Y}_G$ and ${\mathcal{U}_G}^\perp$ as well as between $\mathcal{U}_G$ and ${\mathcal{Y}_G}^\perp$ is zero.

In [1, Theorem 1.1] it has been shown that if H is positive semidefinite then $\|K_S\| = \sin \angle \left(\mathcal{Y}_G, \mathcal{U}_G^\perp \right)$, where the angle function is defined by (see [11])

$$\sin \angle \left(\mathcal{Y}_G, \mathcal{U}_G^\perp \right) = \min \left\{ \left\| P_{\mathcal{U}_G} P_{\mathcal{Y}_G} \right\|, \left\| P_{\mathcal{U}_G^\perp} P_{\mathcal{Y}_G^\perp} \right\| \right\} ;$$

here $P_{\mathcal{M}}$ denotes an orthogonal projector onto $\mathcal{M}$.

Note that if H is positive semidefinite, then Theorem 2.1 has the same geometric interpretation as [1, Theorem 1.1]. However if $H = GJG^*$ is indefinite, we cannot express $\|K_S\|$ in terms of a sine of canonical angles. The following lemma presents the upper bound for $\|K_S\|$ which contains such a sine.

LEMMA 2.2. *Let $H = GJG^*$, J, X and $X_\perp$ be as in Theorem* 2.1. *If $G^*X_1 = 0$ and $G^*X_{\perp,2} = 0$, and if we set $\mathcal{W}_G = \mathcal{R}(JG^*X_\perp)$ $\mathcal{Y}_G = \mathcal{R}(G^*X)$, then*

$$\|K_S\| \le \|U\|\|Y\| \sin\phi, \tag{2.9}$$

where $U = G^*X_{\perp,1}|\Lambda_2|^{-\frac{1}{2}}$, $Y = G^*X_2|\Lambda_1|^{-\frac{1}{2}}$, and $\sin\phi$ is defined by

$$\sin\phi = \sin\angle\left(\mathcal{Y}_G, \mathcal{W}_G^\perp\right) = \min\left\{\left\|P_{\mathcal{W}_G}P_{\mathcal{Y}_G}\right\|, \left\|P_{\mathcal{W}_G^\perp}P_{\mathcal{Y}_G^\perp}\right\|\right\}. \tag{2.10}$$

Proof. From $K = X_{\perp,1}HX_2$ and the definition of K_S, we have

$$\begin{aligned} K_S &= |\Lambda_2|^{-\frac{1}{2}}K|\Lambda_1|^{-\frac{1}{2}} = |\Lambda_2|^{-\frac{1}{2}}X_{\perp,1}^*HX_2|\Lambda_1|^{-\frac{1}{2}} = |\Lambda_2|^{-\frac{1}{2}}X_{\perp,1}^*GJG^*X_2|\Lambda_1|^{-\frac{1}{2}} \\ &= \left(G^*X_{\perp,1}|\Lambda_2|^{-\frac{1}{2}}\right)^* J\left(G^*X_2|\Lambda_1|^{-\frac{1}{2}}\right) = U^*JY = W^*Y, \end{aligned}$$

where $W = JU$. Note that

$$\begin{aligned} Y^*JY &= |\Lambda_1|^{-\frac{1}{2}}X_2^*GJG^*X_2|\Lambda_1|^{-\frac{1}{2}} \\ &= |\Lambda_1|^{-\frac{1}{2}}X_2^*HX_2|\Lambda_1|^{-\frac{1}{2}} = |\Lambda_1|^{-\frac{1}{2}}\Lambda_1|\Lambda_1|^{-\frac{1}{2}} = J_1, \\ W^*JW &= |\Lambda_2|^{-\frac{1}{2}}X_{\perp,1}^*GJG^*X_{\perp,1}|\Lambda_2|^{-\frac{1}{2}} \\ &= |\Lambda_2|^{-\frac{1}{2}}X_{\perp,1}^*HX_{\perp,1}|\Lambda_2|^{-\frac{1}{2}} = |\Lambda_2|^{-\frac{1}{2}}\Lambda_2|\Lambda_2|^{-\frac{1}{2}} = J_2. \end{aligned}$$

This shows that W and Y have J-orthogonal columns. Further, from

$$\begin{aligned} \mathcal{R}(Y) &= \mathcal{R}\left(G^*X_2|\Lambda_1|^{-\frac{1}{2}}\right) = \mathcal{R}\left(G^*X_2\right) \subset \mathcal{R}\left(G^*X\right) = \mathcal{Y}_G, \\ \mathcal{R}(W) &= \mathcal{R}\left(JG^*X_{\perp,1}|\Lambda_2|^{-\frac{1}{2}}\right) = \mathcal{R}\left(JG^*X_\perp\right) \subset \mathcal{R}\left(JG^*X\right) = \mathcal{W}_G, \end{aligned}$$

and from $G^*X_1 = 0$ and $G^*X_{\perp,2} = 0$, it follows that $\mathcal{R}(Y) = \mathcal{Y}_G$ and $\mathcal{R}(W) = \mathcal{W}_G$. Finally, let $W = Q_WR_W$ and $Y = Q_YR_Y$ be QR decompositions of W and Y, respectively. Then

$$\|K_S\| = \|W^*Y\| \le \|R_W^*\|\|R_Y\|\|Q_W^*Q_Y\|. \tag{2.11}$$

The columns of Q_W and Q_Y form orthonormal bases for $\mathcal{W}_G$ and $\mathcal{Y}_G$, respectively. Drmač and Hari have shown in proof of [1, Theorem 1.1] that $\|Q_W^*Q_Y\| = \sin\phi$. Now, using this and the fact that

$$\|R_W^*\| = \|U\|, \qquad \|R_Y\| = \|Y\|,$$

from (2.11) follows (2.9). ∎

Inserting (2.9) into (2.4) and (2.5) we obtain the bound which is a generalization of [1, Theorem 1.1] to indefinite Hermitian matrices, since in the semidefinite case ($J = I$) this bound is equal to $\sin\phi$, and $W \equiv U$ and Y have orthonormal columns.

Note that in the positive definite case the angle function $\angle(\mathcal{Y}_G, \mathcal{Z}_G)$ defined by (2.10) does not depend on G but only on H (see [1]). However, in the indefinite case, this is not true in general. The dependence of the angle function $\angle(\mathcal{Y}_G, \mathcal{Z}_G)$

on the factor G, where $H = GJG^*$, for a nonsingular indefinite matrix H, has been considered in [9]. Now we will present a similar result for the indefinite possible singular matrix H. Let

$$H = G_1 J G_1^* = G_2 J G_2^* \tag{2.12}$$

be decompositions of the matrix H, $i = 1, 2$.

From (2.12) it follows that there exists a nonsingular J-unitary matrix V such that

$$G_2 = G_1 V, \qquad V^* J V = V J V^* = J, \qquad \|V\| = \|V^{-1}\| .$$

Further, let W_i and Y_i be defined as in the proof of Lemma 2.2, that is,

$$W_1 = J G_1^* X_{\perp,1} |\Lambda_2|^{-\frac{1}{2}}, \qquad W_2 = J G_2^* X_{\perp,1} |\Lambda_2|^{-\frac{1}{2}} \tag{2.13}$$

$$Y_1 = G_1^* X_2 |\Lambda_1|^{-\frac{1}{2}}, \qquad Y_1 = G_1^* X_2 |\Lambda_1|^{-\frac{1}{2}} . \tag{2.14}$$

Using the fact that $JV^* = V^{-1}J$, it follows that

$$W_2 = V^{-1} W_1, \qquad Y_2 = V^* Y_1 ,$$

which further implies that the matrix K_S does not depend on decomposition (2.12). Indeed $K_S = W_2^* Y_2 = W_1^* Y_1$.

Let $W_i = Q_{W_i} R_{W_i}$ and $Y_i = Q_{Y_i} R_{Y_i}$ be QR decompositions of W_i and Y_i, respectively, for $i = 1, 2$. Note that (2.11) can be written as $\|K_S\| = \|W_i^* Y_i\|$, $i = 1, 2$. Now, using the simple inequalities

$$\begin{aligned} \sin\phi_2 = \|R_{W_2}^{-*} R_{W_2}^* Q_{W_2}^* Q_{Y_2} R_{Y_2} R_{Y_2}^{-1}\| &\le \|R_{W_2}^{-*}\| \, \|R_{Y_2}^{-1}\| \|K_S\| \\ &\le \|R_{W_2}^{-*}\| \, \|R_{Y_2}^{-1}\| \, \|R_{W_1}\| \, \|R_{Y_1}\| \sin\phi_1 \end{aligned}$$

and similarly,

$$\sin\phi_1 \le \|R_{W_1}^{-*}\| \, \|R_{Y_1}^{-1}\| \, \|R_{W_2}\| \, \|R_{Y_2}\| \sin\phi_2 ,$$

we can write the following bound

$$\|U_1^\dagger\| \, \|Y_1^\dagger\| \, \|U_2\| \, \|Y_2\| \le \frac{\sin\phi_2}{\sin\phi_1} \le \|U_2^\dagger\| \, \|Y_2^\dagger\| \, \|U_1\| \, \|Y_1\| , \tag{2.15}$$

where we have used the fact that $\|R_{W_i}^{-*}\| = \|U_i^\dagger\|$ and $\|R_{Y_i}^{-*}\| = \|Y_i^\dagger\|$, $i = 1, 2$. Here $\dagger$ denotes the generalized inverse.

Further, from (2.13) it follows that $Y_2 = V^* Y_1$ and $U_2 = V^* U_1$. Using this, the upper bound for the right-hand side of (2.15) has the following form:

$$\|U_2^\dagger\| \, \|U_1\| \, \|Y_1\| \, \|Y_2^\dagger\| \le \|V\|^2 \operatorname{cond}(U_1) \operatorname{cond}(Y_1) , \tag{2.16}$$

where $\operatorname{cond}(U) = \|U^\dagger\| \, \|U\|$. On the other hand, the lower bound for the left-hand side has the following form:

$$\|U_2^\dagger\| \, \|U_1\| \, \|Y_1\| \, \|Y_2^\dagger\| \ge \frac{1}{\|V\|^2} \operatorname{cond}(U_1) \operatorname{cond}(Y_1) . \tag{2.17}$$

Now from (2.15), (2.16), and (2.17) if follows that

$$\frac{1}{\|V\|^2}\operatorname{cond}(U_1)\operatorname{cond}(Y_1) \le \frac{\sin\phi_2}{\sin\phi_1} \le \|V\|^2\operatorname{cond}(U_1)\operatorname{cond}(Y_1)\,. \tag{2.18}$$

Recall that $\operatorname{cond}(V) = \|V\|^2$, which means if J-unitary V ($G_2 = V^*G_1$) has a condition number of the modest magnitude, then the corresponding angles will be close, that is, $\sin\phi_1 \approx \sin\phi_2$.

The classes of so-called "well-behaved matrices" for which there exist useful bounds for conditions of V have been considered in [7]. This class includes matrices such as scaled diagonally dominant matrices, block scaled diagonally dominant (BSDD) matrices, and quasi-definite matrices. Details about these bounds can be found in, e.g., [8, Section 3.1] and [7].

3. Quadratic residual bound. In this section we will present a quadratic relative residual bound for the eigenvalues of an indefinite singular Hermitian matrix and compare it with results from the classical perturbation theory.

The main result of this section is a generalization of Drmač and Hari's Theorem 2.1 from [1] to indefinite possibly singular Hermitian matrices.

In the following theorem $\sigma_{\min}(\cdot)$ denotes the smallest singular value of a matrix. We will use the same notation as in Theorem 2.1. For a given nonzero eigenvalue λ of H we shall choose the bases X and $X^\perp$ such that

$$\Lambda_1 = \Xi_\lambda \oplus \widehat{\Xi}_\lambda, \qquad \Lambda_2 = \Omega_\lambda \oplus \widehat{\Omega}_\lambda, \tag{3.1}$$

where the diagonals of Ξ_λ and Ω_λ approximate λ in the sense of Theorem 2.1.

Let Λ_1 and Λ_2 be decomposed as

$$\Lambda_1 = \begin{bmatrix} |\Xi_\lambda|^{1/2} & 0 \\ 0 & |\widehat{\Xi}_\lambda|^{1/2} \end{bmatrix} \begin{bmatrix} J_{11} & 0 \\ 0 & J_{22} \end{bmatrix} \begin{bmatrix} |\Xi_\lambda|^{1/2} & 0 \\ 0 & |\widehat{\Xi}_\lambda|^{1/2} \end{bmatrix}, \tag{3.2}$$

$$\Lambda_2 = \begin{bmatrix} |\Omega_\lambda|^{1/2} & 0 \\ 0 & |\widehat{\Omega}_\lambda|^{1/2} \end{bmatrix} \begin{bmatrix} \bar{J}_{11} & 0 \\ 0 & \bar{J}_{22} \end{bmatrix} \begin{bmatrix} |\Omega_\lambda|^{1/2} & 0 \\ 0 & |\widehat{\Omega}_\lambda|^{1/2} \end{bmatrix}, \tag{3.3}$$

where J_{11}, J_{22}, $\bar{J}_{11}$, $\bar{J}_{22}$ are diagonal matrices with ± 1 on the diagonal. We write $J = J_{11} \oplus J_{22}$, $\bar{J} = \bar{J}_{11} \oplus \bar{J}_{22}$.

THEOREM 3.1. *Let H, $\mathcal{X}$ be as in Theorem* 2.1. *Let $\lambda > 0$ be an eigenvalue of H of multiplicity $n(\lambda)$ (for $\lambda < 0$ we consider $-H$). Let the orthonormal bases of $\mathcal{X}$ and $\mathcal{X}^\perp$ be chosen such that* (3.1) *holds. Write $K_S = |\Lambda_2|^{-1/2} K |\Lambda_1|^{-1/2}$, where K is defined by* (2.3). *Suppose that there exist constants $\alpha > \gamma$ and $\beta > \gamma$ such that*

$$\|\lambda|\Xi_\lambda|^{-1} - J_{11}\| \le \gamma, \qquad \sigma_{\min}(\lambda|\widehat{\Xi}_\lambda|^{-1} - J_{22}) > \alpha, \tag{3.4}$$

$$\|\lambda|\Omega_\lambda|^{-1} - \bar{J}_{11}\| \le \gamma, \qquad \sigma_{\min}(\lambda|\widehat{\Omega}_\lambda|^{-1} - \bar{J}_{22}) > \beta. \tag{3.5}$$

If $\Xi_\lambda \oplus \Omega_\lambda$ is of order $n(\lambda)$ and $\|K_S\| \le \gamma < 1$, then

$$\|\lambda\Xi_\lambda^{-1} - I\| \le \frac{1}{1 - \dfrac{\|K_S\|^2}{\alpha\,\beta}}\,\frac{\|K_S\|^2}{\beta} \le \frac{\|U\|^2\|Y\|^2}{1 - \dfrac{\|U\|^2\|Y\|^2\sin^2\phi}{\alpha\,\beta}}\,\frac{\sin^2\phi}{\beta}, \tag{3.6}$$

$$\|\lambda\Omega_\lambda^{-1} - I\| \le \frac{1}{1 - \dfrac{\|K_S\|^2}{\alpha\,\beta}}\,\frac{\|K_S\|^2}{\alpha} \le \frac{\|U\|^2\|Y\|^2}{1 - \dfrac{\|U\|^2\|Y\|^2\sin^2\phi}{\alpha\,\beta}}\,\frac{\sin^2\phi}{\beta}, \tag{3.7}$$

where $K_S = U^*JY$ and U, Y and $\sin\phi$ are defined as in Lemma 2.2.

Proof. Our proof is similar to the proof of [1, Theorem 2.1], and most of it can be omitted but we include the whole proof for completeness. Without loss of generality we can assume

$$H = \begin{bmatrix} \Lambda_1 & K^* \\ K & \Lambda_2 \end{bmatrix} \begin{matrix} r_M \\ r_N \end{matrix}, \tag{3.8}$$

where Λ_1 and Λ_2 are given by (3.1). Otherwise one can work with $\widehat{H}$ from the proof of Theorem 2.1. Matrix H is a nonsingular matrix of dimension $r \times r$.

Matrices J and $\bar{J}$ are diagonal matrices which contain signs of Λ_1 and Λ_2 from (3.1). It is easy to show that under assumptions (3.4) and (3.5), $J_{11} = I$ and $\bar{J}_{11} = I$, thus $|\Xi_\lambda| = \Xi_\lambda$ and $|\Omega_\lambda| = \Omega_\lambda$. Indeed, let us show that $J_{11} = I$. From (3.4) we have

$$\|\lambda|\Xi_\lambda|^{-1} - J_{11}\| < 1$$

or

$$\max_j \left| \frac{\lambda}{|\lambda_j|} - \operatorname{sign}(\lambda_j) \right| < 1, \quad j = 1, \ldots, \dim(\Xi_\lambda).$$

The last inequality is equivalent to

$$|\lambda - \operatorname{sign}(\lambda_j)|\lambda_j|| < |\lambda_j|,$$

which cannot be obtained for $\lambda_j < 0$, thus $\lambda_j > 0$ for all j, and we conclude that $J_{11} = I$.

By Sylvester's law of inertia, the matrix

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

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

$$H_S(\lambda) = \begin{bmatrix} I - \lambda|\Xi_\lambda|^{-1} & 0 & \left(K_S^{(1,1)}\right)^* & \left(K_S^{(2,1)}\right)^* \\ 0 & J_{22} - \lambda|\widehat{\Xi}_\lambda|^{-1} & \left(K_S^{(1,2)}\right)^* & \left(K_S^{(2,2)}\right)^* \\ \left(K_S^{(1,1)}\right) & \left(K_S^{(1,2)}\right) & I - \lambda|\Omega_\lambda|^{-1} & 0 \\ \left(K_S^{(2,1)}\right) & \left(K_S^{(2,2)}\right) & 0 & \bar{J}_{22} - \lambda|\widehat{\Omega}_\lambda|^{-1} \end{bmatrix}.$$

Let $\widehat{H}_S(\lambda)$ be a matrix similar to $H_S(\lambda)$ defined by

$$\begin{aligned} \widehat{H}_S(\lambda) &= \Pi^T H_S(\lambda) \Pi \\ &= \begin{bmatrix} I - \lambda|\Xi_\lambda^{-1}| & \left(K_S^{(1,1)}\right)^* & 0 & \left(K_S^{(2,1)}\right)^* \\ \left(K_S^{(1,1)}\right) & I - \lambda|\Omega_\lambda|^{-1} & \left(K_S^{(1,2)}\right) & 0 \\ 0 & \left(K_S^{(1,2)}\right)^* & J_{22} - \lambda|\widehat{\Xi}_\lambda|^{-1} & \left(K_S^{(2,2)}\right)^* \\ \left(K_S^{(2,1)}\right) & 0 & \left(K_S^{(2,2)}\right) & \bar{J}_{22} - \lambda|\widehat{\Omega}_\lambda|^{-1} \end{bmatrix}, \end{aligned}$$

where Π denotes an appropriate permutation matrix.

Assumptions (3.4) and (3.5) imply

$$\sigma_{\min}\left(\left(J_{22} - \lambda|\widehat{\Xi}_\lambda|^{-1}\right) \oplus \left(\bar{J}_{22} - \lambda|\widehat{\Omega}_\lambda|^{-1}\right)\right) \geq \min\{\alpha, \beta\} \tag{3.9}$$

$$> \gamma \geq \|K_S\| \geq \max_{1 \leq i,j \leq 2} \left\| K_S^{(i,j)} \right\|. \tag{3.10}$$

Hence matrix

$$C = \begin{bmatrix} J_{22} - \lambda|\widehat{\Xi}_\lambda|^{-1} & \left(K_S^{(2,2)}\right)^* \\ \left(K_S^{(2,2)}\right) & \bar{J}_{22} - \lambda|\widehat{\Omega}_\lambda|^{-1} \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}, \qquad C_{12} = C_{21}^*$$

and its diagonal blocks C_{11} and C_{22} are nonsingular. Therefore (see [2, section 0.7.3]),

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

provided that all matrices in the brackets are nonsingular. However, this follows since these matrices are (signed) Schur complements of C_{11} and C_{22} in C. By the last assumption, C is of order $n - n(\lambda)$ which is also the rank of $\widehat{H}_S(\lambda)$. Since C is nonsingular, its Schur complement in $\widehat{H}_S(\lambda)$ must be zero (see [4, p. 183]). Hence

$$\begin{bmatrix} I - \lambda\Xi_\lambda^{-1} & \left(K_S^{(1,1)}\right)^* \\ \left(K_S^{(1,1)}\right) & I - \lambda\Omega_\lambda^{-1} \end{bmatrix} = \begin{bmatrix} 0 & \left(K_S^{(2,1)}\right)^* \\ \left(K_S^{(2,1)}\right) & 0 \end{bmatrix} C^{-1} \begin{bmatrix} 0 & \left(K_S^{(1,2)}\right)^* \\ \left(K_S^{(1,2)}\right) & 0 \end{bmatrix}. \tag{3.11}$$

From (3.11) we obtain

$$I - \lambda\Xi_\lambda^{-1} = \left(K_S^{(2,1)}\right)^* \left[\bar{J}_{22} - \lambda|\widehat{\Omega}_\lambda|^{-1} - K_S^{(2,2)}\left(J_{22} - \lambda|\widehat{\Xi}_\lambda|^{-1}\right)^{-1}\left(K_S^{(2,2)}\right)^*\right]^{-1} K_S^{(2,1)}$$

$$I - \lambda\Omega_\lambda^{-1} = \left(K_S^{(1,2)}\right)\left[J_{22} - \lambda|\widehat{\Xi}_\lambda|^{-1} - \left(K_S^{(2,2)}\right)^*\left(\bar{J}_{22} - \lambda|\widehat{\Omega}_\lambda|^{-1}\right)^{-1}\left(K_S^{(2,2)}\right)\right]^{-1}\left(K_S^{(1,2)}\right)^*.$$

Now applying a standard 2-norm to the expressions on the left- and right-hand sides we obtain

$$\left\|I - \lambda\Xi_\lambda^{-1}\right\| \le \frac{\left\|K_S^{(2,1)}\right\|^2}{\beta - \dfrac{\left\|K_S^{(2,2)}\right\|^2}{\alpha}}$$

$$\left\|I - \lambda\Omega_\lambda^{-1}\right\| \le \frac{\left\|K_S^{(1,2)}\right\|^2}{\alpha - \dfrac{\left\|K_S^{(2,2)}\right\|^2}{\beta}}.$$

Since

$$\max_{1\le i,j\le 2} \|K_S^{(i,j)}\| \le \|K_S\|,$$

the first inequalities of (3.6) and (3.7) are proved. The upper bounds for (3.6) and (3.7) follow from Lemma 2.2. ∎

Theorem 3.1 is a generalization of [1, Theorem 2.1], since in the positive semidefinite case $J = I$ and bounds (3.6) and (3.7) have the same form as the bound from [1, Theorem 2.1].

The following example is an indefinite version of Example 2.4 from [1], and it shows that for certain Hermitian matrices results from Theorems 1.2 and 1.3 cannot be applicable.

Example 1. Let

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

Then

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

Separation δ from Theorem 1.2 is of order 10^{-5}, and this also holds for separations δ_k and $\widetilde{\delta}_k$ from Theorem 1.3. In Theorem 1.3 $\|R_0\| = \|R\|$. All of this means that these theorems are not applicable.

On the other hand, Theorem 2.1 ensures that for some $j_0 \in \{1,2,3\}$ holds

$$\frac{\lambda_{j_0} - 2\cdot 10^{-5}}{\sqrt{|\lambda_{j_0}|\cdot 2\cdot 10^{-5}}} < 1.24\cdot 10^{-2}. \tag{3.12}$$

Since $\|K_s\| \approx 7.4\cdot 10^{-3}$, we can take $\gamma = 2\cdot 10^{-2}$ in Theorem 2.1. If we consider (3.12), and by taking $\beta = 0.9$, we can assume that conditions from (3.4) are satisfied. Now, the bound from Theorem 3.1 yields

$$\frac{\lambda_{j_0} - 2\cdot 10^{-5}}{\lambda_{j_0}} < 6.1\cdot 10^{-5}.$$

The next example illustrates the theory from the last two paragraphs.

Example 2. Let

$$H = \begin{bmatrix} \frac{10^{-4}}{2} & -\frac{10^{-4}}{2} & \frac{10^{-9}}{\sqrt{3}} & -\frac{10^{-9}}{\sqrt{6}} & \frac{10^{-10}}{2\sqrt{2}} & \frac{\sqrt{3}\,10^{-10}}{2\sqrt{2}} \\ -\frac{10^{-4}}{2} & \frac{10^{-4}}{2} & -\frac{10^{-9}}{\sqrt{3}} & \frac{10^{-9}}{\sqrt{6}} & -\frac{10^{-10}}{2\sqrt{2}} & -\frac{\sqrt{3}\,10^{-10}}{2\sqrt{2}} \\ \frac{10^{-9}}{\sqrt{3}} & -\frac{10^{-9}}{\sqrt{3}} & \frac{(2.510^{6}+3\sqrt{2})\,10^{-10}}{1.5} & \frac{(3-10^{6}\sqrt{2})\,10^{-10}}{3} & \sqrt{3}\cdot 10^{-10} & 3\cdot 10^{-10} \\ -\frac{10^{-9}}{\sqrt{6}} & \frac{10^{-9}}{\sqrt{6}} & \frac{(3-10^{6}\sqrt{2})\,10^{-10}}{3} & \frac{(210^{6}-3\sqrt{2})\,10^{-10}}{1.5} & \frac{\sqrt{3}\,10^{-9}}{5\sqrt{2}} & \frac{3\,10^{-9}}{5\sqrt{2}} \\ \frac{10^{-10}}{2\sqrt{2}} & -\frac{10^{-10}}{2\sqrt{2}} & \sqrt{3}\cdot 10^{-10} & \frac{\sqrt{3}\,10^{-9}}{5\sqrt{2}} & -2.5\cdot 10^{-4} & -\frac{\sqrt{3}\,10^{-3}}{4} \\ \frac{\sqrt{3}\,10^{-10}}{\sqrt{2}} & -\frac{\sqrt{3}\,10^{-10}}{\sqrt{2}} & 3\cdot 10^{-10} & \frac{3\,10^{-9}}{5\sqrt{2}} & -\frac{\sqrt{3}\,10^{-3}}{4} & -7.5\cdot 10^{-4} \end{bmatrix},$$

and let

$$X = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 0 & 0 & 1/\sqrt{3} \\ 0 & 0 & \sqrt{2}/\sqrt{3} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad X_{\perp} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \sqrt{2}/\sqrt{3} & 0 & 0 \\ -1/\sqrt{3} & 0 & 0 \\ 0 & 1/2 & \sqrt{3}/2 \\ 0 & \sqrt{3}/2 & -1/2 \end{bmatrix}.$$

From (2.1), (2.2), and (2.3) it follows that

$$M = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 10^{-4} & 0 \\ 0 & 0 & 10^{-4} \end{bmatrix}, \quad N = \begin{bmatrix} 2\cdot 10^{-4} & 0 & 0 \\ 0 & -10^{-3} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad K = \begin{bmatrix} 10^{-9} & 3\cdot 10^{-10} \\ 10^{-10} & 6\cdot 10^{-10} \end{bmatrix},$$

where $K = X_{\perp,1}^* H X_2$. Note that $X_1 = X(:,1)$, $X_2 = X(:,2:3)$, $X_{\perp,1} = X_\perp(:,1:2)$, and $X_{\perp,2} = X_\perp(:,3)$

Now, we will assume that H is given in factorized forms $H = GJG^*$ and $H = G_1 J_1 G_1^*$, where

$$G = \begin{bmatrix} 6.7967 \cdot 10^{-3} & 1.9505 \cdot 10^{-3} & -2.0328 \cdot 10^{-9} & 1.0000 \cdot 10^{-7} \\ -6.7967 \cdot 10^{-3} & -1.9505 \cdot 10^{-3} & 2.0328 \cdot 10^{-9} & -1.0000 \cdot 10^{-7} \\ 1.5925 \cdot 10^{-3} & -5.5495 \cdot 10^{-3} & -9.9586 \cdot 10^{-9} & 1.1547 \cdot 10^{-2} \\ 2.2524 \cdot 10^{-3} & -7.8482 \cdot 10^{-3} & -1.4084 \cdot 10^{-8} & -8.1649 \cdot 10^{-3} \\ 1.1892 \cdot 10^{-9} & -2.4961 \cdot 10^{-9} & 1.5811 \cdot 10^{-2} & 0 \\ 2.0598 \cdot 10^{-9} & -4.3233 \cdot 10^{-9} & 2.7386 \cdot 10^{-2} & 0 \end{bmatrix},$$

and

$$G_1 = \begin{bmatrix} 7.0711 \cdot 10^{-3} & 0 & 0 & 0 \\ -7.0711 \cdot 10^{-3} & 0 & 0 & 0 \\ 8.1650 \cdot 10^{-8} & 5.7735 \cdot 10^{-3} & 1.1547 \cdot 10^{-2} & 0 \\ -5.7735 \cdot 10^{-8} & 8.1649 \cdot 10^{-3} & -8.1650 \cdot 10^{-3} & 0 \\ 5.0000 \cdot 10^{-9} & 3.0000 \cdot 10^{-8} & -9.8995e-014 & 1.5811 \cdot 10^{-2} \\ 8.6603 \cdot 10^{-9} & 5.1962 \cdot 10^{-8} & -1.7146e-013 & 2.7386 \cdot 10^{-2} \end{bmatrix},$$

and $J = \mathrm{diag}(1,1,1,-1)$, and $J_1 = \mathrm{diag}(1,1,-1,1)$, respectively. All subsequent quantities are displayed properly rounded to the given number of decimal places. Further, let $Y_G = G^* X$ and $W_G = JG^* X_\perp$.

From (2.10) we find that $\sin\phi = 7.4337 \cdot 10^{-6}$ ($\|K_S\| = 7.4337 \cdot 10^{-6}$). Recall that $U = G^* X_{\perp,1} |\Lambda_2|^{-\frac{1}{2}}$ and $Y = G^* X_2 |\Lambda_1|^{-\frac{1}{2}}$, which give $\|U\| \approx 1$ and $\|Y\| \approx 1$. Now from (2.9) and (2.4)–(2.5) it follows that for some $j_0 \in \{1,2,3\}$ holds

$$\frac{\lambda_{j_0} - 10^{-4}}{\sqrt{|\lambda_{j_0}| \cdot 10^{-4}}} < 7.43 \cdot 10^{-6}.$$

Using the above bound and by taking $\beta = 0.5$ we can assume that conditions from (3.5) are satisfied. Now, bound (3.7) from Theorem 3.1 yields

$$\frac{\lambda_{j_0} - 10^{-4}}{\lambda_{j_0}} < 1.1 \cdot 10^{-10}. \tag{3.13}$$

Note that $G_1 = GV$, with

$$V = \begin{bmatrix} 0.96120 & 0.27585 & -1.4764 \cdot 10^{-5} & 7.5213 \cdot 10^{-7} \\ 0.27585 & -0.96120 & 1.7694 \cdot 10^{-7} & -1.5787 \cdot 10^{-6} \\ 2.8748 \cdot 10^{-7} & 1.7249 \cdot 10^{-6} & -6.1662 \cdot 10^{-12} & 1 \\ 1.4142 \cdot 10^{-5} & 4.2426 \cdot 10^{-6} & 1 & 5.2180 \cdot 10^{-12} \end{bmatrix}.$$

The matrix V satisfies $V^* J V = J_1$ and $V J_1 V^* = J$, and $\|V\| = 1.00000175$. Now, this and bound (2.18) insure that $\sin\phi_1 \approx \sin\phi$.

Remark 1. We would like to point out that one can apply bound (3.5) on matrix

$$H_H = \begin{bmatrix} X & X_\perp \end{bmatrix}^* H \begin{bmatrix} X & X_\perp \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 10^{-4} & 0 & 10^{-9} & 10^{-10} & 0 \\ 0 & 0 & 10^{-4} & 3 \cdot 10^{-10} & 6 \cdot 10^{-10} & 0 \\ 0 & 10^{-9} & 3 \cdot 10^{-10} & 2 \cdot 10^{-4} & 0 & 0 \\ 0 & 10^{-10} & 6 \cdot 10^{-10} & 0 & -10^{-3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Since $\|E\| = 1.09 \cdot 10^{-9}$, and the spectral gap between the spectra of $H_H(1:3,1:3)$ and $H_H(4:6,4:6)$ is $\eta = 10^{-4}$, the relative version of the bound (3.5) yields:

$$\frac{|10^{-4} - \widetilde{\lambda}_j|}{10^{-4}} \leq \frac{1}{10^{-4}} \cdot \frac{2\|R_0\|^2}{\eta + \sqrt{\eta^2 + 4\|R_0\|^2}} = 1.19 \cdot 10^{-10}\,, \qquad j = 1,2,3,$$

which is similar to (3.13). Anyway, if one has only the factors of the matrix H (for example G and J), and X and $X_\perp$, then bound (3.13) still holds and at the same time bound (3.5) cannot be applied without forming the matrix H_H.

Acknowledgments. I would like to thank the anonymous referees for very careful reading of the manuscript, valuable comments, and for calling my attention to [3].

REFERENCES

[1] 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.
[2] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1990.
[3] C. K. Li and R. C. Li, *A note on eigenvalues of perturbed H Hermitian matrices*, Linear Algebra Appl., 395 (2005), pp. 183–190.
[4] B. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
[5] R. Mathias, *Quadratic residual bounds for the Hermitian eigenvalue problem*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 541–550.
[6] J. G. Sun, *Eigenvalues of Rayleigh quotient matrices*, Numer. Math., 59 (1991), pp. 603–614.
[7] N. Truhar and R. C. Li, *A* $\sin 2\Theta$ *theorem for graded indefinite Hermitian matrices*, Linear Algebra Appl., 359 (2003), pp. 263–276.
[8] N. Truhar and I. Slapničar, *Relative perturbation bound for invariant subspaces of graded indefinite Hermitian matrices*, Linear Algebra Appl., 301 (1999) pp. 171–185.
[9] N. Truhar and I. Slapničar, *Relative residual bounds for indefinite hermitian matrices*, Linear Algebra Appl., to appear.
[10] K.Veselić and I. Slapničar, *Floating–point perturbations of Hermitian matrices*, Linear Algebra Appl., 195 (1993), pp. 81–116.
[11] P. Å. Wedin, *On angles between subspaces of a finite dimensional inner product space*, in Matrix Pencils, Lecture Notes in Math. 973, B. Kågström and A. Ruhe, eds, Springer, New York, 1983, pp. 263–285.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 4, pp. 961–970

COMPUTATION OF SMALLEST EIGENVALUES IN THE STURM–LIOUVILLE PROBLEM WITH STRONGLY VARYING COEFFICIENTS*

ALEXANDER N. MALYSHEV†

Abstract. An efficient method is developed for computation of eigenvalues and eigenvectors with high relative precision in the Sturm–Liouville problem with strongly varying coefficients. Accuracy of the method is independent of the traditional condition number. New structured condition numbers for nonmultiple eigenvalues are introduced.

Key words. Sturm–Liouville problem, eigenvalue problem, tridiagonal matrix, relative precision, condition number

AMS subject classification. 65F15

DOI. 10.1137/050628131

1. Introduction. Computed approximations $\mathrm{fl}(\lambda)$ to an eigenvalue λ of a self-adjoint matrix A usually satisfy the estimate

$$|\mathrm{fl}(\lambda) - \lambda| \le O(\epsilon_{\mathrm{machine}})\|A\|.$$

When $|\lambda| \ll \|A\|$, computation of the eigenvalue λ with relative precision is generally not feasible without a computer arithmetic of multiple precision. If, for instance, the traditional condition number $\|A^{-1}\|\|A\|$ of a self-adjoint matrix A is of order $O(\epsilon_{\mathrm{machine}}^{-1})$, then its eigenvalue of smallest absolute value can be computed only with poor relative precision in the standard floating point arithmetic.

However, the multiple precision arithmetic can be avoided for some structured eigenvalue problem by using clever discretizations and matrix methods that respect the structure. A widely known example is the computation of singular values of bidiagonal matrices.

In the present work we study one general family of structured eigenvalue problems, a discrete Sturm–Liouville problem, whose eigenvalues we compute with high relative precision.

Throughout the paper we denote by $|A|$ the matrix with the entries $|a_{ij}|$, where A is a matrix with entries a_{ij}.

2. The regular Sturm–Liouville eigenvalue problem. Sturm–Liouville problems are concerned with solutions of the linear, homogeneous, ordinary differential equation

$$-(p(x)y'(x))' + q(x)y(x) = \lambda w(x)y(x), \qquad x \in (a,b), \tag{2.1}$$

where (a,b) is an open interval of the real line $\mathbb{R}$, p, q, w are real-valued coefficients defined on (a,b), and λ is a spectral parameter [14]. We treat only the regular case, where $p(x)$, $q(x)$ and $w(x)$ are sufficiently smooth functions on a finite closed interval

*Received by the editors March 30, 2005; accepted for publication (in revised form) by B. Parlett October 17, 2005; published electronically December 18, 2006.
http://www.siam.org/journals/simax/28-4/62813.html
†Department of Mathematics, University of Bergen, Johannes Brunsgate 12, N-5008 Bergen, Norway (sasha@mi.uib.no).

$[a,b]$. Moreover, we limit ourselves by self-adjoint eigenvalue problems and, therefore, do not admit sign changes in $w(x)$. For certainty, assume that

$$w(x) > 0 \quad \text{for all } x \in [a,b]. \tag{2.2}$$

The boundary conditions are separated (not coupled) and self-adjoint. Namely, at the boundary $x = a$ the solution $y(x)$ satisfies either

$$y(a) = 0 \tag{2.3}$$

or

$$p(a)y'(a) - \gamma_a y(a) = 0, \qquad \gamma_a \in \mathbb{R}. \tag{2.4}$$

Similarly, at $x = b$ we impose the conditions

$$y(b) = 0 \tag{2.5}$$

or

$$p(b)y'(b) + \gamma_b y(b) = 0, \qquad \gamma_b \in \mathbb{R}. \tag{2.6}$$

Let us introduce self-adjoint operators $\mathcal{A}$ and $\mathcal{B}$ by means of the bilinear forms

$$(\mathcal{B}y, \phi) = \int_a^b w(x)y(x)\phi(x)dx \quad \forall \phi(x) \in L^2[a,b], \tag{2.7}$$

$$\begin{aligned}(\mathcal{A}y, \phi) = &\int_a^b [p(x)y'(x)\phi'(x) + q(x)y(x)\phi(x)]dx \\ &+ \gamma_a y(a)\phi(a) + \gamma_b y(b)\phi(b) \quad \forall \phi(x) \in H^1[a,b].\end{aligned} \tag{2.8}$$

When $y(a) = 0$, the test functions ϕ must satisfy $\phi(a) = 0$, and γ_a in (2.8) is set to 0. Analogously, $\phi(b) = 0$ and $\gamma_b = 0$ when $y(b) = 0$. Note that the operator $\mathcal{B}$ takes $L^2[a,b]$ to $L^2[a,b]$ and is positive definite owing to (2.2). The operator $\mathcal{A} : L^2[a,b] \to L^2[a,b]$ is unbounded.

The Sturm–Liouville eigenvalue problem is equivalent to the generalized eigenvalue problem for the operator pencil $\mathcal{A} - \lambda\mathcal{B}$. It turns out that small relative perturbations of the coefficients p, q, w, γ_a and γ_b cause small relative perturbations of the eigenvalues of $\mathcal{A} - \lambda\mathcal{B}$. Our sophisticated numerical method inherits this property and computes the eigenvalues with high relative precision.

Below we sketch a relative perturbation theory for the continuous eigenvalue problem and introduce its condition number. Let λ and $y(x)$ be a nonzero isolated eigenvalue and associated eigenvector of the operator pencil $\mathcal{A} - \lambda\mathcal{B}$. Suppose that perturbations $\delta\mathcal{A}$ and $\delta\mathcal{B}$ are determined by relative perturbations of p, q, w, γ_a and γ_b such that

$$|\delta p| \le \epsilon|p|, \quad |\delta q| \le \epsilon|q|, \quad |\delta w| \le \epsilon|w|, \quad |\delta\gamma_a| \le \epsilon|\gamma_a|, \quad |\delta\gamma_b| \le \epsilon|\gamma_b| \tag{2.9}$$

with infinitely small $\epsilon > 0$. The perturbed eigenpair of the operator pencil $(\mathcal{A}+\delta\mathcal{A}) - \lambda(\mathcal{B} + \delta\mathcal{B})$ is denoted by $\lambda + \delta\lambda$ and $y + \delta y$. It is not difficult to derive the identity

$$\delta\lambda = \frac{(\delta\mathcal{A}y, y) - \lambda(\delta\mathcal{B}y, y)}{(\mathcal{B}y, y)}, \tag{2.10}$$

when ϵ is inifinitely small. Using (2.9) we can obtain the estimates

$$|(\delta\mathcal{A}y, y)| \le \epsilon \left[\int_a^b (|p|y'^2 + |q|y^2)dx + |\gamma_a|y^2(a) + |\gamma_b|y^2(b) \right],$$
$$|(\delta\mathcal{B}y, y)| \le \epsilon \int_a^b wy^2 dx = \epsilon(\mathcal{B}y, y),$$

which together with (2.10) give the estimate

$$\frac{|\delta\lambda|}{|\lambda|} \le \epsilon \left[\frac{\int_a^b (|p|y'^2 + |q|y^2)dx + |\gamma_a|y^2(a) + |\gamma_b|y^2(b)}{|\lambda|(\mathcal{B}y, y)} + 1 \right]. \tag{2.11}$$

Replacing $|\lambda|(\mathcal{B}y, y)$ by $|(\mathcal{A}y, y)|$ in (2.11) we arrive at the inequality

$$\frac{|\delta\lambda|}{|\lambda|} \le \epsilon \cdot \mathrm{relcond}(\lambda),$$

where

$$\mathrm{relcond}(\lambda) = \frac{\int_a^b (|p|y'^2 + |q|y^2)dx + |\gamma_a|y^2(a) + |\gamma_b|y^2(b)}{\left| \int_a^b (py'^2 + qy^2)dx + \gamma_a y^2(a) + \gamma_b y^2(b) \right|} + 1 \tag{2.12}$$

is the relative structured condition number of λ.

If, for instance, $p(x), q(x), \gamma_a, \gamma_b \ge 0$, then relcond $= 2$ for each isolated eigenvalue.

3. Structured finite-difference approximation. Equation (2.1) is discretized on the uniform grid $x_i = a + ih$, $i = 0\colon(n+1)$, $h = (b-a)/(n+1)$:

$$\frac{p(x_{i-1/2})(y_i - y_{i-1}) - p(x_{i+1/2})(y_{i+1} - y_i)}{h^2} + q(x_i)y_i = \lambda w(x_i)y_i, \qquad i = 1\colon n, \tag{3.1}$$

where $x_{i\pm1/2} = x_i \pm h/2$. The left boundary condition (2.3) is fulfilled with $y_0 = 0$. Similarly, the right boundary condition (2.5) is fulfilled with $y_{n+1} = 0$. The condition (2.4) may be discretized as

$$\frac{p(x_{1/2})}{h^2}(y_0 - y_1) + \left[\frac{q(x_0)}{2} + \frac{\gamma_a}{h} \right] y_0 = \lambda \frac{w(x_0)}{2} y_0,$$

and the condition (2.6) as

$$\frac{p(x_{n+1/2})}{h^2}(y_{n+1} - y_n) + \left[\frac{q(x_{n+1})}{2} + \frac{\gamma_b}{h} \right] y_{n+1} = \lambda \frac{w(x_{n+1})}{2} y_{n+1}.$$

All above approximations are second order accurate with respect to h.

Let us introduce diagonal matrices $P = \mathrm{diag}(p_1, \dots, p_{n+1})$, $Q = \mathrm{diag}(q_1, \dots, q_n)$, and $W = \mathrm{diag}(w_1, \dots, w_n)$ with the diagonal entries $p_i = p(x_{i-1/2})/h^2$, $q_i = q(x_i)$, $w_i = w(x_i)$, and bidiagonal matrix

$$B = B_n = \begin{pmatrix} a_1 & b_1 & & & \\ & a_2 & b_2 & & \\ & & \cdot & \cdot & \\ & & & a_n & b_n \end{pmatrix}, \tag{3.2}$$

where $a_i = 1$ and $b_i = -1$.

Henceforth we restrict ourselves only by the Dirichlet boundary conditions

$$y_0 = y_{n+1} = 0. \tag{3.3}$$

Then the vector $Y = [y_1, \dots, y_n]^T$ satisfies the generalized eigenvalue problem

$$BPB^T Y + QY = \lambda WY. \tag{3.4}$$

We emphasize that our numerical algorithm takes advantage of the special structure (3.4). The tridiagonal matrix $BPB^T + Q$ is never formed explicitly. Instead, the factorized representation (3.4) is used.

Remark. Other boundary conditions can be considered analogously. For example, when $y(a) = 0$ and $p(b)y'(b) + \gamma_b y(b) = 0$, the vector $Y_{n+1} = [y_1, \dots, y_{n+1}]^T$ satisfies $B_{n+1}P_{n+1}B_{n+1}^T Y_{n+1} + Q_{n+1}Y_{n+1} = \lambda W_{n+1}Y_{n+1}$ with $P_{n+1} = \operatorname{diag}(p_1, \dots, p_{n+1}, 0)$, $Q_{n+1} = \operatorname{diag}(q_1, \dots, q_n, q_{n+1}/2 + \gamma_b/h)$, $W_{n+1} = \operatorname{diag}(w_1, \dots, w_n, w_{n+1}/2)$.

Remark. Let x_i be a nonuniform grid and $h_i = x_i - x_{i-1}$. A direct modification of (3.1) with zero Dirichlet conditions

$$\frac{2}{h_i + h_{i+1}} \left[p(x_{i-1/2}) \frac{y_i - y_{i-1}}{h_i} - p(x_{i+1/2}) \frac{y_{i+1} - y_i}{h_{i+1}} \right] + q(x_i)y_i = \lambda w(x_i)y_i, \quad i = 1\colon n,$$

leads to the matrix equation $DB\hat{P}B^T Y + QY = \lambda WY$ with the diagonal matrices $D = \operatorname{diag}(\frac{2}{h_1+h_2}, \dots, \frac{2}{h_n+h_{n+1}})$ and $\hat{P} = \operatorname{diag}(p(x_{1/2})/h_1, \dots, p(x_{n+1/2})/h_{n+1})$. Multiplying this equation from the left by D^{-1} and denoting $\hat{Q} = D^{-1}Q$ and $\hat{W} = D^{-1}W$ we obtain the equation $B\hat{P}B^T Y + \hat{Q}Y = \lambda \hat{W}Y$ of the form (3.4). We do not pursue the case of nonuniform grids further in this paper because our primary goal is to reveal the algebraic structure of discretizations which are suitable for floating point computations corresponding to the perturbation theory from the previous section.

4. Relative condition numbers for (3.4). We consider the matrix pencil $BPB^T + Q - \lambda W$, where P, Q and W are diagonal matrices, W is positive definite, and B is the n-by-$n+1$ bidiagonal matrix (3.2). Let λ and v be an eigenpair of this pencil such that the eigenvalue λ is nonmultiple. Assume that the diagonal entries p_i, q_i and w_i of P, Q and W are subject to relative perturbations

$$|\delta p_i| \le \epsilon |p_i|, \quad |\delta q_i| \le \epsilon |q_i|, \quad |\delta w_i| \le \epsilon |w_i|, \tag{4.1}$$

where $\epsilon > 0$ is infinitely small. It is easy to derive that the perturbed eigenvalue $\lambda + \delta\lambda$ satisfies the identity

$$\delta\lambda = \frac{([B\delta P B^T + \delta Q]v, v) - \lambda(\delta W v, v)}{(Wy, y)}, \tag{4.2}$$

where $\delta P = \operatorname{diag}(\delta p_1, \dots, \delta p_{n+1})$, $\delta Q = \operatorname{diag}(\delta q_1, \dots, \delta q_n)$, $\delta W = \operatorname{diag}(\delta w_1, \dots, \delta w_n)$. The following estimate holds owing to (4.1):

$$|\delta\lambda| \le \epsilon \frac{([B|P|B^T + |Q|]v, v) + |\lambda|(Wv, v)}{(Wv, v)}. \tag{4.3}$$

Taking into account the identity $|\lambda|(Wv, v) = |([BPB^T + Q]v, v)|$ we obtain that

$$|\delta\lambda| \le \epsilon |\lambda| \left\{ \frac{([B|P|B^T + |Q|]v, v)}{|([BPB^T + Q]v, v)|} + 1 \right\}. \tag{4.4}$$

Thus the relative structured condition number similar to (2.12) for an isolated eigenvalue λ of (3.4) equals

$$\text{relcond}(\lambda) = \frac{v^T B|P|B^T v + v^T|Q|v}{|v^T BPB^T v + v^T Qv|} + 1, \tag{4.5}$$

where v is an eigenvector corresponding to λ. Relative structured condition numbers of type (4.5) probably first appeared in [5].

In section 6 we will need slightly more general perturbed pencils. Namely, consider a perturbed matrix pencil of the form

$$(I+\Delta)B(P+\delta P)B^T(I+\Delta) + (Q+\delta Q) - \lambda(W+\delta W), \tag{4.6}$$

where δP, δQ and δW are diagonal and their diagonal entries satisfy the inequalities (4.1). The diagonal matrix Δ also satisfies the inequality $\|\Delta\|_2 \le \epsilon$. Applying (4.4) to the pencil

$$B(P+\delta P)B^T + (I+\Delta)^{-2}(Q+\delta Q) - \lambda(I+\Delta)^{-2}(W+\delta W),$$

which has the same eigenvalues as (4.6), we arrive at the estimate

$$|\delta\lambda| \le 3\epsilon|\lambda|\text{relcond}(\lambda). \tag{4.7}$$

5. Relatively stable LDL^T and UDU^T factorizations. In this section we study the LDL^T and UDU^T factorizations for the structured symmetric tridiagonal matrix

$$BPB^T + Q - \mu W, \tag{5.1}$$

where B is an $n \times (n+1)$ bidiagonal matrix as in (3.2), P is an $(n+1)\times(n+1)$ diagonal matrix, Q and W are $n \times n$ diagonal matrices, and μ is a scalar parameter.

The identity $BPB^T + Q - \mu W = LDL^T$, where

$$D = \begin{pmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_n \end{pmatrix}, \qquad L = \begin{pmatrix} 1 & & & \\ l_1 & 1 & & \\ & \ddots & \ddots & \\ & & l_{n-1} & 1 \end{pmatrix}$$

is equivalent to the equalities $d_i + d_{i-1}l_{i-1}^2 = p_i a_i^2 + p_{i+1}b_i^2 + q_i - \mu w_i$ and $d_i l_i = p_{i+1}a_{i+1}b_i$, hence the straightforward algorithm:

$$\begin{aligned} &d_1 = p_1 a_1^2 + p_2 b_1^2 + q_1 - \mu w_1 \\ &\textbf{for } i = 1\colon n-1 \\ &\quad l_i = p_{i+1}a_{i+1}b_i/d_i \\ &\quad d_{i+1} = -d_i l_i^2 + p_{i+1}a_{i+1}^2 + p_{i+2}b_{i+1}^2 + q_{i+1} - \mu w_{i+1} \\ &\textbf{end}. \end{aligned} \tag{5.2}$$

By introducing an auxiliary variable $s_i = d_i - p_{i+1}b_i^2$ the algorithm (5.2) is modified as follows:

$$\begin{aligned} &s_1 = p_1 a_1^2 + q_1 - \mu w_1 \\ &\textbf{for } i = 1\colon n-1 \\ &\quad d_i = s_i + p_{i+1}b_i^2 \\ &\quad t_i = p_{i+1}a_{i+1}/d_i \\ &\quad l_i = t_i b_i \\ &\quad s_{i+1} = t_i a_{i+1} s_i + q_{i+1} - \mu w_{i+1} \\ &\textbf{end} \\ &d_n = s_n + p_{n+1}b_n^2. \end{aligned} \tag{5.3}$$

The algorithm (5.3) is equivalent to (5.2) in exact arithmetic but very different in a floating point arithmetic. Roughly speaking, the effects of rounding errors for (5.3) can be interpreted as relative perturbations of the coefficients p_i, q_i and w_i. Numerical experiments with the examples from section 9 showed that the algorithm (5.2) had no such properties and computed wrong results.

Variants of modification (5.3) were almost simultaneously discovered by Babuška, Degtyarev, Favorskii, and Rutishauser (see [2],[3],[1],[12],[13]). These were mainly aimed at solving tridiagonal systems of linear equations with strongly varying coefficients. Dhillon, Fernando, and Parlett later applied such modifications to eigenvalue problems in [8], [5].

In a perfectly similar manner, the factorization $BPB^T+Q-\mu W = UDU^T$, where

$$D = \begin{pmatrix} d_1 & & & \\ & d_2 & & \\ & & . & \\ & & & d_n \end{pmatrix}, \qquad U = \begin{pmatrix} 1 & u_1 & & \\ & 1 & . & \\ & & . & u_{n-1} \\ & & & 1 \end{pmatrix}$$

leads to the equalities $d_i + d_{i+1}u_i^2 = p_i a_i^2 + p_{i+1}b_i^2 + q_i - \mu w_i$ and $d_{i+1}u_i = p_{i+1}a_{i+1}b_i$, and the straightforward algorithm

$$\begin{array}{l} d_n = p_n a_n^2 + p_{n+1}b_n^2 + q_n - \mu w_n \\ \textbf{for } i = n-1\colon -1\colon 1 \\ \quad u_i = p_{i+1}a_{i+1}b_i/d_{i+1} \\ \quad d_i = -d_{i+1}u_i^2 + p_i a_i^2 + p_{i+1}b_i^2 + q_i - \mu w_i \\ \textbf{end} \end{array} \tag{5.4}$$

is modified by introducing the auxiliary variable $s_i = d_i - p_i a_i^2$:

$$\begin{array}{l} s_n = p_{n+1}b_n^2 + q_n - \mu w_n \\ \textbf{for } i = n-1\colon -1\colon 1 \\ \quad d_{i+1} = s_{i+1} + p_{i+1}a_{i+1}^2 \\ \quad t_{i+1} = p_{i+1}b_i/d_{i+1} \\ \quad u_i = t_{i+1}a_{i+1} \\ \quad s_i = t_{i+1}b_i s_{i+1} + q_i - \mu w_i \\ \textbf{end} \\ d_1 = s_1 + p_1 a_1^2. \end{array} \tag{5.5}$$

The algorithm (5.3) may be called the relatively stable LDL^T factorization of (5.1) and algorithm (5.5) the relatively stable UDU^T factorization of (5.1). These names are due to the mixed relative stability that was first observed in [1] and then more accurately formulated and proved in [8]. We discuss the mixed relative stability in the next section.

6. Mixed relative stability of the relatively stable LDL^T factorization. The computed values of a variable a are denoted by $\mathrm{fl}(a)$ or $\tilde{a}$. The rounding errors for basic arithmetic operations are assumed to satisfy the common model

$$\mathrm{fl}(a \circ b) = (a \circ b)(1+\delta), \text{ where } |\delta| \le \epsilon = \epsilon_{\text{machine}}. \tag{6.1}$$

By the aid of (6.1) we model effects of rounding errors in (5.3) as follows:

$$
\begin{array}{rcll}
\tilde{d}_i & = & \tilde{s}_i(1+\delta_{1,i}) + p_{i+1}b_i^2(1+\delta_{2,i}), & |\delta_{1,i}| \le \epsilon, \\
 & & & |\delta_{2,i}| \le 3\epsilon + O(\epsilon^2), \\
\tilde{t}_i & = & (p_{i+1}a_{i+1}/\tilde{d}_i)(1+\delta_{3,i}), & |\delta_{3,i}| \le 2\epsilon + O(\epsilon^2), \\
\tilde{l}_i & = & \tilde{t}_i b_i(1+\delta_{4,i}), & |\delta_{4,i}| \le \epsilon, \\
\tilde{s}_{i+1} & = & \tilde{t}_i a_{i+1}\tilde{s}_i(1+\delta_{5,i}) & |\delta_{5,i}| \le 4\epsilon + O(\epsilon^2), \\
 & & + q_{i+1}(1+\delta_{6,i}) - \mu w_{i+1}(1+\delta_{7,i}), & |\delta_{6,i}| \le 2\epsilon + O(\epsilon^2), \\
 & & & |\delta_{7,i}| \le 2\epsilon + O(\epsilon^2).
\end{array} \tag{6.2}
$$

Let us eliminate $\tilde{t}_i$ from these formulas and regroup them:

$$
\begin{aligned}
\tilde{d}_i &= \tilde{s}_i(1+\delta_{1,i}) + p_{i+1}b_i^2(1+\delta_{2,i}), \\
\tilde{s}_{i+1}(1+\delta_{1,i+1}) &= \frac{p_{i+1}a_{i+1}^2}{\tilde{d}_i}\tilde{s}_i(1+\delta_{1,i})\frac{(1+\delta_{1,i+1})(1+\delta_{3,i})(1+\delta_{5,i})}{(1+\delta_{1,i})} \\
&\quad + q_{i+1}(1+\delta_{1,i+1})(1+\delta_{6,i}) - \mu w_{i+1}(1+\delta_{1,i+1})(1+\delta_{7,i}), \\
\tilde{l}_i &= \frac{p_{i+1}a_{i+1}b_i}{\tilde{d}_i}(1+\delta_{3,i})(1+\delta_{4,i}).
\end{aligned}
$$

Introducing new variables (which are not computed!)

$$
\begin{array}{ll}
\hat{s}_i = \tilde{s}_i(1+\delta_{1,i}), & \\
\hat{b}_i = b_i\sqrt{1+\delta_{2,i}}, & \hat{a}_{i+1} = a_{i+1}\sqrt{\dfrac{(1+\delta_{1,i+1})(1+\delta_{3,i})(1+\delta_{5,i})}{(1+\delta_{1,i})}}, \\
\hat{q}_{i+1} = q_{i+1}(1+\delta_{1,i+1})(1+\delta_{6,i}), & \hat{w}_{i+1} = w_{i+1}(1+\delta_{1,i+1})(1+\delta_{7,i}),
\end{array} \tag{6.3}
$$

we arrive at the following identities:

$$
\begin{aligned}
\tilde{d}_i &= \hat{s}_i + p_{i+1}\hat{b}_i^2, \\
\hat{s}_{i+1} &= \frac{p_{i+1}\hat{a}_{i+1}^2}{\tilde{d}_i}\hat{s}_i + \hat{q}_{i+1} - \mu\hat{w}_{i+1}, \\
\hat{l}_i &= \frac{p_{i+1}\hat{a}_{i+1}\hat{b}_i}{\tilde{d}_i}, \\
\tilde{l}_i &= \hat{l}_i(1+\delta_{3,i})(1+\delta_{4,i})\sqrt{\frac{(1+\delta_{1,i})}{(1+\delta_{1,i+1})(1+\delta_{2,i})(1+\delta_{3,i})(1+\delta_{5,i})}}.
\end{aligned} \tag{6.4}
$$

Moreover,

$$
\begin{array}{ll}
|\hat{a}_i - a_i| \le |a_i|[4\epsilon + O(\epsilon^2)], & |\hat{b}_i - b_i| \le |b_i|[\frac{3}{2}\epsilon + O(\epsilon^2)], \\
|\hat{q}_i - q_i| \le |q_i|[3\epsilon + O(\epsilon^2)], & |\hat{w}_i - w_i| \le |w_i|[3\epsilon + O(\epsilon^2)], \\
|\tilde{l}_i - \hat{l}_i| \le |\hat{l}_i|[\frac{17}{2}\epsilon + O(\epsilon^2)]. &
\end{array} \tag{6.5}
$$

Let us denote $\tilde{D} = \operatorname{diag}(\tilde{d}_1, \ldots, \tilde{d}_n)$, $\tilde{Q} = \operatorname{diag}(\tilde{q}_1, \ldots, \tilde{q}_n)$, $\tilde{W} = \operatorname{diag}(\tilde{w}_1, \ldots, \tilde{w}_n)$. The bidiagonal matrix $\hat{B}$ is formed from B by replacing a_i with $\hat{a}_i$ and b_i with $\hat{b}_i$. The unit lower triangular matrix $\hat{L}$ is formed from L by replacing l_i with $\hat{l}_i$. In this notation we obtain the identity

$$
\hat{L}\tilde{D}\hat{L}^T = \hat{B}P\hat{B}^T + \tilde{Q} - \mu\tilde{W} \tag{6.6}
$$

from formulas (6.4). The structure of (6.6) is not completely satisfactory for the conditioning theory presented in section 4 because of the availability of the perturbed matrix $\hat{B}$. However, this obstacle is easily removed as shown below.

We recall that bidiagonal matrices are perfect for dealing with relative perturbations of their elements because such perturbations can be interpreted as multiplicative [4, Lemma 5.7]. In our case, this means there exist diagonal matrices Δ and Δ_1 such that $|\Delta| = O(\epsilon)$, $|\Delta_1| = O(\epsilon)$, and $\hat{B} = (I+\Delta)B(I+\Delta_1)$. Thus, (6.6) can be replaced by the representation

$$\hat{L}\hat{D}\hat{L}^T = (I+\Delta)B(I+\Delta_1)P(I+\Delta_1)B^T(I+\Delta) + \tilde{Q} - \mu\tilde{W}. \tag{6.7}$$

Equation (6.7) is now of the form (4.6) with $P+\delta P = (I+\Delta_1)P(I+\Delta_1)$, $Q+\delta Q = \tilde{Q}$, and $W + \delta W = \tilde{W}$. Therefore, estimates of the type (4.7) hold in computations of eigenvalues based on relatively stable LDL^T factorizations.

An analogous result holds for the algorithm (5.5).

7. Computation of eigenvalues. The well-known *bisection* method [4] uses only signs of the diagonal entries of $\mathrm{fl}(D) = \tilde{D}$ computed by the algorithm (5.3). In the previous section we have demonstrated that the signs of $\tilde{d}_i$ depend on P, Q, W in a backward stable manner. Therefore, if an eigenvalue has a low relative structured condition number (4.5), then it can be computed with a high relative accuracy. A faster but more delicate alternative to the bisection is the inverse iteration with shifts.

8. Eigenvectors from twisted factorizations. The twisted (or double) factorization of a tridiagonal matrix T at row r is the factorization $T = Z_r D_r Z_r^T$, where $D_r = \mathrm{diag}\{d_1, \ldots, d_{r-1}, \gamma_r, \check{d}_{r+1}, \ldots, \check{d}_n\}$ is diagonal and

$$Z_r = \begin{pmatrix} 1 & & & & & & & \\ l_1 & 1 & & & & & & \\ & \cdot & \cdot & & & & & \\ & & l_{r-1} & 1 & u_r & & & \\ & & & & 1 & \cdot & & \\ & & & & & \cdot & u_{n-1} & \\ & & & & & & 1 & \end{pmatrix}. \tag{8.1}$$

Assume that $T = LDL^T$ is the LDL^T factorization and $T = U\check{D}U^T$ is the UDU^T factorization. One can show that the elements $l_1, \ldots, l_{r-1}$, $u_r, \ldots, u_{n-1}$, $d_1, \ldots, d_{r-1}$, $\check{d}_{r+1}, \ldots, \check{d}_n$ from the twisted factorization coincide with the corresponding entries of L, U, D, and $\check{D}$. Moreover, the following representations hold for the parameter γ_r:

$$\gamma_r = d_r - T_{rr} + \check{d}_r = d_r - u_r T_{r+1,r}. \tag{8.2}$$

Let e_r be the rth unit coordinate vector. The identity $Z_r e_r = e_r$ implies that $e_r^T T^{-1} e_r = 1/\gamma_r$, and hence $\mathrm{trace}(T^{-1}) = \sum_{r=1}^{n} \frac{1}{\gamma_r}$. If the matrix T is nearly singular, then $\mathrm{trace}(T^{-1})$ is large in most cases. In these cases, $|\gamma_r|$ is tiny for at least one r and the vector

$$v_r = Z_r^{-T} e_r \tag{8.3}$$

satisfies the equation $Tv_r = \gamma_r e_r \approx 0$, i.e., v_r is an approximate singular vector of T corresponding to a nearly zero singular value.

When generating the twisted factorizations of a nearly singular matrix $T = BPB^T + Q - \tilde{\lambda}W$ from the results of (5.3) and (5.5), where $\tilde{\lambda}$ is an accurate approximation to an eigenvalue of the matrix pencil $BPB^T + Q - \lambda W$, the value γ_r is stably computed by each of the two formulas:

$$\gamma_r = d_r - \frac{(p_{r+1}a_{r+1}b_r)^2}{\check{d}_{r+1}} = s_r + \frac{p_{r+1}b_r^2\check{s}_{r+1}}{\check{s}_{r+1} + p_{r+1}a_{r+1}^2}. \tag{8.4}$$

One of the first uses of the twisted factorizations was reported in [10]. Babuška systematically applied them in [1] to the solution of linear systems with tridiagonal matrices and carried out a rather complete rounding error analysis. Godunov with coauthors [9] pioneered the use of the twisted factorizations in reliable computation of eigenvectors of symmetric tridiagonal matrices. Unaware of [9], Fernando in [7] proposed his own variant of computation of an eigenvector of a symmetric tridiagonal matrix by the twisted factorizations. Parlett and Dhillon further developed Fernando's method in [11].

We do not address hard problems with maintaining orthogonality between the computed eigenvectors corresponding to multiple or almost multiple eigenvalues of T here and refer the reader to, e.g., [5],[6]. By the way, the orthogonality between eigenvectors does not hold for matrix pencils with $W \neq \text{const} \cdot I$.

9. Numerical examples. In order to find several smallest eigenvalues and corresponding eigenvectors of (2.1) with high relative accuracy, one has to carry out the following steps:

1. Form the diagonal matrices P, Q, W and bidiagonal B from (3.4).
2. Use (5.3) and the bisection method to compute the desired eigenvalues.
3. For each approximate eigenvalue $\tilde{\lambda}$, run the algorithms (5.3) and (5.5) with $\mu = \tilde{\lambda}$ and compute γ_r by (8.4) for all r. Then choose r for which $|\gamma_r|$ is smallest and compute the eigenvector by (8.3).

Below we demonstrate the power of this procedure in computation of the smallest eigenvalues and associated eigenvectors for very ill-conditioned tridiagonal matrices of the form $T = BPB^T + Q$.

Positive definite matrix T. Consider the Sturm–Liouville problem on the interval $[a, b] = [-1, 1]$ divided into N equal subintervals with

$$p(x) = \exp(-200x^2), \quad q(x) = 0, \quad w(x) = 1, \quad y(-1) = y(1) = 0. \tag{9.1}$$

In this example we have computed the smallest six eigenvalues.

N	λ_1	λ_2	λ_3
10^4	5.544029528624667e−85	8.056984043120979e−82	8.091396479595924e−82
10^5	5.542569829437562e−85	8.060008044877117e−82	8.094455541536701e−82
10^6	5.542555233653751e−85	8.060038353699461e−82	8.094486201448963e−82

N	λ_4	λ_5	λ_6
10^4	2.694830737112778e−81	2.701042250989879e−81	5.651547554516112e−81
10^5	2.699895141887398e−81	2.706131801258254e−81	5.675612519454272e−81
10^6	2.699945802766178e−81	2.706182713474177e−81	5.675852827038452e−81

For $N = 10^6$, all relative structured condition numbers (4.5) are ≈ 2, all residuals $\|Tv - \lambda v\|$ were $\leq 2e{-}93$, and the computed eigenvectors $v_1, v_2, v_3, v_4, v_5, v_6$ were mutually orthogonal with high accuracy. Note that $\|T\|_2 \approx 10^{12}$ for $N = 10^6$.

The eigenvalue functions, represented by the vectors $v_1, \ldots, v_6$, have very narrow boundary layers.

Indefinite matrix T. The smallest 12 eigenvalues of the Sturm–Liouville problem on $[a, b] = [-1, 1]$ divided into 10^6 equal subintervals with

$$p(x) = \exp(-200x^2),\ q(x) = -10^{-80}\cos^2(\pi x),\ w(x) = 1,\ y(-1) = y(1) = 0, \tag{9.2}$$

are given in this table.

λ_1	$-9.193288419740126e{-}81$	λ_7	$-4.246722436372906e{-}81$
λ_2	$-9.192636604107326e{-}81$	λ_8	$-2.632481764148639e{-}82$
λ_3	$-7.304786310492325e{-}81$	λ_9	$-2.386685734031909e{-}82$
λ_4	$-7.297679753112374e{-}81$	λ_{10}	$4.876318269699302e{-}81$
λ_5	$-5.001049216049395e{-}81$	λ_{11}	$4.898522180802763e{-}81$
λ_6	$-4.321230312552760e{-}81$	λ_{12}	$1.109749645012289e{-}80$

The relative structured condition numbers (4.5) for $\lambda_1, \ldots, \lambda_{12}$ are $2.2, 2.2, 2.7, 2.7, 2.2, 4.6, 4.5, 76, 84, 6.1, 6.1, 3.8$, the residuals $\|Tv - \lambda v\|$ were $\leq 9e{-}95$, and the eigenvectors $v_1, v_2, \ldots, v_{12}$ were mutually orthogonal with high accuracy.

REFERENCES

[1] I. Babuška, *Numerical stability in problems of linear algebra*, SIAM J. Numer. Anal., 9 (1972), pp. 53–77.

[2] L. M. Degtyarev and A. P. Favorskii, *A flow variant of the sweep method*, U.S.S.R. Comput. Math. and Math. Phys., 8 (1968), pp. 252–261.

[3] L. M. Degtyarev and A. P. Favorskii, *Flow variant of the sweep method for difference problems with strongly varying coefficients*, U.S.S.R. Comput. Math. and Math. Phys., 9 (1969), pp. 285–294.

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

[5] I. S. Dhillon, *A New $O(N^2)$ Algorithm for the Symmetric Tridiagonal Eigenvalue/Eigenvector Problem*, Ph.D. thesis, University of California, Berkeley, CA, 1997.

[6] I. S. Dhillon and B. N. Parlett, *Orthogonal eigenvectors and relative gaps*, SIAM J. Matrix. Anal. Appl., 25 (2004), pp. 858–899.

[7] K. V. Fernando, *On computing an eigenvector of a tridiagonal matrix. Part* I: *Basic results*, SIAM J. Matrix. Anal. Appl., 18 (1997), pp. 1013–1034.

[8] K. V. Fernando and B. N. Parlett, *Accurate singular values and differential qd algorithms*, Numer. Math., 67 (1994), pp. 191–229.

[9] S. K. Godunov, V. I. Kostin, and A. D. Mitchenko, *Computation of an eigenvalue of symmetric tridiagonal matrices*, Siberian Math. J., 26 (1985), pp. 71–85.

[10] P. Henrici, *Bounds for eigenvalues of certain tridiagonal matrices*, J. Soc. Indust. Appl. Math., 11 (1963), pp. 281–290.

[11] B. N. Parlett and I. S. Dhillon, *Fernando's solution to Wilkinson's problem: An application of double factorization*, Linear Algebra Appl., 267 (1997), pp. 247–279.

[12] H. Rutishauser, *Vorlesungen über Numerische Mathematik*, Birkhäuser, Basel, Switzerland, 1976.

[13] A. A. Samarskii, *The Theory of Difference Schemes*, 2nd ed., Nauka, Moscow, 1989 (in Russian).

[14] A. Zettl, *Sturm-Liouville problems*, in Spectral Theory and Computational Methods of Sturm-Liouville problems, Lecture Notes in Pure and Appl. Math. 191, D. Hinton and P. Schaefer, eds., Marcel Dekker, New York, 1997, pp. 1–104.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 4, pp. 971–1004

VECTOR SPACES OF LINEARIZATIONS FOR MATRIX POLYNOMIALS*

D. STEVEN MACKEY†, NILOUFER MACKEY‡, CHRISTIAN MEHL§, AND VOLKER MEHRMANN§

Abstract. The classical approach to investigating polynomial eigenvalue problems is linearization, where the polynomial is converted into a larger matrix pencil with the same eigenvalues. For any polynomial there are infinitely many linearizations with widely varying properties, but in practice the companion forms are typically used. However, these companion forms are not always entirely satisfactory, and linearizations with special properties may sometimes be required.

Given a matrix polynomial P, we develop a systematic approach to generating large classes of linearizations for P. We show how to simply construct two vector spaces of pencils that generalize the companion forms of P, and prove that almost all of these pencils are linearizations for P. Eigenvectors of these pencils are shown to be closely related to those of P. A distinguished subspace is then isolated, and the special properties of these pencils are investigated. These spaces of pencils provide a convenient arena in which to look for structured linearizations of structured polynomials, as well as to try to optimize the conditioning of linearizations.

Key words. matrix polynomial, matrix pencil, linearization, strong linearization, shifted sum, companion form

AMS subject classifications. 65F15, 15A18, 15A22

DOI. 10.1137/050628350

1. Introduction.

Polynomial eigenvalue problems $P(\lambda)x = 0$, where $P(\lambda) = \sum_{i=0}^{k} \lambda^i A_i$ with real or complex coefficient matrices A_i, form the basis for (among many other applications) the vibration analysis of buildings, machines, and vehicles [5], [9], [21], and numerical methods for the solution of these problems are incorporated into most commercial and noncommercial software packages for structural analysis.

The classical and most widely used approach to solving polynomial eigenvalue problems is *linearization*, i.e., the conversion of $P(\lambda)x = 0$ into a larger size linear eigenvalue problem $L(\lambda)z = (\lambda X + Y)z = 0$ with the same eigenvalues, so that classical methods for linear eigenvalue problems can be pressed into service. The linearizations most commonly commissioned are the companion forms for $P(\lambda)$, one of which is

$$L(\lambda) = \lambda \begin{bmatrix} A_k & 0 & \cdots & 0 \\ 0 & I_n & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & I_n \end{bmatrix} + \begin{bmatrix} A_{k-1} & A_{k-2} & \cdots & A_0 \\ -I_n & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & -I_n & 0 \end{bmatrix}. \tag{1.1}$$

*Received by the editors April 1, 2005; accepted for publication (in revised form) by I. C. F. Ipsen November 16, 2005; published electronically December 18, 2006. This work was supported by the Deutsche Forschungsgemeinschaft through MATHEON, the DFG Research Center *Mathematics for key technologies* in Berlin.

http://www.siam.org/journals/simax/28-4/62835.html

†School of Mathematics, The University of Manchester, Sackville Street, Manchester, M60 1QD, UK (smackey@ma.man.ac.uk). The work of this author was supported by Engineering and Physical Sciences Research Council grant GR/S31693.

‡Department of Mathematics, Western Michigan University, Kalamazoo, MI 49008 (nil.mackey@wmich.edu; http://homepages.wmich.edu/~mackey).

§Institut für Mathematik, MA 4-5, Technische Universität Berlin, D-10623 Berlin, Germany (mehl@math.tu-berlin.de, mehrmann@math.tu-berlin.de; also http://www.math.tu-berlin.de/~mehl, http://www.math.tu-berlin.de/~mehrmann).

Many physical problems lead to matrix polynomials that are structured in some way; for example, the coefficient matrices may all be symmetric [9], or perhaps alternate between symmetric and skew-symmetric [15], or even have palindromic structure [12]. Such structure in the matrix polynomial often forces symmetries or constraints on the spectrum [12], [14], [15], [21] that have physical significance. Numerical methods (in a finite precision environment) that ignore this structure often destroy these qualitatively important spectral symmetries, sometimes even to the point of producing physically meaningless or uninterpretable results [21].

Unfortunately the companion form linearizations do not reflect any structure that may be present in the original polynomial, so their use for numerical computation in such situations may be problematic. It would be preferable if the structural properties of the polynomial were faithfully reflected in the linearization; a structure-preserving numerical method that leaves the qualitative properties of the spectrum intact would then be possible. Examples of such structured linearizations and their concomitant structure-preserving numerical methods can be found in [14] and [15].

An important issue for any computational problem is its conditioning, i.e., its sensitivity to small perturbations. It is known that different linearizations for a given polynomial eigenvalue problem can have very different conditioning [20], [21], so that numerical methods may produce rather different results for each linearization. It would clearly be useful to have available a large class of easily constructible linearizations from which one could always select a linearization guaranteed to be as well-conditioned as the original problem.

A further issue for linearizations concerns eigenvalues at ∞. Much of the literature on polynomial eigenvalue problems considers only polynomials whose leading coefficient matrix A_k is nonsingular (or even the identity), so the issue of infinite eigenvalues does not even arise. However, there are a number of applications, such as constraint multibody systems [2], [16], circuit simulation [3], or optical waveguide design [17], where the leading coefficient is singular. In such cases one must choose a linearization with care, since not all linearizations properly reflect the Jordan structure of the eigenvalue ∞ [13]. It has therefore been suggested [4], [10] that only *strong linearizations*, which are guaranteed to preserve the structure of infinite eigenvalues, can safely be used in such circumstances. Having a large class of linearizations that are known to also be strong linearizations would make this issue of infinite eigenvalues less of a concern in practice.

The aim of this paper is to show how to systematically generate, for any regular matrix polynomial P, large classes of linearizations that address these issues. These linearizations are easy to construct from the data in P, properly handle any infinite eigenvalues, provide a fertile source of structured linearizations for many types of structured polynomials [7], [12], and collectively constitute a well-defined arena in which to look for "optimally" conditioned linearizations [8].

After introducing some basic definitions and notation in section 2, we develop a natural generalization of the companion forms in section 3. The result is two large vector spaces of pencils for each matrix polynomial P, termed $\mathbb{L}_1(P)$ and $\mathbb{L}_2(P)$. Eigenvectors of any pencil from these associated vector spaces are shown to be simply related to the eigenvectors of P, thereby deepening the analogy to the companion forms. While not every pencil in these spaces is a linearization for P, we describe conditions under which these pencils are linearizations in section 4. As a consequence we can then show that almost every pencil in these spaces is in fact a strong linearization for P.

Finally, pencils in $\mathbb{L}_1(P) \cap \mathbb{L}_2(P)$ are considered in sections 5 and 6. For a polynomial P of degree k this intersection, termed $\mathbb{DL}(P)$, is shown to be a subspace of dimension k. Further properties of these special pencils are derived, including an elegant characterization of exactly which pencils in $\mathbb{DL}(P)$ are linearizations for P.

Subsequent papers [7], [8], [12] extend this work to the investigation of the conditioning of linearizations in $\mathbb{DL}(P)$ and the construction of structured linearizations for various types of structured matrix polynomials.

2. Basic definitions and notation. We study $n \times n$ matrix polynomials of the form

$$P(\lambda) = \sum_{i=0}^{k} \lambda^i A_i, \quad A_0, \ldots, A_k \in \mathbb{F}^{n\times n},\ A_k \neq 0, \tag{2.1}$$

where $\mathbb{F}$ denotes the field of real or complex numbers and k is the degree of P.

DEFINITION 2.1. *If $\lambda \in \mathbb{C}$ and nonzero $x \in \mathbb{C}^n$ satisfy $P(\lambda)x = 0$, then x is said to be a right eigenvector of P corresponding to the (finite) eigenvalue λ.*

Following standard usage, we will often abbreviate "right eigenvector" to just "eigenvector" when there is no ambiguity.

Our main concern is with *regular matrix polynomials*, i.e., polynomials $P(\lambda)$ such that $\det P(\lambda)$ is not identically zero for all $\lambda \in \mathbb{C}$; for such polynomials the finite eigenvalues are precisely the roots of the scalar polynomial $\det P(\lambda)$. Note, however, that some of our results also hold for singular matrix polynomials (these are studied in detail in [13], [18]).

It is also useful to allow ∞ as a possible eigenvalue of $P(\lambda)$. The technical device underlying this notion is the correspondence between the eigenvalues of P and those of the polynomial obtained from P by reversing the order of its coefficient matrices.

DEFINITION 2.2 (Reversal of matrix polynomials). *For a matrix polynomial $P(\lambda)$ of degree k as in* (2.1)*, the* reversal *of $P(\lambda)$ is the polynomial*

$$\operatorname{rev} P(\lambda) := \lambda^k P(1/\lambda) = \sum_{i=0}^{k} \lambda^i A_{k-i}. \tag{2.2}$$

Note that the nonzero finite eigenvalues of $\operatorname{rev} P$ are the reciprocals of those of P; the next definition shows how in this context we may also sensibly view 0 and ∞ as reciprocals.

DEFINITION 2.3 (Eigenvalue at ∞). *Let $P(\lambda)$ be a regular matrix polynomial of degree $k \geq 1$. Then $P(\lambda)$ is said to have an eigenvalue at ∞ with eigenvector x if $\operatorname{rev} P(\lambda)$ has the eigenvalue 0 with eigenvector x. The algebraic, geometric, and partial multiplicities of the infinite eigenvalue are defined to be the same as the corresponding multiplicities of the zero eigenvalue of $\operatorname{rev} P(\lambda)$.*

The classical approach to solving and investigating polynomial eigenvalue problems $P(\lambda)x = 0$ is to first perform a *linearization*, that is, to transform the given polynomial into a linear matrix pencil $L(\lambda) = \lambda X + Y$ with the same eigenvalues, and then work with this pencil. This transformation of polynomials to pencils is mediated by *unimodular* matrix polynomials, i.e., matrix polynomials $E(\lambda)$ such that $\det E(\lambda)$ is a nonzero constant, independent of λ.

DEFINITION 2.4 (Linearization [5]). *Let $P(\lambda)$ be an $n \times n$ matrix polynomial of degree k with $k \geq 1$. A pencil $L(\lambda) = \lambda X + Y$ with $X, Y \in \mathbb{F}^{kn\times kn}$ is called a* linearization *of $P(\lambda)$ if there exist unimodular matrix polynomials $E(\lambda), F(\lambda)$ such*

that

$$E(\lambda)L(\lambda)F(\lambda) = \left[\begin{array}{c|c} P(\lambda) & 0 \\ \hline 0 & I_{(k-1)n} \end{array}\right].$$

There are many different possibilities for linearizations, but probably the most important examples in practice have been the so-called companion forms or companion polynomials [5]. Letting

$$X_1 = X_2 = \operatorname{diag}(A_k, I_{(k-1)n}), \tag{2.3a}$$

$$Y_1 = \begin{bmatrix} A_{k-1} & A_{k-2} & \cdots & A_0 \\ -I_n & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & -I_n & 0 \end{bmatrix}, \quad \text{and} \quad Y_2 = \begin{bmatrix} A_{k-1} & -I_n & \cdots & 0 \\ A_{k-2} & 0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & -I_n \\ A_0 & 0 & \cdots & 0 \end{bmatrix}, \tag{2.3b}$$

then $C_1(\lambda) = \lambda X_1 + Y_1$ and $C_2(\lambda) = \lambda X_2 + Y_2$ are, respectively, called the *first* and *second companion forms* for $P(\lambda)$ in (2.1).

The notion of linearization in Definition 2.4 has been designed mainly for matrix polynomials (2.1) with invertible leading coefficient A_k. In this case all the eigenvalues of $P(\lambda)$ are finite, and their Jordan structures (i.e., their partial multiplicities) may be recovered from *any* linearization [5]. However, the situation is somewhat different when the leading coefficient of a regular $P(\lambda)$ is singular, so that ∞ is an eigenvalue with some multiplicity $m > 0$. Although the Jordan structures of all the finite eigenvalues of P are still faithfully recovered from any linearization of P, the eigenvalue ∞ is problematic. Consider, for example, the fact that the identity matrix is a linearization for any unimodular $P(\lambda)$. Indeed, in [10] it is shown that *any* Jordan structure for the eigenvalue ∞ that is compatible with its algebraic multiplicity m can be realized by some linearization for P. Thus linearization in the sense of Definition 2.4 completely fails to reflect the Jordan structure of infinite eigenvalues.

To overcome this deficiency, a modification of Definition 2.4 was introduced in [4], and termed *strong linearization* in [10]. The correspondence between the infinite eigenvalue of a matrix polynomial P and the eigenvalue zero of $\operatorname{rev} P$ is the source of this strengthened definition.

DEFINITION 2.5 (Strong Linearization). *Let $P(\lambda)$ be a matrix polynomial of degree k with $k \geq 1$. If $L(\lambda)$ is a linearization for $P(\lambda)$ and $\operatorname{rev} L(\lambda)$ is a linearization for $\operatorname{rev} P(\lambda)$, then $L(\lambda)$ is said to be a* strong linearization *for $P(\lambda)$.*

For regular polynomials $P(\lambda)$, the additional property that $\operatorname{rev} L(\lambda)$ is a linearization for $\operatorname{rev} P(\lambda)$ ensures that the Jordan structure of the eigenvalue ∞ is preserved by strong linearizations. The first and second companion forms of any regular polynomial P have this additional property [4], and thus are always strong linearizations for P. Most of the pencils we construct in this paper will be shown to be strong linearizations.

The following notation will be used throughout the paper: $I = I_n$ is the $n \times n$ identity, $R = R_k$ denotes the $k \times k$ reverse identity, and $N = N_k$ is the standard $k \times k$ nilpotent Jordan block, i.e.,

$$R = R_k = \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix}, \quad \text{and} \quad N = N_k = \begin{bmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix}. \tag{2.4}$$

The vector $\begin{bmatrix} \lambda^{k-1} & \lambda^{k-2} & \cdots & \lambda & 1 \end{bmatrix}^T$ of decreasing powers of λ is denoted by Λ. We will also sometimes use Λ with an argument, so that $\Lambda(r) = \begin{bmatrix} r^{k-1} & r^{k-2} & \cdots & r & 1 \end{bmatrix}^T$. Denoting the Kronecker product by $\otimes$, the unimodular matrix polynomials

$$T(\lambda) = \begin{bmatrix} 1 & \lambda & \lambda^2 & \cdots & \lambda^{k-1} \\ & 1 & \lambda & \ddots & \vdots \\ & & 1 & \ddots & \lambda^2 \\ & & & \ddots & \lambda \\ & & & & 1 \end{bmatrix} \otimes I \quad \text{and} \quad G(\lambda) = \begin{bmatrix} 1 & & & \lambda^{k-1} \\ & \ddots & & \vdots \\ & & 1 & \lambda \\ & & & 1 \end{bmatrix} \otimes I \tag{2.5}$$

are used in several places in this paper. Observe that the last block-column of $G(\lambda)$ is $\Lambda \otimes I$, and that $T(\lambda)$ may be factored as

$$T(\lambda) = G(\lambda) \begin{bmatrix} I & \lambda I & & & \\ & I & & & \\ & & I & & \\ & & & \ddots & \\ & & & & I \end{bmatrix} \begin{bmatrix} I & & & & \\ & I & \lambda I & & \\ & & I & & \\ & & & \ddots & \\ & & & & I \end{bmatrix} \cdots \begin{bmatrix} I & & & & \\ & \ddots & & & \\ & & I & \lambda I & \\ & & & I & \\ & & & & I \end{bmatrix}. \tag{2.6}$$

3. Vector spaces of "potential" linearizations. The companion forms of a matrix polynomial $P(\lambda)$ have several nice properties that make them attractive as linearizations for P:

- they are immediately constructible from the data in P,
- eigenvectors of P are easily recovered from eigenvectors of the companion forms,
- they are always strong linearizations for P.

However, the companion forms have one significant drawback; they usually do not reflect any structure or eigenvalue symmetry that may be present in the original polynomial P. One would like to be able to draw on a source of linearizations for P that allow for the preservation of structure while sharing as many of the useful properties of companion forms as possible. To this end we introduce vector spaces of pencils that generalize the two companion forms, and analyze some of the properties these pencils have in common with the companion forms.

To motivate the definition of these spaces, let us recall the origin of the first companion form. Imitating the standard procedure for converting a system of higher order linear differential algebraic equations into a first order system (see [5]), introduce the variables $x_1 = \lambda^{k-1}x$, $x_2 = \lambda^{k-2}x, \ldots, x_{k-1} = \lambda x$, $x_k = x$, thereby transforming the $n \times n$ polynomial eigenvalue problem $P(\lambda)x = (\sum_{i=0}^{k} \lambda^i A_i)x = 0$ into

$$A_k(\lambda x_1) + A_{k-1}x_1 + A_{k-2}x_2 + \cdots + A_1 x_{k-1} + A_0 x_k = 0.$$

Then, together with the relations between successive variables, this can all be expressed as the $kn \times kn$ linear eigenvalue problem

$$\underbrace{\left(\lambda \begin{bmatrix} A_k & 0 & \cdots & 0 \\ 0 & I_n & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & I_n \end{bmatrix} + \begin{bmatrix} A_{k-1} & A_{k-2} & \cdots & A_0 \\ -I_n & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & -I_n & 0 \end{bmatrix} \right)}_{= C_1(\lambda)} \begin{bmatrix} x_1 \\ \vdots \\ x_{k-1} \\ x_k \end{bmatrix} = 0. \tag{3.1}$$

Conversely, if we start with (3.1), then the last $k-1$ block rows immediately constrain any solution of (3.1) to have the form

$$\begin{bmatrix} x_1 \\ \vdots \\ x_{k-1} \\ x_k \end{bmatrix} = \begin{bmatrix} \lambda^{k-1}x \\ \vdots \\ \lambda x \\ x \end{bmatrix} = \Lambda \otimes x$$

for some vector $x \in \mathbb{F}^n$. Thus to solve (3.1) it is reasonable to restrict attention to products of the form $C_1(\lambda) \cdot (\Lambda \otimes x)$. However,

$$C_1(\lambda) \cdot \big(\Lambda \otimes x\big) = \Big[\big(P(\lambda)x\big)^T \quad 0 \quad \cdots \quad 0\Big]^T \quad \text{for all} \quad x \in \mathbb{F}^n, \tag{3.2}$$

and so any solution of (3.1) leads to a solution of the original problem $P(\lambda)x = 0$. Now observe that (3.2) is equivalent to the identity

$$C_1(\lambda) \cdot \big(\Lambda \otimes I_n\big) = C_1(\lambda) \begin{bmatrix} \lambda^{k-1}I_n \\ \vdots \\ \lambda I_n \\ I_n \end{bmatrix} = \begin{bmatrix} P(\lambda) \\ 0 \\ \vdots \\ 0 \end{bmatrix} = e_1 \otimes P(\lambda). \tag{3.3}$$

Thus to generalize the companion form we consider the set of all $kn \times kn$ matrix pencils $L(\lambda) = \lambda X + Y$ satisfying the property

$$L(\lambda) \cdot (\Lambda \otimes I_n) = L(\lambda) \begin{bmatrix} \lambda^{k-1}I_n \\ \vdots \\ \lambda I_n \\ I_n \end{bmatrix} = \begin{bmatrix} v_1 P(\lambda) \\ \vdots \\ v_{k-1}P(\lambda) \\ v_k P(\lambda) \end{bmatrix} = v \otimes P(\lambda) \tag{3.4}$$

for some vector $v = [v_1, \ldots, v_k]^T \in \mathbb{F}^k$. This set of pencils will be denoted by $\mathbb{L}_1(P)$ as a reminder that it generalizes the first companion form of P. To work with property (3.4) more effectively we also introduce the notation

$$\mathcal{V}_P = \{v \otimes P(\lambda) \,:\, v \in \mathbb{F}^k\} \tag{3.5}$$

for the set of all possible right-hand sides of (3.4). Thus we have the following definition.

DEFINITION 3.1. $\mathbb{L}_1(P) := \{L(\lambda) = \lambda X + Y : X, Y \in \mathbb{F}^{kn\times kn},\ L(\lambda) \cdot (\Lambda \otimes I_n) \in \mathcal{V}_P\}$.

We will sometimes use the phrase "$L(\lambda)$ satisfies the right ansatz with vector v" or "v is the right ansatz vector for $L(\lambda)$" when $L(\lambda) \in \mathbb{L}_1(P)$ and the vector v in (3.4) is the focus of attention. We say "right" ansatz here because $L(\lambda)$ is multiplied on the right by the block column $\Lambda \otimes I_n$; later we introduce an analogous "left ansatz."

From the properties of the Kronecker product it is easy to see that $\mathcal{V}_P$ is a vector space isomorphic to $\mathbb{F}^k$, and consequently that $\mathbb{L}_1(P)$ is also a vector space.

PROPOSITION 3.2. *For any polynomial $P(\lambda)$, $\mathbb{L}_1(P)$ is a vector space over $\mathbb{F}$.*

Since $C_1(\lambda)$ is always in $\mathbb{L}_1(P)$, we see that $\mathbb{L}_1(P)$ is a nontrivial vector space for any matrix polynomial P.

Our next goal is to show that, like the companion forms, pencils in $\mathbb{L}_1(P)$ are easily constructible from the data in P. A consequence of this construction is a characterization of all the pencils in $\mathbb{L}_1(P)$ and a calculation of $\dim \mathbb{L}_1(P)$. To simplify the

discussion, we introduce the following new operation on block matrices as a convenient tool for working with products of the form $L(\lambda)\cdot(\Lambda\otimes I_n)$.

DEFINITION 3.3 (Column shifted sum). *Let X and Y be block matrices*

$$X=\begin{bmatrix} X_{11} & \cdots & X_{1k}\\ \vdots & & \vdots\\ X_{k1} & \cdots & X_{kk}\end{bmatrix},\qquad Y=\begin{bmatrix} Y_{11} & \cdots & Y_{1k}\\ \vdots & & \vdots\\ Y_{k1} & \cdots & Y_{kk}\end{bmatrix}$$

with blocks $X_{ij},Y_{ij}\in\mathbb{F}^{n\times n}$. Then the column shifted sum *of X and Y is defined to be*

$$X\boxplus\!\!\rightarrow Y:=\begin{bmatrix} X_{11} & \cdots & X_{1k} & 0\\ \vdots & & \vdots & \vdots\\ X_{k1} & \cdots & X_{kk} & 0\end{bmatrix}+\begin{bmatrix} 0 & Y_{11} & \cdots & Y_{1k}\\ \vdots & \vdots & & \vdots\\ 0 & Y_{k1} & \cdots & Y_{kk}\end{bmatrix},$$

where the zero blocks are also $n\times n$.

As an example, for the first companion form $C_1(\lambda)=\lambda X_1+Y_1$ of $P(\lambda)=\sum_{i=0}^{k}\lambda^i A_i$, the column shifted sum $X_1\boxplus\!\!\rightarrow Y_1$ is just

$$\begin{bmatrix} A_k & 0 & \cdots & 0\\ 0 & I_n & \ddots & \vdots\\ \vdots & \ddots & \ddots & 0\\ 0 & \cdots & 0 & I_n\end{bmatrix}\boxplus\!\!\rightarrow\begin{bmatrix} A_{k-1} & A_{k-2} & \cdots & A_0\\ -I_n & 0 & \cdots & 0\\ \vdots & \ddots & \ddots & \vdots\\ 0 & \cdots & -I_n & 0\end{bmatrix}=\begin{bmatrix} A_k & A_{k-1} & \cdots & A_0\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 0\end{bmatrix}.$$

Thus, the property $C_1(\lambda)\cdot(\Lambda\otimes I_n)=e_1\otimes P(\lambda)$ from (3.3) translates in terms of the column shifted sum into $X_1\boxplus\!\!\rightarrow Y_1=e_1\otimes[A_k\ A_{k-1}\ \cdots\ A_0]$. In fact, this shifted sum operation is specifically designed to imitate the product of a pencil $L(\lambda)=\lambda X+Y$ with the block column matrix $\Lambda\otimes I_n$, in the sense of the following lemma.

LEMMA 3.4. *Let $P(\lambda)=\sum_{i=0}^{k}\lambda^i A_i$ be an $n\times n$ matrix polynomial, and $L(\lambda)=\lambda X+Y$ a $kn\times kn$ pencil. Then for $v\in\mathbb{F}^k$,*

$$(\lambda X+Y)\cdot(\Lambda\otimes I_n)=v\otimes P(\lambda)\quad\Longleftrightarrow\quad X\boxplus\!\!\rightarrow Y=v\otimes[A_k\ A_{k-1}\ \cdots\ A_0],\tag{3.6}$$

and so the space $\mathbb{L}_1(P)$ may be alternatively characterized as

$$\mathbb{L}_1(P)=\{\lambda X+Y\,:\,X\boxplus\!\!\rightarrow Y=v\otimes[A_k\ A_{k-1}\ \cdots\ A_0],\ v\in\mathbb{F}^k\}.\tag{3.7}$$

The proof follows from a straightforward calculation which is omitted. The column shifted sum now allows us to directly construct all the pencils in $\mathbb{L}_1(P)$.

THEOREM 3.5 (Characterization of pencils in $\mathbb{L}_1(P)$). *Let $P(\lambda)=\sum_{i=0}^{k}\lambda^i A_i$ be an $n\times n$ matrix polynomial, and $v\in\mathbb{F}^k$ any vector. Then the set of pencils in $\mathbb{L}_1(P)$ with right ansatz vector v consists of all $L(\lambda)=\lambda X+Y$ such that*

$$X=\begin{bmatrix} \overset{n}{v\otimes A_k} & \overset{(k-1)n}{-W}\end{bmatrix}\quad\text{and}\quad Y=\begin{bmatrix} \overset{(k-1)n}{W+\left(v\otimes\begin{bmatrix} A_{k-1} & \cdots & A_1\end{bmatrix}\right)} & \overset{n}{v\otimes A_0}\end{bmatrix},$$

with $W\in\mathbb{F}^{kn\times(k-1)n}$ chosen arbitrarily.

Proof. Consider the *multiplication map* $\mathcal{M}$ that is implicit in the definition of $\mathbb{L}_1(P)$:

$$\begin{aligned}\mathbb{L}_1(P)&\xrightarrow{\ \mathcal{M}\ }\mathcal{V}_P,\\ L(\lambda)&\longmapsto L(\lambda)\,(\Lambda\otimes I_n).\end{aligned}$$

Clearly $\mathcal{M}$ is linear. To see that $\mathcal{M}$ is surjective, let $v \otimes P(\lambda)$ be an arbitrary element of $\mathcal{V}_P$ and construct

$$X_v = \begin{bmatrix} \overset{n}{v \otimes A_k} & \overset{(k-1)n}{0} \end{bmatrix} \quad \text{and} \quad Y_v = \Big[v \otimes \begin{bmatrix} \overset{(k-1)n}{A_{k-1} & \cdots & A_1} \end{bmatrix} \quad \overset{n}{v \otimes A_0} \Big].$$

Then $X_v \boxplus Y_v = v \otimes [A_k \ A_{k-1} \ \cdots \ A_0]$, so by Lemma 3.4, $L_v(\lambda) := \lambda X_v + Y_v$ is an $\mathcal{M}$-preimage of $v \otimes P(\lambda)$. The set of all $\mathcal{M}$-preimages of $v \otimes P(\lambda)$ is then $L_v(\lambda) + \ker \mathcal{M}$, so all that remains is to compute $\ker \mathcal{M}$. By (3.6), the kernel of $\mathcal{M}$ consists of all pencils $\lambda X + Y$ satisfying $X \boxplus Y = 0$. The definition of the shifted sum then implies that X and Y must have the form

$$X = \begin{bmatrix} \overset{n}{0} & \overset{(k-1)n}{-W} \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} \overset{(k-1)n}{W} & \overset{n}{0} \end{bmatrix},$$

where $W \in \mathbb{F}^{kn \times (k-1)n}$ is arbitrary. This completes the proof. ∎

COROLLARY 3.6. $\dim \mathbb{L}_1(P) = k(k-1)n^2 + k$.

Proof. Since $\mathcal{M}$ is surjective, $\dim \mathbb{L}_1(P) = \dim \ker \mathcal{M} + \dim \mathcal{V}_P = k(k-1)n^2 + k$. ∎

Thus we see that $\mathbb{L}_1(P)$ is a relatively large subspace of the full pencil space (with dimension $2k^2n^2$), yet the pencils in $\mathbb{L}_1(P)$ are still easy to construct from the data in P. The next corollary isolates a special case of Theorem 3.5 that plays an important role in section 4.

COROLLARY 3.7. *Suppose $L(\lambda) = \lambda X + Y \in \mathbb{L}_1(P)$ has right ansatz vector $v = \alpha e_1$. Then*

$$X = \left[\begin{array}{c|c} \alpha A_k & X_{12} \\ \hline 0 & -Z \end{array}\right] \quad \textit{and} \quad Y = \left[\begin{array}{c|c} Y_{11} & \alpha A_0 \\ \hline Z & 0 \end{array}\right] \tag{3.8}$$

for some $Z \in \mathbb{F}^{(k-1)n \times (k-1)n}$.

Note that $C_1(\lambda)$ fits the pattern in Corollary 3.7 with $v = e_1$ and $Z = -I_{(k-1)n}$.

The second important property of the companion form is the simple relationship between its eigenvectors and those of the polynomial P that it linearizes. From the discussion following (3.1) it is evident that every eigenvector of $C_1(\lambda)$ has the form $\Lambda \otimes x$, where x is an eigenvector of P. Thus eigenvectors of P are recovered simply by extracting the last n coordinates from eigenvectors of the companion form. Our next result shows that linearizations in $\mathbb{L}_1(P)$ also have this property.

THEOREM 3.8 (Eigenvector Recovery Property for $\mathbb{L}_1(P)$). *Let $P(\lambda)$ be an $n \times n$ matrix polynomial of degree k, and $L(\lambda)$ any pencil in $\mathbb{L}_1(P)$ with nonzero right ansatz vector v. Then $x \in \mathbb{C}^n$ is an eigenvector for $P(\lambda)$ with finite eigenvalue $\lambda \in \mathbb{C}$ if and only if $\Lambda \otimes x$ is an eigenvector for $L(\lambda)$ with eigenvalue λ. If, in addition, P is regular and $L \in \mathbb{L}_1(P)$ is a linearization for P, then every eigenvector of L with finite eigenvalue λ is of the form $\Lambda \otimes x$ for some eigenvector x of P.*

Proof. The first statement follows immediately from the identity

$$L(\lambda)(\Lambda \otimes x) = L(\lambda)(\Lambda \otimes I_n)(1 \otimes x) = (v \otimes P(\lambda))(1 \otimes x) = v \otimes (P(\lambda)x).$$

For the second statement, assume that $\lambda \in \mathbb{C}$ is a finite eigenvalue of $L(\lambda)$ with geometric multiplicity m, and let $y \in \mathbb{C}^{kn}$ be any eigenvector of $L(\lambda)$ associated with λ. Since $L(\lambda)$ is a linearization of $P(\lambda)$, the geometric multiplicity of λ for $P(\lambda)$ is

also m. Let $x_1, \ldots, x_m$ be linearly independent eigenvectors of $P(\lambda)$ associated with λ, and define $y_i = \Lambda \otimes x_i$ for $i = 1, \ldots, m$. Then $y_1, \ldots, y_m$ are linearly independent eigenvectors for $L(\lambda)$ with eigenvalue λ, and so y must be a linear combination of $y_1, \ldots, y_m$. Thus y has the form $y = \Lambda \otimes x$ for some eigenvector $x \in \mathbb{C}^n$ for P. ∎

A result analogous to Theorem 3.8 is also valid for the eigenvalue ∞. Because additional arguments are needed, this will be deferred until section 4.

The above development and analysis of the pencil space $\mathbb{L}_1(P)$ has a parallel version in which the starting point is the second companion form $C_2(\lambda) = \lambda X_2 + Y_2$, as in (2.3). The analogue of (3.3) is the identity

$$\begin{bmatrix} \lambda^{k-1} I_n & \cdots & \lambda I_n & I_n \end{bmatrix} \cdot C_2(\lambda) \;=\; \begin{bmatrix} P(\lambda) & 0 & \cdots & 0 \end{bmatrix},$$

expressed more compactly as $(\Lambda^T \otimes I_n) \cdot C_2(\lambda) = e_1^T \otimes P(\lambda)$. This leads us to consider pencils $L(\lambda) = \lambda X + Y$ satisfying the "left ansatz"

$$\big(\Lambda^T \otimes I_n\big) \cdot L(\lambda) = w^T \otimes P(\lambda), \tag{3.9}$$

and to a corresponding vector space $\mathbb{L}_2(P)$. The vector w in (3.9) will be referred to as the "left ansatz vector" for $L(\lambda)$.

Definition 3.9. *With $\mathcal{W}_P = \{w^T \otimes P(\lambda) : w \in \mathbb{F}^k\}$, we define*

$$\mathbb{L}_2(P) = \{L(\lambda) = \lambda X + Y :\, X, Y \in \mathbb{F}^{kn \times kn},\, \big(\Lambda^T \otimes I_n\big) \cdot L(\lambda) \in \mathcal{W}_P\}.$$

The analysis of $\mathbb{L}_2(P)$ is aided by the introduction of the following block matrix operation.

Definition 3.10 (Row shifted sum). *Let X and Y be block matrices*

$$X = \begin{bmatrix} X_{11} & \cdots & X_{1k} \\ \vdots & & \vdots \\ X_{k1} & \cdots & X_{kk} \end{bmatrix}, \qquad Y = \begin{bmatrix} Y_{11} & \cdots & Y_{1k} \\ \vdots & & \vdots \\ Y_{k1} & \cdots & Y_{kk} \end{bmatrix}$$

with blocks $X_{ij}, Y_{ij} \in \mathbb{F}^{n \times n}$. Then the row shifted sum *of X and Y is defined to be*

$$X \underset{\downarrow}{\boxplus} Y := \begin{bmatrix} X_{11} & \cdots & X_{1k} \\ \vdots & & \vdots \\ X_{k1} & \cdots & X_{kk} \\ 0 & \cdots & 0 \end{bmatrix} + \begin{bmatrix} 0 & \cdots & 0 \\ Y_{11} & \cdots & Y_{1k} \\ \vdots & & \vdots \\ Y_{k1} & \cdots & Y_{kk} \end{bmatrix},$$

where the zero blocks are also $n \times n$.

The following analogue of Lemma 3.4 establishes the correspondence between the left ansatz and row shifted sums.

Lemma 3.11. *Let $P(\lambda) = \sum_{i=0}^{k} \lambda^i A_i$ be an $n \times n$ matrix polynomial, and $L(\lambda) = \lambda X + Y$ a $kn \times kn$ pencil. Then for any $w \in \mathbb{F}^k$,*

$$(\Lambda^T \otimes I_n) \cdot (\lambda X + Y) = w^T \otimes P(\lambda) \quad \Longleftrightarrow \quad X \underset{\downarrow}{\boxplus} Y = w^T \otimes \begin{bmatrix} A_k \\ \vdots \\ A_0 \end{bmatrix}. \tag{3.10}$$

Using these tools, one can characterize the pencils in $\mathbb{L}_2(P)$ in a manner completely analogous to Theorem 3.5, and thus conclude that

$$\dim \mathbb{L}_2(P) = \dim \mathbb{L}_1(P) = k(k-1)n^2 + k. \tag{3.11}$$

It is also not difficult to establish a stronger relationship between the spaces $\mathbb{L}_1(P)$ and $\mathbb{L}_2(P)$, which again immediately implies (3.11). Here for a polynomial $P(\lambda) = \sum_{i=0}^{k} \lambda^i A_i$, we use P^T to denote the polynomial $\sum_{i=0}^{k} \lambda^i A_i^T$; by extension, if $\mathcal{S}$ is any set of polynomials, then $\mathcal{S}^T$ is the set $\{P^T : P \in \mathcal{S}\}$.

PROPOSITION 3.12. $\mathbb{L}_2(P) = [\mathbb{L}_1(P^T)]^T$.

Proof.

$$\begin{aligned} L \in \mathbb{L}_1(P^T) &\Leftrightarrow L(\lambda) \cdot (\Lambda \otimes I_n) = v \otimes P^T(\lambda) \\ &\Leftrightarrow \left(\Lambda^T \otimes I_n\right) \cdot L^T(\lambda) = v^T \otimes P(\lambda) \;\Leftrightarrow\; L^T \in \mathbb{L}_2(P). \end{aligned}$$ ∎

The analogue of Theorem 3.8 for pencils in $\mathbb{L}_2(P)$ involves left eigenvectors of $P(\lambda)$ rather than right eigenvectors. Since the definition of a left eigenvector of a matrix polynomial does not seem to be completely standardized, we include here the definition used in this paper.

DEFINITION 3.13 (Left eigenvectors). *A left eigenvector of an $n \times n$ matrix polynomial P associated with a finite eigenvalue λ is a nonzero vector $y \in \mathbb{C}^n$ such that $y^* P(\lambda) = 0$. A left eigenvector for P corresponding to the eigenvalue ∞ is a left eigenvector for* $\operatorname{rev} P$ *associated with the eigenvalue* 0.

This definition differs from the one adopted in [5], although it is compatible with the usual definition for left eigenvectors of a matrix [6], [19]. We have chosen Definition 3.13 here because it is the one typically used in formulas for condition numbers of eigenvalues, a topic investigated in [8]. The following result shows that left eigenvectors of P are easily recovered from linearizations in $\mathbb{L}_2(P)$. The proof is completely analogous to that given for Theorem 3.8.

THEOREM 3.14 (Eigenvector Recovery Property for $\mathbb{L}_2(P)$). *Let $P(\lambda)$ be an $n \times n$ matrix polynomial of degree k, and $L(\lambda)$ any pencil in $\mathbb{L}_2(P)$ with nonzero left ansatz vector w. Then $y \in \mathbb{C}^n$ is a left eigenvector for $P(\lambda)$ with finite eigenvalue $\lambda \in \mathbb{C}$ if and only if $\overline{\Lambda} \otimes y$ is a left eigenvector for $L(\lambda)$ with eigenvalue λ. If, in addition, P is regular and $L \in \mathbb{L}_2(P)$ is a linearization for P, then every left eigenvector of L with finite eigenvalue λ is of the form $\overline{\Lambda} \otimes y$ for some left eigenvector y of P.*

Just as for Theorem 3.8, there is an analogous result for the eigenvalue ∞ that can be found in section 4.

In this section we have seen that pencils in $\mathbb{L}_1(P)$ and $\mathbb{L}_2(P)$ closely resemble the companion forms, and have eigenvectors that are simply related to those of P. Thus one can reasonably view $\mathbb{L}_1(P)$ and $\mathbb{L}_2(P)$ as large classes of "potential" linearizations for $P(\lambda)$. So far, though, we have not shown any of these "good candidates" to actually be linearizations. It is to this question that we turn next.

4. When is a pencil in $\mathbb{L}_1(P)$ a linearization? It is clear that not all pencils in the spaces $\mathbb{L}_1(P)$ and $\mathbb{L}_2(P)$ are linearizations of P—consider, for example, any pencil in $\mathbb{L}_1(P)$ with right ansatz vector $v = 0$. In this section we focus on $\mathbb{L}_1(P)$ and obtain criteria for deciding whether a pencil from $\mathbb{L}_1(P)$ is a linearization for P or not. We show, for example, that for any given $L \in \mathbb{L}_1(P)$ there is typically a condition (specific to L) on the coefficient matrices of P that must be satisfied in order to guarantee that L is actually a linearization for P. Specific examples of such "linearization conditions" can be found in section 4.1 and in the tables in section 5. Analogues of all the results in this section also hold for $\mathbb{L}_2(P)$, with very similar arguments.

We begin with a result concerning the special case of the right ansatz (3.4) considered in Corollary 3.7. Note that P is not assumed here to be regular.

THEOREM 4.1. *Suppose that* $P(\lambda) = \sum_{i=0}^{k} \lambda^i A_i$ *with* $A_k \neq 0$ *is an* $n \times n$ *matrix polynomial, and* $L(\lambda) = \lambda X + Y \in \mathbb{L}_1(P)$ *has nonzero right ansatz vector* $v = \alpha e_1$*, so that*

$$L(\lambda) \cdot (\Lambda \otimes I_n) = \alpha e_1 \otimes P(\lambda). \tag{4.1}$$

Partition X *and* Y *as in* (3.8) *so that*

$$L(\lambda) \;=\; \lambda X + Y \;=\; \lambda \left[\begin{array}{c|c} \alpha A_k & X_{12} \\ \hline 0 & -Z \end{array}\right] + \left[\begin{array}{c|c} Y_{11} & \alpha A_0 \\ \hline Z & 0 \end{array}\right], \tag{4.2}$$

where $Z \in \mathbb{F}^{(k-1)n \times (k-1)n}$*. Then* Z *nonsingular implies that* $L(\lambda)$ *is a strong linearization of* $P(\lambda)$*.*

Proof. We show first that $L(\lambda)$ is a linearization of $P(\lambda)$. Begin the reduction of $L(\lambda)$ to $\operatorname{diag}(P(\lambda), I_{(k-1)n})$ using the unimodular matrix polynomials $T(\lambda)$ and $G(\lambda)$ defined in (2.5). In the product $L(\lambda)G(\lambda)$, clearly the first $k-1$ block-columns are the same as those of $L(\lambda)$; because the last block-column of $G(\lambda)$ is $\Lambda \otimes I$, we see from (4.1) that the last block-column of $L(\lambda)G(\lambda)$ is $\alpha e_1 \otimes P(\lambda)$. Partitioning Z in (4.2) into block columns $[Z_1\ Z_2\ \cdots\ Z_{k-1}]$, where $Z_i \in \mathbb{F}^{(k-1)n \times n}$, we thus obtain

$$\begin{aligned} L(\lambda)G(\lambda) \;&=\; \begin{bmatrix} * & * & \dots & * & * \\ Z_1 & (Z_2 - \lambda Z_1) & \dots & (Z_{k-1} - \lambda Z_{k-2}) & -\lambda Z_{k-1} \end{bmatrix} G(\lambda) \\ &=\; \begin{bmatrix} * & * & \dots & * & \alpha P(\lambda) \\ Z_1 & (Z_2 - \lambda Z_1) & \dots & (Z_{k-1} - \lambda Z_{k-2}) & 0 \end{bmatrix}. \end{aligned}$$

Further transformation by block-column operations yields

$$L(\lambda)T(\lambda) = L(\lambda)\, \underbrace{G(\lambda) \begin{bmatrix} I & \lambda I & & & \\ & I & & & \\ & & I & & \\ & & & \ddots & \\ & & & & I \end{bmatrix} \begin{bmatrix} I & & & & \\ & I & \lambda I & & \\ & & I & & \\ & & & \ddots & \\ & & & & I \end{bmatrix} \cdots \begin{bmatrix} I & & & & \\ & \ddots & & & \\ & & I & \lambda I & \\ & & & I & \\ & & & & I \end{bmatrix}}_{= T(\lambda)} = \left[\begin{array}{c|c} * & \alpha P(\lambda) \\ \hline Z & 0 \end{array}\right].$$

Scaling and block-column permutations on $L(\lambda)T(\lambda)$ show that there exists a unimodular matrix polynomial $F(\lambda)$ such that

$$L(\lambda)F(\lambda) = \left[\begin{array}{c|c} P(\lambda) & W(\lambda) \\ \hline 0 & Z \end{array}\right]$$

for some matrix polynomial $W(\lambda)$. (Note that we have reached this point without any assumptions about Z.) Now if Z is nonsingular, then $L(\lambda)$ is a linearization for $P(\lambda)$, since

$$\left[\begin{array}{c|c} I & -W(\lambda)Z^{-1} \\ \hline 0 & Z^{-1} \end{array}\right] L(\lambda)F(\lambda) = \left[\begin{array}{c|c} P(\lambda) & 0 \\ \hline 0 & I_{(k-1)n} \end{array}\right].$$

To show that $L(\lambda)$ is also a strong linearization for $P(\lambda)$, it remains to show that $\operatorname{rev} L(\lambda) = \lambda Y + X$ is a linearization for $\operatorname{rev} P(\lambda)$. Now it would be nice if $\operatorname{rev} L(\lambda)$ were a pencil in $\mathbb{L}_1(\operatorname{rev} P)$, but it is not; however, a small modification of $\operatorname{rev} L(\lambda)$

is in $\mathbb{L}_1(\mathrm{rev}P)$. Observe that $\lambda^{k-1}\cdot\Lambda(1/\lambda) = [1,\lambda,\ldots,\lambda^{k-2},\lambda^{k-1}]^T = R_k\Lambda$, where R_k denotes the $k\times k$ reverse identity matrix. Thus replacing λ by $1/\lambda$ in (4.1) and multiplying both sides by λ^k yields

$$\lambda L(1/\lambda)\cdot\big(\lambda^{k-1}\Lambda(1/\lambda)\otimes I\big) = \alpha e_1\otimes\lambda^k P(1/\lambda),$$

or equivalently, $\mathrm{rev}L(\lambda)\cdot\big((R_k\Lambda)\otimes I\big) = \alpha e_1\otimes\mathrm{rev}P(\lambda)$. Thus, $\widehat{L}(\lambda) := \mathrm{rev}L(\lambda)\cdot(R_k\otimes I)$ satisfies

$$\widehat{L}(\lambda)\cdot(\Lambda\otimes I) = \alpha e_1\otimes\mathrm{rev}P(\lambda), \tag{4.3}$$

and so $\widehat{L}\in\mathbb{L}_1(\mathrm{rev}P)$. (Observe that $\widehat{L}(\lambda)$ is just $\mathrm{rev}L(\lambda) = \lambda Y + X$ with the block-columns of Y and X arranged in reverse order.) Since $\widehat{L}$ and $\mathrm{rev}L$ are equivalent pencils, the proof will be complete once we show that $\lambda\widehat{X}+\widehat{Y} := \widehat{L}(\lambda)$ is a linearization for $\mathrm{rev}P(\lambda)$. However, $\widehat{X} = Y\cdot(R_k\otimes I)$ and $\widehat{Y} = X\cdot(R_k\otimes I)$, and hence from (4.2) it follows that

$$\widehat{X} = \left[\begin{array}{c|c} \alpha A_0 & \widehat{X}_{12} \\ \hline 0 & -\widehat{Z} \end{array}\right] \quad\text{and}\quad \widehat{Y} = \left[\begin{array}{c|c} \widehat{Y}_{11} & \alpha A_k \\ \hline \widehat{Z} & 0 \end{array}\right],$$

where $\widehat{Z} = -Z\cdot(R_{k-1}\otimes I)$. Clearly $\widehat{Z}$ is nonsingular if Z is, and so by the part of the theorem that has already been proved, $\widehat{L}$ (and therefore also $\mathrm{rev}L$) is a linearization for $\mathrm{rev}P(\lambda)$. ∎

Remark 4.2. The fact (first proved in [4]) that the first companion form of any polynomial is always a strong linearization is a special case of Theorem 4.1.

When a matrix polynomial $P(\lambda)$ is regular, then it is easy to see from Definition 2.4 that any linearization for $P(\lambda)$ must also be regular. The next result shows something rather surprising: when a pencil L is in $\mathbb{L}_1(P)$ this minimal necessary condition of regularity is actually sufficient to guarantee that L is a linearization for P. This result serves to emphasize just how close a pencil is to being a linearization for P, even a strong linearization for P, once it satisfies the ansatz (3.4).

THEOREM 4.3 (Strong Linearization Theorem). *Let $P(\lambda)$ be a regular matrix polynomial, and let $L(\lambda)\in\mathbb{L}_1(P)$. Then the following statements are equivalent:*

(i) *$L(\lambda)$ is a linearization for $P(\lambda)$.*
(ii) *$L(\lambda)$ is a regular pencil.*
(iii) *$L(\lambda)$ is a strong linearization for $P(\lambda)$.*

Proof. "(i) ⇒ (ii)": If $L(\lambda)$ is a linearization for $P(\lambda)$, then there exist unimodular matrix polynomials $E(\lambda)$, $F(\lambda)$ such that

$$E(\lambda)L(\lambda)F(\lambda) = \begin{bmatrix} P(\lambda) & 0 \\ 0 & I_{(k-1)n} \end{bmatrix}.$$

Thus the regularity of $P(\lambda)$ implies the regularity of $L(\lambda)$.

"(ii) ⇒ (iii)": Since $L(\lambda)\in\mathbb{L}_1(P)$, we know that $L(\lambda)\cdot(\Lambda\otimes I_n) = v\otimes P(\lambda)$ for some $v\in\mathbb{F}^k$. However, $L(\lambda)$ is regular, and so v is nonzero. Let $M\in\mathbb{F}^{k\times k}$ be any nonsingular matrix such that $Mv = \alpha e_1$. Then the regular pencil $\widetilde{L}(\lambda) := (M\otimes I_n)\cdot L(\lambda)$ is in $\mathbb{L}_1(P)$ with right ansatz vector αe_1, since

$$\begin{aligned}\widetilde{L}(\lambda)(\Lambda\otimes I_n) = (M\otimes I_n)L(\lambda)(\Lambda\otimes I_n) &= (M\otimes I_n)(v\otimes P(\lambda)) \\ &= Mv\otimes P(\lambda) \\ &= \alpha e_1\otimes P(\lambda).\end{aligned}$$

Hence by Corollary 3.7 the matrices $\widetilde{X}$ and $\widetilde{Y}$ in $\widetilde{L}(\lambda) := \lambda\widetilde{X} + \widetilde{Y}$ have the forms

$$\widetilde{X} = \begin{array}{c} \begin{matrix} \scriptstyle n & \scriptstyle (k-1)n \end{matrix} \\ \left[\begin{array}{c|c} \alpha A_k & \widetilde{X}_{12} \\ \hline 0 & -\widetilde{Z} \end{array}\right] \end{array} \begin{matrix} \scriptstyle n \\ \scriptstyle (k-1)n \end{matrix} \quad \text{and} \quad \widetilde{Y} = \begin{array}{c} \begin{matrix} \scriptstyle (k-1)n & \scriptstyle n \end{matrix} \\ \left[\begin{array}{c|c} \widetilde{Y}_{11} & \alpha A_0 \\ \hline \widetilde{Z} & 0 \end{array}\right] \end{array} \begin{matrix} \scriptstyle n \\ \scriptstyle (k-1)n \end{matrix}.$$

Now if $\widetilde{Z}$ was singular, there would exist a nonzero vector $w \in \mathbb{F}^{(k-1)n}$ such that $w^T\widetilde{Z} = 0$. But this would imply that

$$\begin{bmatrix} 0 & w^T \end{bmatrix}(\lambda\widetilde{X} + \widetilde{Y}) \;=\; 0 \quad \text{for all} \;\; \lambda \in \mathbb{F},$$

contradicting the regularity of $\widetilde{L}(\lambda)$. Thus $\widetilde{Z}$ is nonsingular, and so by Theorem 4.1 we know that $\widetilde{L}(\lambda)$, and hence also $L(\lambda)$, is a strong linearization for $P(\lambda)$.

"(iii) $\Rightarrow$ (i)" is trivial. ∎

Now recall from Definitions 2.3 and 3.13 that a vector $x \in \mathbb{C}^n$ is a right (left) eigenvector for a polynomial P with eigenvalue ∞ if and only if x is a right (left) eigenvector for $\mathrm{rev}P$ with eigenvalue 0. Translating statements about infinite eigenvalues to ones about zero eigenvalues allows us to use Theorems 3.8, 3.14, and 4.3 to extend the eigenvector recovery properties of $\mathbb{L}_1(P)$ and $\mathbb{L}_2(P)$ to the eigenvalue ∞.

THEOREM 4.4 (Eigenvector Recovery at ∞). *Let $P(\lambda)$ be an $n \times n$ matrix polynomial of degree k, and $L(\lambda)$ any pencil in $\mathbb{L}_1(P)$ (resp., $\mathbb{L}_2(P)$) with nonzero right (left) ansatz vector v. Then $x \in \mathbb{C}^n$ is a right (left) eigenvector for $P(\lambda)$ with eigenvalue ∞ if and only if $e_1 \otimes x$ is a right (left) eigenvector for $L(\lambda)$ with eigenvalue ∞. If, in addition, P is regular and $L \in \mathbb{L}_1(P)$ (resp., $\mathbb{L}_2(P)$) is a linearization for P, then every right (left) eigenvector of L with eigenvalue ∞ is of the form $e_1 \otimes x$ for some right (left) eigenvector x of P with eigenvalue ∞.*

Proof. We give the proof only for right eigenvectors of $L \in \mathbb{L}_1(P)$ here. The argument for recovery of left eigenvectors of $L \in \mathbb{L}_2(P)$ is essentially the same, given the analogues of Theorems 4.1 and 4.3 for $\mathbb{L}_2(P)$.

For any $L(\lambda)$ define $\widehat{L}(\lambda) := \mathrm{rev}L(\lambda) \cdot (R_k \otimes I)$. Then the reasoning used in Theorem 4.1 to obtain (4.3) shows that $L \in \mathbb{L}_1(P) \Rightarrow \widehat{L} \in \mathbb{L}_1(\mathrm{rev}P)$, with the same nonzero right ansatz vector v. By Theorem 3.8 we know that x is a right eigenvector for $\mathrm{rev}P$ with eigenvalue 0 if and only if $\Lambda \otimes x = e_k \otimes x$ is a right eigenvector for $\widehat{L}$ with eigenvalue 0. However, $e_k \otimes x$ is a right eigenvector for $\widehat{L}$ if and only if $e_1 \otimes x = (R_k \otimes I)(e_k \otimes x)$ is a right eigenvector for $\mathrm{rev}L$, both with eigenvalue 0. This establishes the first part of the theorem.

If P is regular and $L \in \mathbb{L}_1(P)$ is a linearization for P, then by Theorem 4.3 $\widehat{L} \in \mathbb{L}_1(\mathrm{rev}P)$ is a linearization for $\mathrm{rev}P$. Theorem 3.8 then implies that every right eigenvector of $\widehat{L}$ with eigenvalue 0 is of the form $e_k \otimes x$, where x is a right eigenvector of $\mathrm{rev}P$ with eigenvalue 0; equivalently every right eigenvector of $\mathrm{rev}L$ with eigenvalue 0 is of the form $e_1 \otimes x$ for some right eigenvector x of $\mathrm{rev}P$ with eigenvalue 0. This establishes the second part of the theorem. ∎

4.1. Linearization conditions. A useful by-product of the proof of Theorem 4.3 is a simple procedure for generating a symbolic "linearization condition" for any given pencil $L \in \mathbb{L}_1(P)$, i.e., a necessary and sufficient condition (in terms of the data in P) for L to be a linearization for P. We describe this procedure and then illustrate with some examples.

PROCEDURE TO DETERMINE THE LINEARIZATION CONDITION FOR A PENCIL IN $\mathbb{L}_1(P)$:

(1) Suppose that $P(\lambda)$ is a regular polynomial and $L(\lambda) = \lambda X + Y \in \mathbb{L}_1(P)$ has nonzero right ansatz vector $v \in \mathbb{F}^k$, i.e., $L(\lambda) \cdot (\Lambda \otimes I_n) = v \otimes P(\lambda)$.
(2) Select any nonsingular matrix M such that $Mv = \alpha e_1$.
(3) Apply the corresponding block-transformation $M \otimes I_n$ to $L(\lambda)$ to produce $\widetilde{L}(\lambda) := (M \otimes I_n)L(\lambda)$, which must be of the form

$$\widetilde{L}(\lambda) = \lambda \widetilde{X} + \widetilde{Y} = \lambda \left[\begin{array}{c|c} \widetilde{X}_{11} & \widetilde{X}_{12} \\ \hline 0 & -Z \end{array}\right] + \left[\begin{array}{c|c} \widetilde{Y}_{11} & \widetilde{Y}_{12} \\ \hline Z & 0 \end{array}\right], \tag{4.4}$$

where $\widetilde{X}_{11}$ and $\widetilde{Y}_{12}$ are $n \times n$. Since only Z is of interest here, it suffices to form just $\widetilde{Y} = (M \otimes I_n)Y$.
(4) Extract $\boxed{\det Z \neq 0}$, the linearization condition for $L(\lambda)$.

Note that this procedure can readily be implemented as a numerical algorithm to check if a pencil in $\mathbb{L}_1(P)$ is a linearization: choose M to be unitary, e.g., a Householder reflector, then use a rank revealing factorization such as the QR-decomposition with column pivoting or the singular value decomposition to check if Z is nonsingular.

Example 4.5. Consider the general quadratic polynomial $P(\lambda) = \lambda^2 A + \lambda B + C$ (assumed to be regular) and the following pencils in $\mathbb{L}_1(P)$:

$$L_1(\lambda) = \lambda \begin{bmatrix} A & B+C \\ A & 2B-A \end{bmatrix} + \begin{bmatrix} -C & C \\ A-B & C \end{bmatrix}, \quad L_2(\lambda) = \lambda \begin{bmatrix} 0 & -B \\ A & B-C \end{bmatrix} + \begin{bmatrix} B & 0 \\ C & C \end{bmatrix}.$$

Since

$$\begin{bmatrix} A & B+C \\ A & 2B-A \end{bmatrix} \boxplus \begin{bmatrix} -C & C \\ A-B & C \end{bmatrix} = \begin{bmatrix} A & B & C \\ A & B & C \end{bmatrix},$$

we have $L_1(\lambda) \in \mathbb{L}_1(P)$ with right ansatz vector $v = \begin{bmatrix} 1 & 1 \end{bmatrix}^T$. Subtracting the first entry from the second reduces v to e_1, and the corresponding block-row-operation on Y yields

$$\widetilde{Y} = \begin{bmatrix} -C & C \\ A-B+C & 0 \end{bmatrix}.$$

Hence $Z = A - B + C$, and $\det(A - B + C) = \det P(-1) \neq 0$ is the linearization condition. Thus $L_1(\lambda)$ is a linearization for P if and only if $\lambda = -1$ is *not* an eigenvalue of P. On the other hand, for $L_2(\lambda)$ we have

$$\begin{bmatrix} 0 & -B \\ A & B-C \end{bmatrix} \boxplus \begin{bmatrix} B & 0 \\ C & C \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ A & B & C \end{bmatrix},$$

and so $L_2(\lambda) \in \mathbb{L}_1(P)$ with $v = [0 \;\; 1]^T$. Permuting the entries of v gives e_1, and applying the analogous block-row-permutation to Y yields

$$\widetilde{Y} = \begin{bmatrix} C & C \\ B & 0 \end{bmatrix}.$$

Thus $Z = \widetilde{Y}_{21} = B$, and so $\det B \neq 0$ is the linearization condition for $L_2(\lambda)$.

The next example shows that the linearization condition for a pencil in $\mathbb{L}_1(P)$ may depend on some nonlinear combination of the data in P, and thus its meaning may not be so easy to interpret.

Example 4.6. Consider the general cubic polynomial $P(\lambda) = \lambda^3 A + \lambda^2 B + \lambda C + D$ (again assumed to be regular) and the pencil

$$L_3(\lambda) = \lambda X + Y = \lambda \begin{bmatrix} A & 0 & 2C \\ -2A & -B-C & D-4C \\ 0 & A & -I \end{bmatrix} + \begin{bmatrix} B & -C & D \\ C-B & 2C-D & -2D \\ -A & I & 0 \end{bmatrix}$$

in $\mathbb{L}_1(P)$. Since $X \boxplus\!\!\rightarrow Y = \begin{bmatrix} 1 & -2 & 0 \end{bmatrix}^T \otimes \begin{bmatrix} A & B & C & D \end{bmatrix}$, we have $v = \begin{bmatrix} 1 & -2 & 0 \end{bmatrix}^T$. Adding twice the first block-row of Y to the second block-row of Y gives

$$Z = \begin{bmatrix} B+C & -D \\ -A & I \end{bmatrix},$$

and hence the linearization condition $\det Z = \det(B + C - DA) \neq 0$. (Recall that for $n \times n$ blocks W, X, Y, Z with $YZ = ZY$, we have $\det \left[\begin{smallmatrix} W & X \\ Y & Z \end{smallmatrix}\right] = \det(WZ - XY)$. See [11].)

We have seen in this section that each pencil in $\mathbb{L}_1(P)$ has its own particular condition on the coefficient matrices of P that must be satisfied in order for the pencil to be a linearization for P. From this point of view it seems conceivable that there could be polynomials P for which very few of the pencils in $\mathbb{L}_1(P)$ are actually linearizations for P. However, the following result shows that this never happens; when P is regular the "bad" pencils in $\mathbb{L}_1(P)$ always form a very sparse subset of $\mathbb{L}_1(P)$.

THEOREM 4.7 (Linearizations Are Generic in $\mathbb{L}_1(P)$). *For any regular $n \times n$ matrix polynomial $P(\lambda)$ of degree k, almost every pencil in $\mathbb{L}_1(P)$ is a linearization for $P(\lambda)$. (Here by "almost every" we mean for all but a closed, nowhere dense set of measure zero in $\mathbb{L}_1(P)$.)*

Proof. Let $d = \dim \mathbb{L}_1(P) = k + (k-1)kn^2$, and let $L_1(\lambda), L_2(\lambda), \ldots, L_d(\lambda)$ be any fixed basis for $\mathbb{L}_1(P)$. Since any $L(\lambda) \in \mathbb{L}_1(P)$ can be uniquely expressed as a linear combination

$$L(\lambda) \;=\; \beta_1 L_1(\lambda) + \beta_2 L_2(\lambda) + \cdots + \beta_d L_d(\lambda),$$

we can view $\det L(\lambda)$ as a polynomial in λ whose coefficients $c_0, c_1, c_2, \ldots, c_{kn}$ are each polynomial functions of $\beta_1, \ldots, \beta_d$, that is, $c_i = c_i(\beta_1, \ldots, \beta_d)$.

Now by Theorem 4.3 we know that $L(\lambda) \in \mathbb{L}_1(P)$ *fails* to be a linearization for $P(\lambda)$ if and only if $\det L(\lambda) \equiv 0$, equivalently if all the coefficients c_i are zero. Thus the subset of pencils in $\mathbb{L}_1(P)$ that are not linearizations for $P(\lambda)$ can be characterized as the common zero set $\mathcal{Z}$ of the polynomials $\{c_i(\beta_1, \beta_2, \ldots, \beta_d) : 0 \leq i \leq kn\}$, i.e., as an algebraic subset of $\mathbb{F}^d$.

Since proper algebraic subsets of $\mathbb{F}^d$ are well known to be closed, nowhere dense subsets of measure zero, the proof will be complete once we show that $\mathcal{Z}$ is a proper subset of $\mathbb{F}^d$, or equivalently, that there is a pencil in $\mathbb{L}_1(P)$ that is a linearization for P. But this is immediate: the first companion form $C_1(\lambda)$ for $P(\lambda)$ is in $\mathbb{L}_1(P)$ and is always a linearization for P (see [5] or Remark 4.2). ∎

Although $\mathbb{L}_1(P)$ and $\mathbb{L}_2(P)$ contain a large supply of linearizations for P, there do exist simple linearizations for P that are neither in $\mathbb{L}_1(P)$ nor in $\mathbb{L}_2(P)$. We illustrate this with a recent example from [1].

Example 4.8. For the cubic matrix polynomial $P(\lambda) = \lambda^3 A_3 + \lambda^2 A_2 + \lambda A_1 + A_0$, the pencil

$$L(\lambda) = \lambda \begin{bmatrix} 0 & A_3 & 0 \\ I & A_2 & 0 \\ 0 & 0 & I \end{bmatrix} + \begin{bmatrix} -I & 0 & 0 \\ 0 & A_1 & A_0 \\ 0 & -I & 0 \end{bmatrix}$$

is shown in [1] to be a linearization for P. Using shifted sums, it is easy to see that $L(\lambda)$ is in neither $\mathbb{L}_1(P)$ nor $\mathbb{L}_2(P)$.

4.2. Another view of $\mathbb{L}_1(P)$ and $\mathbb{L}_2(P)$. In section 3 we defined the pencil space $\mathbb{L}_1(P)$ by generalizing one particular property of the first companion form $C_1(\lambda)$ of P. A different connection between $\mathbb{L}_1(P)$ and $C_1(\lambda)$ can be established, which gives an alternative insight into why the pencils in $\mathbb{L}_1(P)$ retain so many of the attractive features of $C_1(\lambda)$. Using the first three steps of the procedure in section 4.1, together with the characterization of $\mathbb{L}_1(P)$-pencils given in Theorem 3.5 and Corollary 3.7, one can show that any $L(\lambda) \in \mathbb{L}_1(P)$ can be factored (non-uniquely) in the form

$$L(\lambda) \;=\; (K \otimes I_n) \left[\begin{array}{c|c} \alpha I_n & U \\ \hline 0 & -Z \end{array}\right] C_1(\lambda), \tag{4.5}$$

where $Z \in \mathbb{F}^{(k-1)n \times (k-1)n}$ is the same as the block Z in Corollary 3.7 and (4.4), and $K \in \mathbb{F}^{k \times k}$ is nonsingular. Note that the scalar $\alpha \in \mathbb{F}$ is zero if and only if the right ansatz vector v of $L(\lambda)$ is zero. This factorization gives another reason why the *right* eigenvectors of pencils in $\mathbb{L}_1(P)$ have the same Kronecker product structure as those of $C_1(\lambda)$, and why pencils in $\mathbb{L}_1(P)$ are either strong linearizations of P (like $C_1(\lambda)$) or singular pencils, depending on the nonsingularity or singularity of the block Z and the scalar α.

In a completely analogous fashion one can factor any $L(\lambda) \in \mathbb{L}_2(P)$ as

$$L(\lambda) \;=\; C_2(\lambda) \left[\begin{array}{c|c} \beta I_n & 0 \\ \hline T & -V \end{array}\right] (H \otimes I_n), \tag{4.6}$$

thus providing a different insight into the *left* eigenvector structure of pencils in $\mathbb{L}_2(P)$, and the fact that almost all pencils in $\mathbb{L}_2(P)$ are strong linearizations for P (like $C_2(\lambda)$).

On the other hand, certain aspects of $\mathbb{L}_1(P)$ and $\mathbb{L}_2(P)$ are less apparent from the point of view of these factorizations. For example, the fact that $\mathbb{L}_1(P)$ and $\mathbb{L}_2(P)$ are vector spaces is no longer so obvious. In addition, the criterion for a pencil to be an element of $\mathbb{L}_1(P)$ or $\mathbb{L}_2(P)$ is now implicit rather than explicit and is therefore rather harder to verify.

We are also interested in the possibility of the existence of pencils that are simultaneously in $\mathbb{L}_1(P)$ and $\mathbb{L}_2(P)$. The factored forms (4.5) and (4.6) might make it seem rather unlikely that there could be any nontrivial pencils in this intersection. However, in the next section we will see (using shifted sums) that this is an erroneous impression.

Finally, it is worth pointing out that the ansatz equations (3.4) and (3.9) enjoy the advantage of being identities in the variable λ, and so can be treated analytically as well as algebraically. This property is exploited in the analysis of the conditioning of eigenvalues of linearizations [8].

5. Double ansatz spaces. So far we have constructed two large vector spaces of pencils $\mathbb{L}_1(P)$ and $\mathbb{L}_2(P)$ for any given matrix polynomial $P(\lambda)$ and shown that when P is regular, almost all of these pencils are linearizations for P. Indeed, these spaces are so large that for any choice of (right or left) ansatz vector there are many degrees of freedom available for choosing a potential linearization in $\mathbb{L}_1(P)$ or $\mathbb{L}_2(P)$ with the given ansatz vector (see Theorem 3.5). This suggests that it might be possible to identify special subspaces of pencils in $\mathbb{L}_1(P)$ or $\mathbb{L}_2(P)$ with additional useful properties.

Recall that one of the key advantages of linearizations in $\mathbb{L}_1(P)$ is that right eigenvectors of P are easily recovered from right eigenvectors of the linearization. $\mathbb{L}_2(P)$ offers a similar advantage for recovery of left eigenvectors. Thus it seems natural to consider pencils in the intersection of $\mathbb{L}_1(P)$ and $\mathbb{L}_2(P)$; for these pencils we can simply relate both the right and left eigenvectors of the pencil to those of the original polynomial P. This simultaneous eigenvector recovery property is particularly important in the investigation of the conditioning of linearizations [8]. Therefore we make the following definition.

DEFINITION 5.1 (Double ansatz spaces). *For any $n \times n$ matrix polynomial P of degree k, the double ansatz space of P is*

$$\mathbb{DL}(P) := \mathbb{L}_1(P) \cap \mathbb{L}_2(P),$$

i.e., the set of $kn \times kn$ pencils $L(\lambda)$ that simultaneously satisfy

$$\text{(5.1)} \quad \text{a "right ansatz"} \qquad L(\lambda) \cdot (\Lambda \otimes I) \;=\; v \otimes P(\lambda) \;\; \text{for some} \;\; v \in \mathbb{F}^k,$$

$$\text{(5.2)} \quad \text{and a "left ansatz"} \qquad (\Lambda^T \otimes I) \cdot L(\lambda) \;=\; w^T \otimes P(\lambda) \;\; \text{for some} \;\; w \in \mathbb{F}^k.$$

The rest of this paper is devoted to developing some of the basic properties of $\mathbb{DL}(P)$-spaces, additional aspects of which are explored in [7], [8], [12]. In this section we characterize $\mathbb{DL}(P)$ and show how all the pencils in $\mathbb{DL}(P)$ may be constructed. In section 6 we reconsider the "linearization condition" discussed in section 4. As illustrated by Example 4.6, the intrinsic meaning of this condition can sometimes be rather obscure. However, we will see that for pencils in $\mathbb{DL}(P)$ this condition can always be expressed in a way that makes its meaning transparent.

A priori, the right and left ansatz vectors of a pencil in $\mathbb{DL}(P)$ may be any pair $v, w \in \mathbb{F}^k$. However, it turns out that only pairs with $v = w$ can ever be realized by a $\mathbb{DL}(P)$-pencil. To show this, we first need to determine when the equations $X \boxplus\!\!\rightarrow Y = S$ and $X \underset{\downarrow}{\boxplus} Y = T$ can be solved simultaneously for X and Y.

PROPOSITION 5.2. *Let $S = [S_{ij}]$ and $T = [T_{ji}]$ be block matrices of size $kn \times (k+1)n$ and $(k+1)n \times kn$, respectively, where $S_{ij}, T_{ji} \in \mathbb{F}^{n\times n}$ for $i = 1, \ldots, k$ and $j = 1, \ldots, k+1$. Then there exist block $k \times k$ matrices $X = [X_{ij}]$, $Y = [Y_{ij}]$ with blocks $X_{ij}, Y_{ij} \in \mathbb{F}^{n\times n}$ for $i, j = 1, \ldots, k$ such that*

$$\text{(5.3)} \qquad X \boxplus\!\!\rightarrow Y = S \quad \text{and} \quad X \underset{\downarrow}{\boxplus} Y = T$$

if and only if for $j = 1, \ldots, k$ the blocks of S and T satisfy the compatibility conditions

$$\text{(5.4)} \qquad T_{jj} + \sum_{\mu=1}^{j-1} (T_{\mu,2j-\mu} - S_{\mu,2j-\mu}) = S_{jj} + \sum_{\mu=1}^{j-1} (S_{2j-\mu,\mu} - T_{2j-\mu,\mu})$$

and

$$\text{(5.5)} \qquad \sum_{\mu=1}^{j} (S_{\mu,2j+1-\mu} - T_{\mu,2j+1-\mu}) = \sum_{\mu=1}^{j} (T_{2j+1-\mu,\mu} - S_{2j+1-\mu,\mu}).$$

(*Here, $S_{\nu,\eta} = 0 = T_{\eta,\nu}$ whenever $(\nu,\eta) \notin \{1,\dots,k\} \times \{1,\dots,k+1\}$.*) *If* (5.3) *has a solution, then X and Y are uniquely determined by the formulas*

$$X_{ij} = T_{ij} + \sum_{\mu=1}^{i-1}(T_{\mu,j+i-\mu} - S_{\mu,j+i-\mu}), Y_{ij} = \sum_{\mu=1}^{i}(S_{\mu,j+i+1-\mu} - T_{\mu,j+i+1-\mu}), \tag{5.6}$$

$$X_{ji} = S_{ji} + \sum_{\mu=1}^{i-1}(S_{j+i-\mu,\mu} - T_{j+i-\mu,\mu}), Y_{ji} = \sum_{\mu=1}^{i}(T_{j+i+1-\mu,\mu} - S_{j+i+1-\mu,\mu}), \tag{5.7}$$

for $i,j = 1,\dots,k$ and $j \geq i$.

Proof. Due to its technical nature, the proof is provided in Appendix A. ∎

We are now in a position to show not only that any $\mathbb{DL}(P)$-pencil has its right ansatz vector equal to its left ansatz vector, but also that every $v \in \mathbb{F}^k$ is actually realized as the ansatz vector of a pencil in $\mathbb{DL}(P)$, indeed of a unique pencil in $\mathbb{DL}(P)$. Note that this result does not require any regularity assumption on P.

THEOREM 5.3. *Let $P(\lambda) = \sum_{i=0}^{k} \lambda^i A_i$ be a matrix polynomial with coefficients in $\mathbb{F}^{n\times n}$ and $A_k \neq 0$. Then for vectors $v = (v_1,\dots,v_k)^T$ and $w = (w_1,\dots,w_k)^T$ in $\mathbb{F}^k$ there exists a $kn \times kn$ matrix pencil $L(\lambda) = \lambda X + Y$ that simultaneously satisfies*

$$L(\lambda)\cdot(\Lambda \otimes I) = v \otimes P(\lambda) \quad \text{and} \quad (\Lambda^T \otimes I)\cdot L(\lambda) = w^T \otimes P(\lambda) \tag{5.8}$$

if and only if $v = w$. In this case, if $X = [X_{ij}]$ and $Y = [Y_{ij}]$ are written as block matrices with $n \times n$ blocks X_{ij} and Y_{ij}, then X and Y are uniquely determined by v. In particular, setting $v_0 := 0$, $v_\mu := 0$, and $A_\mu := 0 \in \mathbb{F}^{n\times n}$ for $\mu < 0$ or $\mu > k$, the blocks of X and Y satisfy the formulas

$$X_{ij} = v_{\max(i,j)} A_{k+1-\min(i,j)} + \sum_{\mu=1}^{\min(i-1,j-1)} (v_{j+i-\mu}A_{k+1-\mu} - v_\mu A_{k+1-j-i+\mu}), \tag{5.9}$$

$$Y_{ij} = \sum_{\mu=1}^{\min(i,j)} (v_\mu A_{k-j-i+\mu} - v_{j+i+1-\mu}A_{k+1-\mu}), \qquad i,j = 1,\dots,k. \tag{5.10}$$

Proof. See Appendix B for the proof. ∎

In light of the results in Theorem 5.3, we no longer need to refer separately to the right and left ansatz vectors of a pencil in $\mathbb{DL}(P)$. It suffices to say *the* ansatz vector v of $L \in \mathbb{DL}(P)$, and it is to be understood that v plays both roles.

We can also concisely summarize the result of Theorem 5.3 in a slightly different way. Viewing $\mathbb{DL}(P)$ as a special subspace of $\mathbb{L}_1(P)$, consider the multiplication map $\mathcal{M}$ (introduced in the proof of Theorem 3.5) restricted to the subspace $\mathbb{DL}(P)$. Then the following is an immediate corollary of Theorem 5.3.

COROLLARY 5.4. *For any polynomial P, the map $\mathbb{DL}(P) \xrightarrow{\mathcal{M}} \mathcal{V}_P$ is an isomorphism.*

Thus once an ansatz vector v has been chosen, a pencil from $\mathbb{DL}(P)$ is uniquely determined and can be computed using the formulas of Theorem 5.3.

Another significant property of $\mathbb{DL}(P)$ is worth mentioning here. A matrix polynomial is *symmetric* when all its coefficient matrices are symmetric. For symmetric P, a simple argument shows that every pencil in $\mathbb{DL}(P)$ is also symmetric: $L \in \mathbb{DL}(P)$ with ansatz vector v implies that L^T is also in $\mathbb{DL}(P)$ with the *same* ansatz vector v, and then $L = L^T$ follows from the uniqueness statement of Theorem 5.3.

TABLE 1
Some pencils in $\mathbb{DL}(P)$ for the general quadratic $P(\lambda) = \lambda^2 A + \lambda B + C$. Linearization condition found using procedure in section 4.1.

v	$L(\lambda) \in \mathbb{DL}(P)$ for given v	Linearization condition
$\begin{bmatrix}1\\0\end{bmatrix}$	$\lambda\begin{bmatrix}A & 0\\0 & -C\end{bmatrix} + \begin{bmatrix}B & C\\C & 0\end{bmatrix}$	$\det(C) \neq 0$
$\begin{bmatrix}0\\1\end{bmatrix}$	$\lambda\begin{bmatrix}0 & A\\A & B\end{bmatrix} + \begin{bmatrix}-A & 0\\0 & C\end{bmatrix}$	$\det(A) \neq 0$
$\begin{bmatrix}1\\1\end{bmatrix}$	$\lambda\begin{bmatrix}A & A\\A & B-C\end{bmatrix} + \begin{bmatrix}B-A & C\\C & C\end{bmatrix}$	$\det(A-B+C) = \det[P(-1)] \neq 0$
$\begin{bmatrix}\alpha\\\beta\end{bmatrix}$	$\lambda\begin{bmatrix}\alpha A & \beta A\\\beta A & \beta B-\alpha C\end{bmatrix} + \begin{bmatrix}\alpha B-\beta A & \alpha C\\\alpha C & \beta C\end{bmatrix}$	$\det(\beta^2 A - \alpha\beta B + \alpha^2 C) \neq 0$; equivalently, $\det\left[P(-\frac{\beta}{\alpha})\right] \neq 0$

TABLE 2
Some pencils in $\mathbb{DL}(P)$ for the general cubic $P(\lambda) = \lambda^3 A + \lambda^2 B + \lambda C + D$. Linearization condition found using procedure in section 4.1.

v	$L(\lambda) \in \mathbb{DL}(P)$ for given v	Linearization condition
$\begin{bmatrix}1\\0\\0\end{bmatrix}$	$\lambda\begin{bmatrix}A & 0 & 0\\0 & -C & -D\\0 & -D & 0\end{bmatrix} + \begin{bmatrix}B & C & D\\C & D & 0\\D & 0 & 0\end{bmatrix}$	$\det D \neq 0$
$\begin{bmatrix}0\\1\\0\end{bmatrix}$	$\lambda\begin{bmatrix}0 & A & 0\\A & B & 0\\0 & 0 & -D\end{bmatrix} + \begin{bmatrix}-A & 0 & 0\\0 & C & D\\0 & D & 0\end{bmatrix}$	$\det A \cdot \det D \neq 0$
$\begin{bmatrix}0\\0\\1\end{bmatrix}$	$\lambda\begin{bmatrix}0 & 0 & A\\0 & A & B\\A & B & C\end{bmatrix} + \begin{bmatrix}0 & -A & 0\\-A & -B & 0\\0 & 0 & D\end{bmatrix}$	$\det A \neq 0$
$\begin{bmatrix}1\\0\\-1\end{bmatrix}$	$\lambda\begin{bmatrix}A & 0 & -A\\0 & -A-C & -B-D\\-A & -B-D & -C\end{bmatrix} + \begin{bmatrix}B & A+C & D\\A+C & B+D & 0\\D & 0 & -D\end{bmatrix}$	$\det\begin{bmatrix}A+C & B+D\\B+D & A+C\end{bmatrix} \neq 0$
$\begin{bmatrix}1\\1\\1\end{bmatrix}$	$\lambda\begin{bmatrix}A & A & A\\A & A+B-C & B-D\\A & B-D & C-D\end{bmatrix} + \begin{bmatrix}B-A & C-A & D\\C-A & C+D-B & D\\D & D & D\end{bmatrix}$	$\det\begin{bmatrix}C-B & A-B+D\\A-B+D & A-C+D\end{bmatrix} \neq 0$

Examples of pencils in $\mathbb{DL}(P)$ for $k = 2$ and $k = 3$ may be found in Tables 1 and 2. Using shifted sums, one easily verifies that these examples are indeed in both $\mathbb{L}_1(P)$ *and* $\mathbb{L}_2(P)$, with the *same* right and left ansatz vector v. Note that if A, B, C, and D are symmetric, then so are all the pencils in these examples. Symmetric linearizations are studied in more detail in [7].

Perhaps the most striking property of the space $\mathbb{DL}(P)$ is that the linearization condition for each pencil in $\mathbb{DL}(P)$ can be directly linked to its ansatz vector v, as will be seen in the next section.

6. The eigenvalue exclusion theorem. We now establish a connection between the linearization condition of any pencil $L \in \mathbb{DL}(P)$ and the ansatz vector v that defines L. For example, consider the cubic polynomial $P(\lambda) = \lambda^3 A + \lambda^2 B + \lambda C + D$ and the pencil

$$L(\lambda) = \lambda \begin{bmatrix} A & 0 & -A \\ 0 & -A-C & -B-D \\ -A & -B-D & -C \end{bmatrix} + \begin{bmatrix} B & A+C & D \\ A+C & B+D & 0 \\ D & 0 & -D \end{bmatrix}$$

in $\mathbb{DL}(P)$ with ansatz vector $v = \begin{bmatrix} 1 & 0 & -1 \end{bmatrix}^T$. Using the procedure in section 4.1, one easily finds that

$$\det \begin{bmatrix} A+C & B+D \\ B+D & A+C \end{bmatrix} \neq 0 \tag{6.1}$$

is the linearization condition for $L(\lambda)$. (See also Table 2.) Now it is not immediately clear what the meaning of this condition is, or even whether it has any intrinsic meaning at all. However, the identity

$$\begin{bmatrix} 0 & I \\ I & I \end{bmatrix} \begin{bmatrix} A+C & B+D \\ B+D & A+C \end{bmatrix} \begin{bmatrix} I & 0 \\ -I & I \end{bmatrix}$$
$$= \begin{bmatrix} -A+B-C+D & A+C \\ 0 & A+B+C+D \end{bmatrix} = \begin{bmatrix} P(-1) & A+C \\ 0 & P(+1) \end{bmatrix}$$

shows that condition (6.1) is equivalent to saying that neither -1 nor $+1$ is an eigenvalue of the matrix polynomial $P(\lambda)$. Thus in this example we can reinterpret the linearization condition from section 4.1 as an "eigenvalue exclusion" condition.

Why do these particular eigenvalues need to be excluded? And what role, if any, does the ansatz vector $v = \begin{bmatrix} 1 & 0 & -1 \end{bmatrix}^T$ play here? Observe that if we interpret the components of v as the coefficients of a scalar polynomial, then we obtain $x^2 - 1$, whose roots are exactly the eigenvalues that have to be excluded in order to guarantee that $L(\lambda)$ is a linearization for $P(\lambda)$. The goal of this section is to show that this is not merely a coincidence, but rather an instance of a general phenomenon described by the "eigenvalue exclusion theorem."

The main technical result needed to prove this theorem is an explicit formula for the determinant of a pencil $L(\lambda)$ in $\mathbb{DL}(P)$. To aid in the development of this formula we first introduce some notation to be used throughout this section. As before, $P(\lambda) = \sum_{i=0}^{k} \lambda^i A_i$ is an $n \times n$ matrix polynomial with nonzero leading coefficient A_k. The pencil $L(\lambda) \in \mathbb{DL}(P)$ under consideration has ansatz vector $v = [v_1, v_2, \ldots, v_k]^T$, with an associated scalar polynomial defined as follows.

DEFINITION 6.1 (v-polynomial). *With a vector $v = [v_1, v_2, \ldots, v_k]^T \in \mathbb{F}^k$ associate the scalar polynomial*

$$\mathsf{p}(x\,;v) \;=\; v_1 x^{k-1} + v_2 x^{k-2} + \cdots + v_{k-1} x + v_k,$$

*referred to as the "*v*-polynomial" of the vector v. We adopt the convention that $\mathsf{p}(x\,;v)$ has a root at ∞ whenever $v_1 = 0$.*

We also need to introduce the notion of the "Horner shifts" of a polynomial.

DEFINITION 6.2 (Horner shifts). *For any polynomial $p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$ and $0 \leq \ell \leq n$, the "*degree ℓ Horner shift of $p(x)$*" is $p_\ell(x) := a_n x^\ell + a_{n-1} x^{\ell-1} + \cdots + a_{n-\ell+1} x + a_{n-\ell}$.*

Remark 6.3. The polynomials in Definition 6.2 satisfy the recurrence relation

$$\begin{aligned} p_0(x) &= a_n, \\ p_{\ell+1}(x) &= xp_\ell(x) + a_{n-\ell-1} \quad \text{for } 0 \le \ell \le n-1, \\ p_n(x) &= p(x), \end{aligned}$$

and are precisely the polynomials appearing in Horner's method for evaluating the polynomial $p(x)$.

We have seen in Theorem 5.3 that $L(\lambda) \in \mathbb{DL}(P)$ is uniquely determined by the vector v and the polynomial P, so it is not surprising that one can also specify the columns of $L(\lambda)$ in terms of this data. This is done in the next lemma, where a block-column–wise description of $L(\lambda)$ is given. In this description we make extensive use of the standard $k \times k$ nilpotent Jordan block N from (2.4) in the matrix $N \otimes I$, employed here as a block-shift operator.

LEMMA 6.4 (Block-column structure of pencils in $\mathbb{DL}(P)$). *Suppose that $L(\lambda) = \lambda X + Y$ is in $\mathbb{DL}(P)$ with ansatz vector v. Partition X and Y as*

$$X = \begin{bmatrix} X_1 & X_2 & \cdots & X_k \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} Y_1 & \cdots & Y_{k-1} & Y_k \end{bmatrix},$$

where $X_\ell, Y_\ell \in \mathbb{F}^{nk \times n}$, $\ell = 1, \ldots, k$. Then with $Y_0 := 0$, the block-columns Y_ℓ satisfy the recurrence

$$Y_\ell = (N \otimes I)(Y_{\ell-1} - v \otimes A_{k-\ell+1}) + v_\ell \begin{bmatrix} A_{k-1} \\ \vdots \\ A_0 \end{bmatrix}, \quad 1 \le \ell \le k-1, \tag{6.2}$$

$$Y_k = v \otimes A_0. \tag{6.3}$$

The block-columns of X are determined by $X_\ell = -Y_{\ell-1} + v \otimes A_{k-\ell+1}$ for $1 \le \ell \le k$, and the pencil $L(\lambda)$ has the columnwise description

$$L(\lambda) = \left[\begin{array}{c|c|c|c|c} Y_1 & Y_2 - \lambda Y_1 & \cdots & Y_{k-1} - \lambda Y_{k-2} & v \otimes A_0 - \lambda Y_{k-1} \\ +\lambda v \otimes A_k & +\lambda v \otimes A_{k-1} & & +\lambda v \otimes A_2 & +\lambda v \otimes A_1 \end{array}\right]. \tag{6.4}$$

Proof. Let $Y_0 = [Y_{i0}] := 0$, $X_\ell = [X_{i\ell}]$, and $Y_\ell = [Y_{i\ell}]$ for $n \times n$ blocks $Y_{i0}, X_{i\ell}, Y_{i\ell}$, where $i = 1, \ldots, k$. Then we obtain from (5.10) for $1 \le i < \ell \le k-1$ that

$$\begin{aligned} Y_{i\ell} &= \sum_{\mu=1}^{i} (v_\mu A_{k-\ell-i+\mu} - v_{\ell+i+1-\mu} A_{k+1-\mu}) \\ &= \sum_{\mu=1}^{i+1} (v_\mu A_{k-\ell-i+\mu} - v_{\ell+i+1-\mu} A_{k+1-\mu}) - v_{i+1} A_{k+1-\ell} + v_\ell A_{k-i} \\ &= Y_{i+1,\ell-1} - v_{i+1} A_{k+1-\ell} + v_\ell A_{k-i}. \end{aligned}$$

Analogously, we obtain for $1 \le \ell \le i \le k-1$ that

$$\begin{aligned} Y_{i\ell} &= \sum_{\mu=1}^{\ell} (v_\mu A_{k-\ell-i+\mu} - v_{\ell+i+1-\mu} A_{k+1-\mu}) \\ &= \sum_{\mu=1}^{\ell-1} (v_\mu A_{k-\ell-i+\mu} - v_{\ell+i+1-\mu} A_{k+1-\mu}) + v_\ell A_{k-i} - v_{i+1} A_{k+1-\ell} \\ &= Y_{i+1,\ell-1} - v_{i+1} A_{k+1-\ell} + v_\ell A_{k-i}. \end{aligned}$$

Since formula (5.10) also implies $Y_{k\ell} = v_\ell A_0$, we obtain

$$Y_\ell = \begin{bmatrix} Y_{1\ell} \\ \vdots \\ Y_{k-1,\ell} \\ Y_{k\ell} \end{bmatrix} = \begin{bmatrix} Y_{2,\ell-1} \\ \vdots \\ Y_{k,\ell-1} \\ 0 \end{bmatrix} - \begin{bmatrix} v_2 A_{k-\ell+1} \\ \vdots \\ v_k A_{k-\ell+1} \\ 0 \end{bmatrix} + \begin{bmatrix} v_\ell A_k \\ \vdots \\ v_\ell A_1 \\ v_\ell A_0 \end{bmatrix}$$

$$= (N \otimes I)(Y_{\ell-1} - v \otimes A_{k-\ell+1}) + v_\ell \begin{bmatrix} A_{k-1} \\ \vdots \\ A_0 \end{bmatrix}$$

for $\ell = 1, \ldots, k-1$. Noting that (3.6) implies $Y_k = v \otimes A_0$ and $X_\ell + Y_{\ell-1} = v \otimes A_{k-\ell+1}$ for $\ell = 1, \ldots, k$, we immediately obtain (6.4). ∎

Using (6.2), we can now develop a concise formula describing the action of the block-row $\Lambda^T(x) \otimes I$ on the block-column Y_ℓ, where x is a scalar variable taking values in $\mathbb{C}$ and $\Lambda^T(x) := \begin{bmatrix} x^{k-1} & x^{k-2} & \cdots & x & 1 \end{bmatrix}$. This formula will be used repeatedly and plays a central role in the proof of Theorem 6.6. (Note that $\Lambda^T(x)v$ is the same as the scalar v-polynomial $\mathsf{p}(x\,;v)$.)

LEMMA 6.5. *Suppose that $L(\lambda) \in \mathbb{DL}(P)$ with ansatz vector v, and $\mathsf{p}(x\,;v)$ is the v-polynomial of v. Let Y_ℓ denote the ℓth block column of Y in $L(\lambda) = \lambda X + Y$, where $1 \le \ell \le k-1$. Then*

$$\left(\Lambda^T(x) \otimes I\right) Y_\ell \;=\; \mathsf{p}_{\ell-1}(x\,;v)P(x) - x\,\mathsf{p}(x\,;v)P_{\ell-1}(x), \tag{6.5}$$

where $\mathsf{p}_{\ell-1}(x\,;v)$ and $P_{\ell-1}(\lambda)$ are the degree $\ell-1$ Horner shifts of $\mathsf{p}(x\,;v)$ and $P(\lambda)$, respectively.

Proof. The proof will proceed by induction on ℓ. First note that for the $k \times k$ nilpotent Jordan block N, it is easy to check that $\Lambda^T(x)N = \begin{bmatrix} 0 & x^{k-1} & \cdots & x \end{bmatrix} = x\Lambda^T(x) - x^k e_1^T$.

$\boxed{\ell = 1}$: Using (6.2), we have

$$\left(\Lambda^T(x) \otimes I\right) Y_1 \;=\; \left(\Lambda^T(x) \otimes I\right) \left(v_1 \begin{bmatrix} A_{k-1} \\ \vdots \\ A_0 \end{bmatrix} - (N \otimes I)(v \otimes A_k) \right).$$

Simplifying this gives

$$\begin{aligned} \left(\Lambda^T(x) \otimes I\right) Y_1 &= v_1 \left(P(x) - x^k A_k\right) - \left(\Lambda^T(x)N \otimes I\right)(v \otimes A_k) \\ &= v_1 P(x) - v_1 x^k A_k - \left(\left(x\Lambda^T(x) - x^k e_1^T\right)v \otimes A_k\right) \\ &= \mathsf{p}_0(x\,;v)P(x) - v_1 x^k A_k - \left(x\Lambda^T(x)v\right)A_k + \left(x^k e_1^T v\right)A_k \\ &= \mathsf{p}_0(x\,;v)P(x) - v_1 x^k A_k - x\,\mathsf{p}(x\,;v)A_k + v_1 x^k A_k \\ &= \mathsf{p}_0(x\,;v)P(x) - x\,\mathsf{p}(x\,;v)P_0(x), \end{aligned}$$

which establishes (6.5) for $\ell = 1$. The induction hypothesis is now the following:

$$\left(\Lambda^T(x) \otimes I\right) Y_{\ell-1} \;=\; \mathsf{p}_{\ell-2}(x\,;v)P(x) - x\,\mathsf{p}(x\,;v)P_{\ell-2}(x). \tag{6.6}$$

$\boxed{\ell-1 \Rightarrow \ell}$: Starting again with (6.2), we have

$$\begin{aligned}
\left(\Lambda^T(x)\otimes I\right) Y_\ell &= \left(\Lambda^T(x)\otimes I\right)\left((N\otimes I)(Y_{\ell-1} - v\otimes A_{k-\ell+1}) + v_\ell \begin{bmatrix} A_{k-1} \\ \vdots \\ A_0 \end{bmatrix}\right) \\
&= \left(\Lambda^T(x)N\otimes I\right)(Y_{\ell-1} - v\otimes A_{k-\ell+1}) + v_\ell \left(\Lambda^T(x)\otimes I\right)\begin{bmatrix} A_{k-1} \\ \vdots \\ A_0 \end{bmatrix} \\
&= \left(\left(x\Lambda^T(x) - x^k e_1^T\right)\otimes I\right)(Y_{\ell-1} - v\otimes A_{k-\ell+1}) + v_\ell \left(P(x) - x^k A_k\right) \\
&= x\left(\Lambda^T(x)\otimes I\right)Y_{\ell-1} - x^k\left(e_1^T\otimes I\right)Y_{\ell-1} - \left(x\Lambda^T(x)v\right)A_{k-\ell+1} \\
&\qquad + v_1 x^k A_{k-\ell+1} + v_\ell\, P(x) - v_\ell\, x^k A_k.
\end{aligned}$$

Note that $\left(e_1^T\otimes I\right)Y_{\ell-1}$ is the topmost block in $Y_{\ell-1}$ and is equal to $v_1 A_{k-\ell+1} - v_\ell A_k$, by (5.10). Finally, invoking the induction hypothesis (6.6) gives

$$\begin{aligned}
\left(\Lambda^T(x)\otimes I\right) Y_\ell &= x\,\mathsf{p}_{\ell-2}(x\,;v)P(x) - x^2\,\mathsf{p}(x\,;v)P_{\ell-2}(x) - v_1 x^k A_{k-\ell+1} + v_\ell\, x^k A_k \\
&\qquad - x\,\mathsf{p}(x\,;v)A_{k-\ell+1} + v_1 x^k A_{k-\ell+1} + v_\ell\, P(x) - v_\ell\, x^k A_k \\
&= \left(x\,\mathsf{p}_{\ell-2}(x\,;v) + v_\ell\right)P(x) - x\,\mathsf{p}(x\,;v)\left(xP_{\ell-2}(x) + A_{k-\ell+1}\right) \\
&= \mathsf{p}_{\ell-1}(x\,;v)P(x) - x\,\mathsf{p}(x\,;v)P_{\ell-1}(x).
\end{aligned}$$

This completes the proof. ∎

THEOREM 6.6 (Determinant formula for pencils in $\mathbb{DL}(P)$). *Suppose that $L(\lambda)$ is in $\mathbb{DL}(P)$ with nonzero ansatz vector $v = [v_1, v_2, \ldots, v_k]^T$. Assume that v has m leading zeroes with $0 \le m \le k-1$, so that $v_1 = v_2 = \cdots = v_m = 0$, $v_{m+1} \neq 0$ is the first nonzero coefficient of $\mathsf{p}(x\,;v)$, and $\mathsf{p}(x\,;v)$ has $k-m-1$ finite roots in $\mathbb{C}$, counted with multiplicities, denoted here by $r_1, r_2, \ldots, r_{k-m-1}$. Then we have*

(6.7)
$$\det L(\lambda) = \begin{cases} (-1)^{n\cdot\left\lfloor\frac{k}{2}\right\rfloor}(v_1)^{kn}\det\Big(P(r_1)P(r_2)\cdots P(r_{k-1})\Big)\det P(\lambda) & \text{if } m=0, \\ (-1)^s (v_{m+1})^{kn}(\det A_k)^m \det\Big(P(r_1)\cdots P(r_{k-m-1})\Big)\det P(\lambda) & \text{if } m>0, \end{cases}$$

where $s = n\left(m + \left\lfloor\frac{m}{2}\right\rfloor + \left\lfloor\frac{k-m}{2}\right\rfloor\right)$.

Proof. The proof proceeds in three parts.

Part 1. We first consider the case when $m = 0$ (i.e., $v_1 \neq 0$) and $\mathsf{p}(x\,;v)$ has $k-1$ *distinct* finite roots. The strategy of the proof is to reduce $L(\lambda)$ by a sequence of equivalence transformations to a point where the determinant can just be read off.

We begin the reduction process by right-multiplying $L(\lambda)$ by the block-Toeplitz matrix $T(\lambda)$. Recall that $T(\lambda)$ and $G(\lambda)$ denote the unimodular matrix polynomials defined in (2.5), and are related to each other via the factorization in (2.6). Using (6.4) for the description of $L(\lambda)$, an argument very similar to the one used in the proof of Theorem 4.1 yields the block-column–wise description

$$L(\lambda)G(\lambda) = \left[\begin{array}{l|l|c|l} Y_1 + \lambda v\otimes A_k & Y_2 - \lambda Y_1 + \lambda\, v\otimes A_{k-1} & \cdots & Y_{k-1} - \lambda Y_{k-2} + \lambda\, v\otimes A_2 \end{array}\middle|\ v\otimes P(\lambda)\right],$$

and hence

(6.8)
$$L(\lambda)T(\lambda) = \left[\begin{array}{l|l|c|l} Y_1 + \lambda v\otimes P_0(\lambda) & Y_2 + \lambda v\otimes P_1(\lambda) & \cdots & Y_{k-1} + \lambda v\otimes P_{k-2}(\lambda) \end{array}\middle|\ v\otimes P(\lambda)\right].$$

Next we left-multiply by a constant (nonsingular) "Vandermonde-like" matrix M, built block-row–wise from $\Lambda^T(x) := [\,x^{k-1}\,x^{k-2}\,\cdots\,x\;1]$ evaluated at each of the roots of $\mathsf{p}(x\,;v)$,

$$M \;:=\; \begin{bmatrix} e_1^T \\ \Lambda^T(r_1) \\ \Lambda^T(r_2) \\ \vdots \\ \Lambda^T(r_{k-1}) \end{bmatrix} \otimes I \;=\; \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ r_1^{k-1} & r_1^{k-2} & \cdots & r_1 & 1 \\ r_2^{k-1} & r_2^{k-2} & \cdots & r_2 & 1 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ r_{k-1}^{k-1} & r_{k-1}^{k-2} & \cdots & r_{k-1} & 1 \end{bmatrix} \otimes I. \tag{6.9}$$

Using Lemma 6.5 and the fact that $\Lambda^T(r_j)v = \mathsf{p}(r_j\,;v)$, we obtain that

$$\begin{aligned} \big(\Lambda^T(r_j)\otimes I\big)\,(Y_\ell + \lambda v\otimes P_{\ell-1}(\lambda)) \\ = \;\mathsf{p}_{\ell-1}(r_j\,;v)P(r_j) - r_j\,\mathsf{p}(r_j\,;v)P_{\ell-1}(r_j) + \lambda\,\mathsf{p}(r_j\,;v)P_{\ell-1}(\lambda). \end{aligned}$$

Since $r_1,\dots,r_{k-1}$ are the roots of $\mathsf{p}(x\,;v)$, the product $ML(\lambda)T(\lambda)$ simplifies to

$$\left[\begin{array}{cccc|c} * & * & \cdots & * & v_1P(\lambda) \\ \hline \mathsf{p}_0(r_1\,;v)P(r_1) & \mathsf{p}_1(r_1\,;v)P(r_1) & \cdots & \mathsf{p}_{k-2}(r_1\,;v)P(r_1) & 0 \\ \mathsf{p}_0(r_2\,;v)P(r_2) & \mathsf{p}_1(r_2\,;v)P(r_2) & \cdots & \mathsf{p}_{k-2}(r_2\,;v)P(r_2) & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathsf{p}_0(r_{k-1}\,;v)P(r_{k-1}) & \mathsf{p}_1(r_{k-1}\,;v)P(r_{k-1}) & \cdots & \mathsf{p}_{k-2}(r_{k-1}\,;v)P(r_{k-1}) & 0 \end{array}\right].$$

This matrix now factors into

$$\underbrace{\left[\begin{array}{c|ccc} I & & & \\ \hline & P(r_1) & & \\ & & \ddots & \\ & & & P(r_{k-1}) \end{array}\right]}_{=:\,W} \left[\begin{array}{ccc|c} * & \cdots & * & v_1P(\lambda) \\ \hline \mathsf{p}_0(r_1\,;v)I & \cdots & \mathsf{p}_{k-2}(r_1\,;v)I & 0 \\ \vdots & \ddots & \vdots & \vdots \\ \mathsf{p}_0(r_{k-1}\,;v)I & \cdots & \mathsf{p}_{k-2}(r_{k-1}\,;v)I & 0 \end{array}\right],$$

and after reversing the order of the block-columns using $R\otimes I$, we have

$$ML(\lambda)T(\lambda)(R\otimes I) \;=\; W \left[\begin{array}{c|c} v_1P(\lambda) & * \\ \hline \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} & V\otimes I \end{array}\right], \tag{6.10}$$

where

$$\begin{aligned} V &= \begin{bmatrix} \mathsf{p}_{k-2}(r_1\,;v) & \cdots & \mathsf{p}_1(r_1\,;v) & \mathsf{p}_0(r_1\,;v) \\ \vdots & \vdots & \vdots & \vdots \\ \mathsf{p}_{k-2}(r_{k-1}\,;v) & \cdots & \mathsf{p}_1(r_{k-1}\,;v) & \mathsf{p}_0(r_{k-1}\,;v) \end{bmatrix} \\ &= \begin{bmatrix} (v_1r_1^{k-2}+\cdots+\,v_{k-2}r_1\,+\,v_{k-1}) & \cdots & (v_1r_1\,+\,v_2) & v_1 \\ \vdots & \vdots & \vdots & \vdots \\ (v_1r_{k-1}^{k-2}+\cdots+v_{k-2}r_{k-1}+\,v_{k-1}) & \cdots & (v_1r_{k-1}+\,v_2) & v_1 \end{bmatrix}. \end{aligned}$$

All that remains is to observe that V can be reduced by ($\det = +1$) column operations to

$$v_1 \cdot \begin{bmatrix} r_1^{k-2} & r_1^{k-3} & \cdots & r_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{k-1}^{k-2} & r_{k-1}^{k-3} & \cdots & r_{k-1} & 1 \end{bmatrix}, \tag{6.11}$$

so $\det(V \otimes I) = v_1^{(k-1)n} \det M$. Taking determinants on both sides of (6.10) now gives

$$\begin{aligned} \det M \ \cdot \ & \det L(\lambda) \cdot \det T(\lambda) \cdot \det(R \otimes I) \\ & = \det\left(P(r_1)P(r_2)\cdots P(r_{k-1})\right) \cdot \det\left(v_1 P(\lambda)\right) \cdot \det(V \otimes I). \end{aligned}$$

Since

$$\det(R \otimes I) = \det(R_k \otimes I_n) = (\det R_k)^n (\det I_n)^k = (-1)^{n \cdot \lfloor \frac{k}{2} \rfloor} \tag{6.12}$$

and $\det T(\lambda) = +1$, this simplifies to the desired result

$$\det L(\lambda) \ = \ (-1)^{n \cdot \lfloor \frac{k}{2} \rfloor} (v_1)^{kn} \det\left(P(r_1)P(r_2)\cdots P(r_{k-1})\right) \det P(\lambda). \tag{6.13}$$

This completes the argument for the case when $m = 0$ and the $k-1$ roots of $\mathsf{p}(x\,;v)$ are all distinct.

Part 2. We now describe how to modify this argument to handle $m > 0$, i.e., the first nonzero coefficient of $\mathsf{p}(x\,;v)$ is v_{m+1}. We will continue to assume that the $k-m-1$ finite roots of $\mathsf{p}(x\,;v)$ are all distinct.

We start out the same way as before, postmultiplying $L(\lambda)$ by $T(\lambda)$ to get (6.8). But then, instead of M in (6.9), we use all available finite roots of $\mathsf{p}(x\,;v)$ to define the following modified version of M:

$$\widehat{M} := \begin{bmatrix} e_1^T \\ \vdots \\ e_{m+1}^T \\ \Lambda^T(r_1) \\ \vdots \\ \Lambda^T(r_{k-m-1}) \end{bmatrix} \otimes I_n \ = \ \left[\begin{array}{ccccc} & I_{m+1} & & 0 & \\ \hline r_1^{k-1} & r_1^{k-2} & \cdots & r_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{k-m-1}^{k-1} & r_{k-m-1}^{k-2} & \cdots & r_{k-m-1} & 1 \end{array}\right] \otimes I_n. \tag{6.14}$$

Now simplify the product $\widehat{M}L(\lambda)T(\lambda)$ using Lemma 6.5 and $\Lambda^T(r_\ell)v = \mathsf{p}(r_\ell\,;v) = 0$ as before, as well as the fact that $v_1 = v_2 = \cdots = v_m = 0$, which implies that $\mathsf{p}_0(x\,;v)$, $\mathsf{p}_1(x\,;v)$, …, $\mathsf{p}_{m-1}(x\,;v)$ are all zero polynomials. Then we obtain

$$\widehat{M}L(\lambda)T(\lambda)$$

$$= \left[\begin{array}{ccc|c} & & & 0 \\ & * & & \vdots \\ & & & 0 \\ \hline * & \cdots & * & v_{m+1}P(\lambda) \\ \hline \mathsf{p}_0(r_1\,;v)P(r_1) & \cdots & \mathsf{p}_{k-2}(r_1\,;v)P(r_1) & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \mathsf{p}_0(r_{k-m-1}\,;v)P(r_{k-m-1}) & \cdots & \mathsf{p}_{k-2}(r_{k-m-1}\,;v)P(r_{k-m-1}) & 0 \end{array}\right]$$

$$= \left[\begin{array}{c|ccc|c} & & & & 0 \\ B & & * & & \vdots \\ & & & & 0 \\ \hline * & * & \cdots & * & v_{m+1}P(\lambda) \\ \hline & \mathsf{p}_m(r_1\,;v)P(r_1) & \cdots & \mathsf{p}_{k-2}(r_1\,;v)P(r_1) & 0 \\ 0 & \vdots & \vdots & \vdots & \vdots \\ & \mathsf{p}_m(r_{k-m-1}\,;v)P(r_{k-m-1}) & \cdots & \mathsf{p}_{k-2}(r_{k-m-1}\,;v)P(r_{k-m-1}) & 0 \end{array}\right], \quad \begin{array}{c} \\ \\ \\ \\ \\ \\ n \end{array}$$

$$\begin{array}{ccc} mn & & (k-m-1)n \end{array}$$

where the $mn \times mn$ block B can also be seen to have some further structure. First note that because of the structure of $\widehat{M}$, the block B in $\widehat{M}L(\lambda)T(\lambda)$ is exactly the same as the corresponding block in $L(\lambda)T(\lambda)$ in (6.8), which is just the first mn rows of

$$\left[\begin{array}{c|c|c|c} Y_1 & Y_2 & \cdots & Y_m \\ +\lambda v \otimes P_0(\lambda) & +\lambda v \otimes P_1(\lambda) & & +\lambda v \otimes P_{m-1}(\lambda) \end{array}\right].$$

But because $v_1 = v_2 = \cdots = v_m = 0$, the terms $\lambda v \otimes P_i(\lambda)$ make no contribution to these first mn rows. So B is the same as the first mn rows of

$$[Y_1|Y_2|\cdots|Y_m].$$

Using the recurrence (6.2) from Lemma 6.4 with $1 \le \ell \le m$, we can now show that B is actually block anti-triangular. When $\ell = 1$ we have $Y_1 = -Nv \otimes A_k$. Since the first m entries of Nv are $[v_2,\, v_3, \ldots,\, v_{m+1}]^T = [0,\, 0, \ldots, v_{m+1}]^T$, we see that the first block-column of B is $[0, \ldots, 0, -v_{m+1}A_k^T]^T$. With $\ell = 2$ we have $Y_2 = (N \otimes I)Y_1 - Nv \otimes A_{k-1}$, whose first mn rows are

$$\begin{bmatrix} 0 \\ \vdots \\ 0 \\ -v_{m+1}A_k \\ * \end{bmatrix} - \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ -v_{m+1}A_{k-1} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -v_{m+1}A_k \\ * \end{bmatrix}.$$

By induction, we then see that the first mn rows of Y_ℓ for $1 \le \ell \le m$ look like

$$\left[0, \ldots, 0,\, -v_{m+1}A_k^T,\, *, \ldots, *\right]^T,$$

with $m-\ell$ leading blocks of zeroes. Thus B has the block anti-triangular form

$$B = -v_{m+1} \cdot \begin{bmatrix} 0 & \cdots & 0 & A_k \\ \vdots & \iddots & \iddots & * \\ 0 & A_k & \iddots & \vdots \\ A_k & * & \cdots & * \end{bmatrix},$$

and so $\widehat{M}L(\lambda)T(\lambda)$ is equal to

$$\left[\begin{array}{ccc|ccc|c} 0 & & -v_{m+1}A_k & & & & 0 \\ & \iddots & & & * & & \vdots \\ -v_{m+1}A_k & & * & & & & 0 \\ \hline * & & & * & \cdots & * & v_{m+1}P(\lambda) \\ \hline & 0 & & \mathsf{p}_m(r_1\,;v)P(r_1) & \cdots & \mathsf{p}_{k-2}(r_1\,;v)P(r_1) & 0 \\ & & & \vdots & \vdots & \vdots & \vdots \\ & & & \mathsf{p}_m(r_{k-m-1}\,;v)P(r_{k-m-1}) & \cdots & \mathsf{p}_{k-2}(r_{k-m-1}\,;v)P(r_{k-m-1}) & 0 \end{array}\right].$$

Performing some block-column permutations gives us

(6.15)

$$\widehat{M}L(\lambda)T(\lambda)\left((R_m \oplus R_{k-m}) \otimes I_n\right)$$

$$= \left[\begin{array}{ccc|c|ccc} -v_{m+1}A_k & & 0 & 0 & & & \\ & \ddots & & \vdots & & * & \\ * & & -v_{m+1}A_k & 0 & & & \\ \hline & * & & v_{m+1}P(\lambda) & * & \cdots & * \\ \hline & & & 0 & \mathsf{p}_{k-2}(r_1\,;v)P(r_1) & \cdots & \mathsf{p}_m(r_1\,;v)P(r_1) \\ & 0 & & \vdots & \vdots & \vdots & \vdots \\ & & & 0 & \mathsf{p}_{k-2}(r_{k-m-1}\,;v)P(r_{k-m-1}) & \cdots & \mathsf{p}_m(r_{k-m-1}\,;v)P(r_{k-m-1}) \end{array}\right],$$

which after factoring becomes

$$\text{(6.16)} \quad \left[\begin{array}{c|c|c} (-v_{m+1}I_m)\otimes I_n & 0 & 0 \\ \hline 0 & v_{m+1}I_n & 0 \\ \hline 0 & 0 & \widehat{W} \end{array}\right] \left[\begin{array}{ccc|c|c} A_k & & 0 & & \\ & \ddots & & 0 & * \\ * & & A_k & & \\ \hline & 0 & & P(\lambda) & * \\ \hline & 0 & & 0 & \widehat{V}\otimes I_n \end{array}\right],$$

where $\widehat{W} = \operatorname{diag}(P(r_1), \ldots, P(r_{k-m-1}))$ and

$$\widehat{V} = \begin{bmatrix} \mathsf{p}_{k-2}(r_1\,;v) & \cdots & \mathsf{p}_m(r_1\,;v) \\ \vdots & \vdots & \vdots \\ \mathsf{p}_{k-2}(r_{k-m-1}\,;v) & \cdots & \mathsf{p}_m(r_{k-m-1}\,;v) \end{bmatrix}$$

$$= \begin{bmatrix} (v_{m+1}r_1^{k-m-2} + \cdots + v_{k-1}) & \cdots & (v_{m+1}r_1 + v_{m+2}) & v_{m+1} \\ \vdots & \vdots & \vdots & \vdots \\ (v_{m+1}r_{k-m-1}^{k-m-2} + \cdots + v_{k-1}) & \cdots & (v_{m+1}r_{k-m-1} + v_{m+2}) & v_{m+1} \end{bmatrix}.$$

Since $v_{m+1} \neq 0$, this $(k-m-1)\times(k-m-1)$ matrix $\widehat{V}$ can be reduced by (det $= +1$) column operations in a manner analogous to the reduction of V in (6.11), so we see that

$$\det(\widehat{V} \otimes I_n) \;=\; (v_{m+1})^{(k-m-1)n} \det \widehat{M}\,. \tag{6.17}$$

Now taking determinants on both sides of (6.15) and using the factorization (6.16) gives

$$\begin{aligned}&\det \widehat{M} \cdot \det L(\lambda) \cdot \det T(\lambda) \cdot \det(R_m \otimes I_n) \cdot \det(R_{k-m} \otimes I_n)\\ &\quad = \det\left(P(r_1)P(r_2)\cdots P(r_{k-m-1})\right) \cdot \det(-v_{m+1}A_k)^m \cdot \det\left(v_{m+1}P(\lambda)\right) \cdot \det(\widehat{V}\otimes I_n).\end{aligned}$$

Canceling $\det \widehat{M}$ on both sides using (6.17), and using $\det T(\lambda) = +1$ together with the fact that $\det(R \otimes I)$ is its own inverse, we get

$$\begin{aligned}\det L(\lambda) = \det\left(P(r_1)P(r_2)\cdots P(r_{k-m-1})\right) &\cdot (-1)^{mn} \cdot (v_{m+1})^{kn} \cdot (\det A_k)^m\\ &\cdot \det P(\lambda) \cdot \det(R_m \otimes I_n) \cdot \det(R_{k-m} \otimes I_n).\end{aligned}$$

Finally, substituting $\det(R_m \otimes I_n) = (-1)^{n\cdot\left\lfloor \frac{m}{2} \right\rfloor}$ and $\det(R_{k-m} \otimes I_n) = (-1)^{n\cdot\left\lfloor \frac{k-m}{2} \right\rfloor}$ from (6.12) yields the desired formula (6.7). Note that this is consistent with formula (6.13) derived for the $m = 0$ case, as long as we interpret the term $(\det A_k)^m$ to be equal to $+1$ whenever $m = 0$, regardless of whether $\det A_k$ is zero or nonzero.

Part 3. Now that we know that (6.7) holds for any $v \in \mathbb{F}^k$ such that the corresponding $\mathsf{p}(x\,;v)$ has *distinct* finite roots, we can leverage this result to the general case by a continuity argument. For every *fixed* m and fixed polynomial $P(\lambda)$, the formula on the right-hand side of (6.7) is clearly a continuous function of the leading coefficient v_{m+1} and the roots $r_1, r_2, \ldots, r_{k-m-1}$ of $\mathsf{p}(x\,;v)$, and is defined for all lists in the set $\mathcal{D} = \big\{(v_{m+1}, r_1, r_2, \ldots, r_{k-m-1}) : v_{m+1} \neq 0\big\}$, regardless of whether the numbers $r_1, r_2, \ldots, r_{k-m-1}$ are distinct or not.

The left-hand side of (6.7) can also be viewed as a function defined and continuous for all lists in $\mathcal{D}$. To see this, first observe that the map

$$(v_{m+1}, r_1, r_2, \ldots, r_{k-m-1}) \;\mapsto\; (v_{m+1}, v_{m+2}, \ldots, v_k)$$

taking the leading coefficient and roots of the polynomial $\mathsf{p}(x\,;v)$ to the coefficients of the same polynomial $\mathsf{p}(x\,;v)$ is defined and continuous on $\mathcal{D}$, as well as being surjective. Next note that because of the isomorphism in Corollary 5.4, the unique pencil $L(\lambda) \in \mathbb{DL}(P)$ corresponding to $v = (0, 0, \ldots, 0, v_{m+1}, \ldots, v_k)^T$ can be expressed as a linear combination

$$L(\lambda) = v_{m+1}L_{m+1}(\lambda) + \cdots + v_kL_k(\lambda)$$

of the *fixed* pencils $L_i(\lambda)$ corresponding to $v = e_i$. Thus $\det L(\lambda)$ is a continuous function of $(v_{m+1}, v_{m+2}, \ldots, v_k)$, and hence also of $(v_{m+1}, r_1, r_2, \ldots, r_{k-m-1})$.

In summary, the two sides of (6.7) are continuous functions defined on the same domain $\mathcal{D}$ and have been shown to be equal on a *dense* subset

$$\big\{(v_{m+1}, r_1, r_2, \ldots, r_{k-m-1}) : v_{m+1} \neq 0 \;\text{ and }\; r_1, r_2, \ldots, r_{k-m-1} \;\text{ are distinct}\big\}$$

of $\mathcal{D}$. Therefore by continuity the two sides of (6.7) must be equal on all of $\mathcal{D}$. Since this argument holds for each m with $0 \leq m \leq k-1$, the desired result is established for all nonzero $v \in \mathbb{F}^k$. ∎

We now have all the ingredients needed to prove the main result of this section. Keep in mind our convention that the "roots of $\mathsf{p}(x\,;v)$" includes a root at ∞ whenever $v_1 = 0$.

THEOREM 6.7 (Eigenvalue Exclusion Theorem). *Suppose that $P(\lambda)$ is a regular matrix polynomial and $L(\lambda)$ is in $\mathbb{DL}(P)$ with nonzero ansatz vector v. Then $L(\lambda)$ is a linearization for $P(\lambda)$ if and only if no root of the* v*-polynomial $\mathsf{p}(x\,;v)$ is an eigenvalue of $P(\lambda)$. (Note that this statement includes ∞ as one of the possible roots of $\mathsf{p}(x\,;v)$ or possible eigenvalues of $P(\lambda)$.)*

Proof. By Theorem 4.3, $L(\lambda)$ is a linearization for $P(\lambda)$ if and only if $L(\lambda)$ is regular. However, from the determinant formula (6.7) it follows that $L(\lambda)$ is regular if and only if no root of $\mathsf{p}(x\,;v)$ is an eigenvalue of $P(\lambda)$. ∎

Using Theorem 6.7, we can now show that almost every pencil in $\mathbb{DL}(P)$ is a linearization for P. Although the same property was proved in Theorem 4.7 for pencils in $\mathbb{L}_1(P)$, the result for $\mathbb{DL}(P)$ is not a consequence of Theorem 4.7, since $\mathbb{DL}(P)$ is itself a closed, nowhere dense subset of measure zero in $\mathbb{L}_1(P)$. Neither can the proof of Theorem 4.7 be directly generalized in any simple way; hence the need for a different argument in the following result.

THEOREM 6.8 (Linearizations Are Generic in $\mathbb{DL}(P)$). *For any regular matrix polynomial $P(\lambda)$, pencils in $\mathbb{DL}(P)$ are linearizations of $P(\lambda)$ for almost all $v \in \mathbb{F}^k$. (Here "almost all" means for all but a closed, nowhere dense set of measure zero in $\mathbb{F}^k$.)*

Proof. Recall that the resultant [22] $\operatorname{res}(f,g)$ of two polynomials $f(x)$ and $g(x)$ is a polynomial in the coefficients of f and g with the property that $\operatorname{res}(f,g) = 0$ if and only if $f(x)$ and $g(x)$ have a common (finite) root. Now consider $\operatorname{res}(\mathsf{p}(x\,;v), \det P(x))$, which, because $P(\lambda)$ is fixed, can be viewed as a polynomial $r(v_1, v_2, \ldots, v_k)$ in the components of $v \in \mathbb{F}^k$. The zero set $\mathcal{Z}(r) = \left\{v \in \mathbb{F}^k : r(v_1, v_2, \ldots, v_k) = 0\right\}$, then, is exactly the set of $v \in \mathbb{F}^k$ for which some finite root of $\mathsf{p}(x\,;v)$ is an eigenvalue of $P(\lambda)$, together with the point $v = 0$. Recall that by our convention the v-polynomial $\mathsf{p}(x\,;v)$ has ∞ as a root exactly for $v \in \mathbb{F}^k$ lying in the hyperplane $v_1 = 0$. Thus by Theorem 6.7 the set of vectors $v \in \mathbb{F}^k$ for which the corresponding pencil $L(\lambda) \in \mathbb{DL}(P)$ is *not* a linearization of $P(\lambda)$ is either the proper algebraic set $\mathcal{Z}(r)$ or the union of two proper algebraic sets, $\mathcal{Z}(r)$ and the hyperplane $v_1 = 0$. However, the union of any finite number of proper algebraic sets is always a closed, nowhere dense set of measure zero in $\mathbb{F}^k$. ∎

How far can the eigenvalue exclusion theorem be extended from $\mathbb{DL}(P)$-pencils to other pencils in $\mathbb{L}_1(P)$? Let us say that a pencil $L \in \mathbb{L}_1(P)$ with right ansatz vector v has the *eigenvalue exclusion property* if the statement "no root of the v-polynomial $\mathsf{p}(x\,;v)$ is an eigenvalue of $P(\lambda)$" is equivalent to the linearization condition for L. That there are pencils in $\mathbb{L}_1(P)$ with the eigenvalue exclusion property that are not in $\mathbb{DL}(P)$ is shown by the pencil $L_1(\lambda)$ in Example 4.5. The following variation of Example 4.6, though, is easily shown not to have the eigenvalue exclusion property.

Example 6.9. For the general cubic polynomial $P(\lambda) = \lambda^3 A + \lambda^2 B + \lambda C + D$ consider the pencil

$$L(\lambda) = \lambda X + Y = \lambda \begin{bmatrix} A & 0 & 2C \\ -2A & -B-C & A-4C \\ 0 & A & 0 \end{bmatrix} + \begin{bmatrix} B & -C & D \\ C-B & 2C-A & -2D \\ -A & 0 & 0 \end{bmatrix}$$

that is in $\mathbb{L}_1(P)$ but not in $\mathbb{DL}(P)$. Since $X \boxplus Y = \begin{bmatrix} 1 & -2 & 0 \end{bmatrix}^T \otimes \begin{bmatrix} A & B & C & D \end{bmatrix}$, the right ansatz vector is $v = \begin{bmatrix} 1 & -2 & 0 \end{bmatrix}^T$ with v-polynomial $\mathsf{p}(x\,;v) = x^2 - 2x$ and

roots 0 and 2. On the other hand, applying the procedure described in section 4.1 gives

$$Z = \begin{bmatrix} B+C & -A \\ -A & 0 \end{bmatrix},$$

and hence the linearization condition $\det Z = \det(-A^2) \neq 0$, equivalently $\det A \neq 0$. Thus $L(\lambda)$ is a linearization for $P(\lambda)$ if and only if ∞ is not an eigenvalue of $P(\lambda)$. In this example, then, the roots of the v-polynomial do not correctly predict the linearization condition for L.

The first companion form of a polynomial P is another example where the eigenvalue exclusion property is easily seen not to hold. Characterizing the set of pencils in $\mathbb{L}_1(P)$ for which the eigenvalue exclusion property does hold is an open problem.

7. Concluding remarks. By generalizing the first and second companion form linearizations for a matrix polynomial $P(\lambda)$, we have introduced two large vector spaces of pencils, $\mathbb{L}_1(P)$ and $\mathbb{L}_2(P)$, which serve as sources of potential linearizations for $P(\lambda)$. The mild hypothesis that $P(\lambda)$ is regular makes almost every pencil in these spaces a linearization for $P(\lambda)$.

A number of properties enjoyed by the companion forms extend to the linearizations in $\mathbb{L}_1(P)$ and $\mathbb{L}_2(P)$: they are strong linearizations, are readily constructed from the coefficient matrices of $P(\lambda)$, and have eigenvectors that reveal those of $P(\lambda)$. Furthermore, a simple procedure can be used to test when a pencil in $\mathbb{L}_1(P)$ or $\mathbb{L}_2(P)$ is a linearization of $P(\lambda)$.

The intersection of $\mathbb{L}_1(P)$ and $\mathbb{L}_2(P)$, denoted by $\mathbb{DL}(P)$, is of particular significance. Pencils in $\mathbb{L}_1(P)$ reveal only right eigenvectors of $P(\lambda)$, while those in $\mathbb{L}_2(P)$ lead to left eigenvectors of $P(\lambda)$. Pencils in $\mathbb{DL}(P)$ therefore simultaneously reveal right as well as left eigenvectors of P. An isomorphism between $\mathbb{DL}(P)$ and $\mathbb{F}^k$ allows the association of a unique scalar polynomial of degree $k-1$ to each pencil in $\mathbb{DL}(P)$. Linearizations in $\mathbb{DL}(P)$ can then be characterized by an eigenvalue exclusion property—a pencil in this distinguished subspace is a linearization precisely when no root of its associated scalar polynomial is an eigenvalue of P.

As remarked earlier, the first and second companion form linearizations have a significant drawback—they usually do not reflect any structure that may be present in $P(\lambda)$. Different linearizations can also exhibit very different conditioning. By systematizing the construction of large classes of linearizations that generalize the companion forms, we have provided a rich arena in which linearizations with additional properties like structure preservation or improved conditioning can be found. This is the subject of the work in [7], [8], [12].

Appendix A. Proof of Proposition 5.2.

Proof. "⇒": Assume that (5.3) holds. First, we show by induction on k that the formulas (5.6)–(5.7) hold.

$\boxed{k=1}$: In this case, we have

$$X \boxplus\!\!\!\rightarrow Y = S = \begin{bmatrix} S_{11} & S_{12} \end{bmatrix}, \qquad X \boxplus\!\!\!\downarrow Y = T = \begin{bmatrix} T_{11} \\ T_{21} \end{bmatrix}$$

and hence $X = S_{11} = T_{11}$ and $Y = S_{12} = T_{21}$, which coincides with (5.6)–(5.7).

$\boxed{k-1 \Rightarrow k}$: By the definition of the column and row shifted sums, (5.3) implies

$$Y_{ik} = S_{i,k+1} \quad \text{and} \quad Y_{ki} = T_{k+1,i} \tag{A.1}$$

as well as $X_{ji} + Y_{j,i-1} = S_{ji}$ and $X_{ij} + Y_{i-1,j} = T_{ij}$ for $j = 1, \dots, k$ and $i = 2, \dots, k$, which together with (A.1) gives

$$X_{ki} = S_{ki} - T_{k+1,i-1} \quad \text{and} \quad X_{ik} = T_{ik} - S_{i-1,k+1} \tag{A.2}$$

for $i = 1, \dots, k$. (Remember that $S_{0,k+1} = 0 = T_{k+1,0}$ by convention.) In order to be able to use the induction hypothesis, let us partition X and Y as

$$X = \left[\begin{array}{ccc|c} & \widetilde{X} & & \begin{matrix} X_{1k} \\ \vdots \\ X_{k-1,k} \end{matrix} \\ \hline X_{k1} & \dots & X_{k,k-1} & X_{kk} \end{array}\right], \quad Y = \left[\begin{array}{ccc|c} & \widetilde{Y} & & \begin{matrix} Y_{1k} \\ \vdots \\ Y_{k-1,k} \end{matrix} \\ \hline Y_{k1} & \dots & Y_{k,k-1} & Y_{kk} \end{array}\right],$$

with $(n-1)k \times (n-1)k$ matrices $\widetilde{X}$ and $\widetilde{Y}$. Then we obtain

$$\widetilde{X} \boxplus\!\!\rightarrow \widetilde{Y} = \begin{bmatrix} S_{11} & \dots & S_{1,k-1} & S_{1k} - X_{1k} \\ S_{21} & \dots & S_{2,k-1} & S_{2k} - X_{2k} \\ \vdots & \ddots & \vdots & \vdots \\ S_{k-1,1} & \dots & S_{k-1,k-1} & S_{k-1,k} - X_{k-1,k} \end{bmatrix} \tag{A.3}$$

$$= \begin{bmatrix} S_{11} & \dots & S_{1,k-1} & S_{1k} - T_{1k} \\ S_{21} & \dots & S_{2,k-1} & S_{2k} - T_{2k} + S_{1,k+1} \\ \vdots & \ddots & \vdots & \vdots \\ S_{k-1,1} & \dots & S_{k-1,k-1} & S_{k-1,k} - T_{k-1,k} + S_{k-2,k+1} \end{bmatrix} =: \widetilde{S}. \tag{A.4}$$

Analogously,

(A.5)

$$\widetilde{X} \boxplus\!\!\downarrow \widetilde{Y} = \underbrace{\begin{bmatrix} T_{11} & T_{12} & \dots & T_{1,k-1} \\ \vdots & \vdots & \ddots & \vdots \\ T_{k-1,1} & T_{k-1,2} & \dots & T_{k-1,k-1} \\ T_{k1} - S_{k1} & T_{k2} - S_{k2} + T_{k+1,1} & \dots & T_{k,k-1} - S_{k,k-1} + T_{k+1,k-2} \end{bmatrix}}_{=: \widetilde{T}}.$$

Writing $\widetilde{S} = [\widetilde{S}_{ij}]$ and $\widetilde{T} = [\widetilde{T}_{ij}]$ with $n \times n$ blocks $\widetilde{S}_{ij}, \widetilde{T}_{ij}$ and using the induction hypothesis for $\widetilde{X} = [X_{ij}]$ and $\widetilde{Y} = [Y_{ij}]$, we then obtain for $i, j = 1, \dots, k-1$ and $j \geq i$ that

$$X_{ij} = \widetilde{T}_{ij} + \sum_{\mu=1}^{i-1} (\widetilde{T}_{\mu,j+i-\mu} - \widetilde{S}_{\mu,j+i-\mu}), \quad Y_{ij} = \sum_{\mu=1}^{i} (\widetilde{S}_{\mu,j+i+1-\mu} - \widetilde{T}_{\mu,j+i+1-\mu}), \tag{A.6}$$

$$X_{ji} = \widetilde{S}_{ji} + \sum_{\mu=1}^{i-1} (\widetilde{S}_{j+i-\mu,\mu} - \widetilde{T}_{j+i-\mu,\mu}), \quad Y_{ji} = \sum_{\mu=1}^{i} (\widetilde{T}_{j+i+1-\mu,\mu} - \widetilde{S}_{j+i+1-\mu,\mu}), \tag{A.7}$$

where $\widetilde{S}_{\nu\eta} = 0 = \widetilde{T}_{\eta\nu}$ whenever $(\nu, \eta) \notin \{1, \dots, k-1\} \times \{1, \dots, k\}$. We claim that together with (A.1) and (A.2), the formulas (A.6)–(A.7) coincide with the formulas (5.6)–(5.7). We show this in detail for the first formula in (5.6); for the other

formulas there is a similar proof that is omitted. If $j+i \le k$, then the block forms of $\widetilde{S}$ and $\widetilde{T}$ given in (A.4) and (A.5) immediately yield

$$X_{ij} = \widetilde{T}_{ij} + \sum_{\mu=1}^{i-1}(\widetilde{T}_{\mu,j+i-\mu} - \widetilde{S}_{\mu,j+i-\mu}) = T_{ij} + \sum_{\mu=1}^{i-1}(T_{\mu,j+i-\mu} - S_{\mu,j+i-\mu}).$$

If $j+i > k$ and $i,j < k$, then $j+i-m = k$ for some $m \ge 1$; using $S_{\nu\eta} = 0 = T_{\eta\nu}$ for $(\nu,\eta) \notin \{1,\dots,k\} \times \{1,\dots,k+1\}$, we obtain

$$\begin{aligned} X_{ij} &= \widetilde{T}_{ij} + \sum_{\mu=1}^{i-1}(\widetilde{T}_{\mu,j+i-\mu} - \widetilde{S}_{\mu,j+i-\mu}) \\ &= \widetilde{T}_{ij} + \sum_{\mu=m+1}^{i-1}(\widetilde{T}_{\mu,j+i-\mu} - \widetilde{S}_{\mu,j+i-\mu}) - \widetilde{S}_{mk} \\ &= T_{ij} + \sum_{\mu=m+1}^{i-1}(T_{\mu,j+i-\mu} - S_{\mu,j+i-\mu}) - S_{mk} + T_{mk} - S_{m-1,k+1} \\ &= T_{ij} + \sum_{\mu=1}^{i-1}(T_{\mu,j+i-\mu} - S_{\mu,j+i-\mu}). \end{aligned}$$

Finally, for $i = k$ or $j = k$ the statement follows immediately from (A.1) or (A.2). This concludes the inductive proof of the formulas (5.6)–(5.7). In particular, this implies that X and Y are uniquely determined by S and T. Note that X_{ii} and Y_{ii} now satisfy two distinct formulas for $i = 1,\dots,n$. Since both right-hand sides in the formulas (5.6)–(5.7) must be equal in this case, we directly obtain (5.4) and (5.5).

"$\Leftarrow$": We have to show the existence of block-matrices $X = [X_{ij}]$ and $Y = [Y_{ij}]$ such that $X \boxplus\!\!\!\rightarrow Y = S$ and $X \underset{\downarrow}{\boxplus} Y = T$. Define X_{ij} and Y_{ij} by the formulas (5.6)–(5.7). Because of (5.4) and (5.5), X and Y are well defined. We will now show in detail that $X \boxplus\!\!\!\rightarrow Y = S$. (The proof of $X \underset{\downarrow}{\boxplus} Y = T$ is similar and will be omitted.) Indeed, formulas (5.6)–(5.7) imply $X_{j1} = S_{j1}$ and $Y_{ik} = S_{ik}$ for $i,j = 1,\dots,k$. Moreover, we obtain for $i = 1,\dots,k$ and $j = 2,\dots,k$ that

$$\begin{aligned} X_{ij} + Y_{i,j-1} &= T_{ij} + \sum_{\mu=1}^{i-1}(T_{\mu,j+i-\mu} - S_{\mu,j+i-\mu}) + \sum_{\mu=1}^{i}(S_{\mu,j+i-\mu} - T_{\mu,j+i-\mu}) \\ &= T_{ij} + S_{ij} - T_{ij} = S_{ij} \end{aligned}$$

if $j-1 \ge i$, and that

$$X_{ij} + Y_{i,j-1} = S_{ij} + \sum_{\mu=1}^{j-1}(S_{j+i-\mu,\mu} - T_{j+i-\mu,\mu}) + \sum_{\mu=1}^{j-1}(T_{j+i-\mu,\mu} - S_{j+i-\mu,\mu}) = S_{ij}$$

if $j-1 < i$. This shows $X \boxplus\!\!\!\rightarrow Y = S$ and concludes the proof. ∎

Appendix B. Proof of Theorem 5.3.

Proof. Note that (5.8) is equivalent to $X \boxplus\!\!\!\rightarrow Y = S$ and $X \underset{\downarrow}{\boxplus} Y = T$, where $S = (S_{ij})$ and $T = (T_{ji})$ are block $k \times k$ matrices such that

$$S_{ij} = v_i A_{k+1-j}, \quad T_{ji} = w_i A_{k+1-j}, \qquad i = 1,\dots,k, \quad j = 1,\dots,k+1. \tag{B.1}$$

Then by Proposition 5.2, X and Y satisfying (5.8) exist if and only if

$$\text{(B.2)}\qquad \begin{aligned} &w_j A_{k+1-j} + \sum_{\mu=1}^{j-1} \left(w_{2j-\mu} A_{k+1-\mu} - v_\mu A_{k+1-2j+\mu} \right) \\ &\quad = v_j A_{k+1-j} + \sum_{\mu=1}^{j-1} \left(v_{2j-\mu} A_{k+1-\mu} - w_\mu A_{k+1-2j+\mu} \right) \end{aligned}$$

and

$$\text{(B.3)}\qquad \sum_{\mu=1}^{j} \left(v_\mu A_{k-2j+\mu} - w_{2j+1-\mu} A_{k+1-\mu} \right) = \sum_{\mu=1}^{j} \left(w_\mu A_{k-2j+\mu} - v_{2j+1-\mu} A_{k+1-\mu} \right)$$

for $j = 1, \ldots, k$. (Here $v_0 := w_0 := 0$, and for $\mu < 0$ or $\mu > k$, $v_\mu := w_\mu := 0$ and $A_\mu := 0 \in \mathbb{F}^{n \times n}$.) Hence, it is sufficient to prove the statement

$$v = w \quad \Longleftrightarrow \quad \text{(B.2) and (B.3) are satisfied.}$$

"$\Rightarrow$": If $v = w$, then (B.2) and (B.3) are obviously true.

"$\Leftarrow$": We show $v_m = w_m$ for $m = 1, \ldots, k$ by induction on m.

$\boxed{m = 1}$: (B.2) for $j = 1$ yields $v_1 A_k = w_1 A_k$. Since $A_k \neq 0$, this implies $v_1 = w_1$.

$\boxed{m = 2}$: (B.3) for $j = 1$ yields $v_1 A_{k-1} - w_2 A_k = w_1 A_{k-1} - v_2 A_k$. Since $v_1 = w_1$ and $A_k \neq 0$, this implies $v_2 = w_2$.

$\boxed{m - 1 \Rightarrow m}$: Assume first that m is odd, so that $m = 2j - 1$ for some $j \geq 2$. Since by the induction hypothesis we have $v_i = w_i$ for $i = 1, \ldots, 2j - 2$, we obtain from (B.2) that $w_{2j-1} A_k = v_{2j-1} A_k$. This implies $w_{2j-1} = v_{2j-1}$ because $A_k \neq 0$. Next assume that m is even, i.e., $m = 2j$ for some $j \geq 2$. Again, since $v_i = w_i$ for $i = 1, \ldots, 2j - 1$ by the induction hypothesis, we obtain from (B.3) that $w_{2j} A_k = v_{2j} A_k$. This implies $w_{2j} = v_{2j}$ because $A_k \neq 0$.

This concludes the induction. Hence we have $v = w$.

The uniqueness of X and Y and the formulas (5.9) and (5.10) follow directly from Proposition 5.2, the formulas (5.6) and (5.7), and (B.1). ∎

Acknowledgments. We thank the mathematics departments of the universities of Manchester and TU Berlin, and the Banff International Research Station for giving us the opportunity to carry out this joint research. We also thank Ralph Byers, Peter Benner, Nick Higham, Françoise Tisseur, and Hongguo Xu for enlightening discussions on the topic, and an anonymous referee for pointing out the factorization in (4.5).

REFERENCES

[1] E. N. Antoniou and S. Vologiannidis, *A new family of companion forms of polynomial matrices*, Electron. J. Linear Algebra, 11 (2004), pp. 78–87.

[2] E. Eich-Soellner and C. Führer, *Numerical Methods in Multibody Systems*, B. G. Teubner, Stuttgart, Germany, 1998.

[3] R. W. Freund, *Krylov-subspace methods for reduced-order modeling in circuit simulation*, J. Comput. Appl. Math., 123 (2000), pp. 395–421.

[4] I. Gohberg, M. A. Kaashoek, and P. Lancaster, *General theory of regular matrix polynomials and band Toeplitz operators*, Integral Equations Operator Theory, 11 (1988), pp. 776–882.

[5] I. Gohberg, P. Lancaster, and L. Rodman, *Matrix Polynomials*, Academic Press, New York, 1982.
[6] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[7] N. J. Higham, D. S. Mackey, N. Mackey, and F. Tisseur, *Symmetric Linearizations for Matrix Polynomials*, MIMS EPrint 2005.25, Manchester Institute for Mathematical Sciences, The University of Manchester, UK, 2005.
[8] 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.
[9] P. Lancaster, *Lambda-Matrices and Vibrating Systems*, Pergamon Press, Oxford, UK, 1966.
[10] P. Lancaster and P. Psarrakos, *A Note on Weak and Strong Linearizations of Regular Matrix Polynomials*, Numerical Analysis Report No. 470, Manchester Centre for Computational Mathematics, Manchester, UK, 2005.
[11] D. S. Mackey, *The characteristic polynomial of a partitioned matrix*, in Linear Algebra Gems, D. Carlson, C. R. Johnson, D. C. Lay, and A. D. Porter, eds., MAA Notes #59, Mathematical Association of America, Washington, DC, 2002, pp. 13–14.
[12] D. S. Mackey, N. Mackey, C. Mehl, and V. Mehrmann, *Palindromic Polynomial Eigenvalue Problems: Good Vibrations from Good Linearizations*, Technical Report 239, DFG Research Center Matheon, "Mathematics for key technologies" in Berlin, TU Berlin, Berlin, Germany, 2005; available online at http://www.matheon.de/.
[13] V. Mehrmann and C. Shi, *Analysis of Higher Order Linear Differential-Algebraic Systems*, Preprint 2004/17, Institut für Mathematik, TU Berlin, D-10623 Berlin, Germany, 2004; available online from http://www.math.tu-berlin.de/preprints/.
[14] 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.
[15] V. Mehrmann and D. Watkins, *Polynomial eigenvalue problems with Hamiltonian structure*, Electron. Trans. Numer. Anal., 13 (2002), pp. 106–118.
[16] W. Schiehlen, *Advanced Multibody System Dynamics*, Kluwer Academic Publishers, Stuttgart, Germany, 1993.
[17] F. Schmidt, T. Friese, L. Zschiedrich, and P. Deuflhard, *Adaptive Multigrid Methods for the Vectorial Maxwell Eigenvalue Problem for Optical Waveguide Design*, in Mathematics. Key Technology for the Future, W. Jäger and H.-J. Krebs, eds., Springer-Verlag, 2003, pp. 279–292.
[18] C. Shi, *Linear Differential-Algebraic Equations of Higher-Order and the Regularity or Singularity of Matrix Polynomials*, Ph.D. thesis, Institut für Mathematik, Technische Universität, Berlin, 2004.
[19] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.
[20] F. Tisseur, *Backward error and condition of polynomial eigenvalue problems*, Linear Algebra Appl., 309 (2000), pp. 339–361.
[21] F. Tisseur and K. Meerbergen, *The quadratic eigenvalue problem*, SIAM Rev., 43 (2001), pp. 235–286.
[22] B. L. van der Waerden, *Modern Algebra*, vol. 1, 2nd ed., Frederick Ungar Publishing, New York, 1953.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 4, pp. 1005–1028

THE CONDITIONING OF LINEARIZATIONS OF MATRIX POLYNOMIALS*

NICHOLAS J. HIGHAM†, D. STEVEN MACKEY†, AND FRANÇOISE TISSEUR†

Abstract. The standard way of solving the polynomial eigenvalue problem of degree m in $n \times n$ matrices is to "linearize" to a pencil in $mn \times mn$ matrices and solve the generalized eigenvalue problem. For a given polynomial, P, infinitely many linearizations exist and they can have widely varying eigenvalue condition numbers. We investigate the conditioning of linearizations from a vector space $\mathbb{DL}(P)$ of pencils recently identified and studied by Mackey, Mackey, Mehl, and Mehrmann. We look for the best conditioned linearization and compare the conditioning with that of the original polynomial. Two particular pencils are shown always to be almost optimal over linearizations in $\mathbb{DL}(P)$ for eigenvalues of modulus greater than or less than 1, respectively, provided that the problem is not too badly scaled and that the pencils are linearizations. Moreover, under this scaling assumption, these pencils are shown to be about as well conditioned as the original polynomial. For quadratic eigenvalue problems that are not too heavily damped, a simple scaling is shown to convert the problem to one that is well scaled. We also analyze the eigenvalue conditioning of the widely used first and second companion linearizations. The conditioning of the first companion linearization relative to that of P is shown to depend on the coefficient matrix norms, the eigenvalue, and the left eigenvectors of the linearization and of P. The companion form is found to be potentially much more ill conditioned than P, but if the 2-norms of the coefficient matrices are all approximately 1 then the companion form and P are guaranteed to have similar condition numbers. Analogous results hold for the second companion form. Our results are phrased in terms of both the standard relative condition number and the condition number of Dedieu and Tisseur [*Linear Algebra Appl.*, 358 (2003), pp. 71–94] for the problem in homogeneous form, this latter condition number having the advantage of applying to zero and infinite eigenvalues.

Key words. matrix polynomial, matrix pencil, linearization, companion form, condition number, homogeneous form, quadratic eigenvalue problem, vector space, scaling

AMS subject classifications. 65F15, 15A18

DOI. 10.1137/050628283

1. Introduction.

Consider the matrix polynomial of degree m

$$P(\lambda) = \sum_{i=0}^{m} \lambda^i A_i, \qquad A_i \in \mathbb{C}^{n\times n}, \quad A_m \neq 0. \tag{1.1}$$

We will assume throughout that P is regular, that is, $\det P(\lambda) \not\equiv 0$. The eigenproblem for P—the polynomial eigenvalue problem—is to find scalars λ and nonzero vectors x and y satisfying $P(\lambda)x = 0$ and $y^*P(\lambda) = 0$; x and y are right and left eigenvectors corresponding to the eigenvalue λ.

A standard way of solving the eigenproblem is to convert P into a linear polynomial

$$L(\lambda) = \lambda X + Y, \qquad X, Y \in \mathbb{C}^{mn\times mn}$$

*Received by the editors April 1, 2005; accepted for publication (in revised form) by J. Barlow October 11, 2005; published electronically December 18, 2006. This work was supported by Engineering and Physical Sciences Research Council grant GR/S31693.

http://www.siam.org/journals/simax/28-4/62828.html

†School of Mathematics, The University of Manchester, Sackville Street, Manchester, M60 1QD, UK (higham@ma.man.ac.uk, http://www.ma.man.ac.uk/~higham/, smackey@ma.man.ac.uk, ftisseur@ma.man.ac.uk, http://www.ma.man.ac.uk/~ftisseur/). The work of the first author was supported by a Royal Society-Wolfson Research Merit Award. The work of the third author was supported by Engineering and Physical Sciences Research Council grant GR/R45079.

with the same spectrum as P and solve the resulting generalized eigenproblem $L(\lambda)x = 0$, which is usually done by the QZ algorithm for small to medium size problems or a Krylov method for large sparse problems. The aim of this work is to provide guidance on how to choose from among the infinitely many possible pencils $L(\lambda)$.

We are interested in pencils $L(\lambda)$ that are linearizations of $P(\lambda)$ in the following sense: they satisfy

$$E(\lambda)L(\lambda)F(\lambda) = \begin{bmatrix} P(\lambda) & 0 \\ 0 & I_{(m-1)n} \end{bmatrix} \tag{1.2}$$

for some unimodular $E(\lambda)$ and $F(\lambda)$ (that is, $\det(E(\lambda))$ is a nonzero constant, independent of λ, and likewise for F) [6, sect. 7.2]. This definition implies that $\gamma \det(L(\lambda)) = \det(P(\lambda))$ for some nonzero constant γ, so that L and P have the same spectrum. As an example, the pencil

$$C_1(\lambda) = \lambda \begin{bmatrix} A & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} B & C \\ -I & 0 \end{bmatrix} \tag{1.3}$$

can be shown to be a linearization for the quadratic $Q(\lambda) = \lambda^2 A + \lambda B + C$; it is known as the first companion form linearization (see section 7).

Two important sets of potential linearizations are identified and studied by Mackey, Mackey, Mehl, and Mehrmann [13]. With the notation

$$\Lambda = [\lambda^{m-1}, \lambda^{m-2}, \dots, 1]^T, \tag{1.4}$$

the sets are

$$\mathbb{L}_1(P) = \left\{ L(\lambda) : L(\lambda)(\Lambda \otimes I_n) = v \otimes P(\lambda),\ v \in \mathbb{C}^m \right\}, \tag{1.5}$$

$$\mathbb{L}_2(P) = \left\{ L(\lambda) : (\Lambda^T \otimes I_n)L(\lambda) = w^T \otimes P(\lambda),\ w \in \mathbb{C}^m \right\}. \tag{1.6}$$

It is easy to check that $C_1(\lambda)$ in (1.3) belongs to $\mathbb{L}_1(Q)$ (with $v = e_1$)[1]; so the pencils in $\mathbb{L}_1$ can be thought of as generalizations of the first companion form. It is proved in [13, Prop. 3.2, Prop. 3.12, Thm. 4.7] that $\mathbb{L}_1(P)$ and $\mathbb{L}_2(P)$ are vector spaces and that almost all pencils in these spaces are linearizations of P.

One of the underlying reasons for the interest in $\mathbb{L}_1$ and $\mathbb{L}_2$ is that eigenvectors of P can be directly recovered from eigenvectors of linearizations in $\mathbb{L}_1$ and $\mathbb{L}_2$. Specifically, if L is any pencil in $\mathbb{L}_1(P)$ with nonzero vector v, then x is a right eigenvector of P with eigenvalue λ if and only if $\Lambda \otimes x$ (if λ is finite) or $e_1 \otimes x$ (if $\lambda = \infty$) is a right eigenvector for L with eigenvalue λ. Moreover, if this $L \in \mathbb{L}_1(P)$ is a linearization for P, then *every* right eigenvector of L has one of these two Kronecker product forms; hence some right eigenvector of P can be recovered from every right eigenvector of L. Similarly, if L is any pencil in $\mathbb{L}_2(P)$ with nonzero vector w, then y is a left eigenvector for P with eigenvalue λ if and only if $\overline{\Lambda} \otimes y$ (if λ is finite) or $e_1 \otimes y$ (if $\lambda = \infty$) is a left eigenvector for L with eigenvalue λ. Again, if this $L \in \mathbb{L}_2(P)$ is a linearization for P, then *every* left eigenvector of L is of the form $\overline{\Lambda} \otimes y$ or $e_1 \otimes y$, and so every left eigenvector of L produces a left eigenvector for P. Some insight can be gained from the proof of the first part of these results. For any $L \in \mathbb{L}_1(P)$, postmultiplying the equation in (1.5) defining $\mathbb{L}_1$ by $1 \otimes x$ gives

$$L(\lambda)(\Lambda \otimes x) = v \otimes P(\lambda)x.$$

[1] e_i denotes the ith column of the identity matrix I_m.

Hence for finite λ and $v \neq 0$, (x, λ) is an eigenpair for P if and only if $(\Lambda \otimes x, \lambda)$ is an eigenpair for L. The complete proofs of these results can be found in [13, Thm. 3.8, Thm. 3.14, Thm. 4.4].

It is natural to concentrate attention on the pencils that lie in

$$\mathbb{DL}(P) = \mathbb{L}_1(P) \cap \mathbb{L}_2(P), \tag{1.7}$$

because there is a simultaneous correspondence between left *and* right eigenvectors of P and of pencils in $\mathbb{DL}(P)$. It is shown in [13, Thm. 5.3] and in [7, Thm. 3.4] that $L \in \mathbb{DL}(P)$ only if L satisfies the conditions in (1.5) and (1.6) with $w = v$, and that, for any $v \in \mathbb{C}^m$, there are uniquely determined X and Y such that $L(\lambda) = \lambda X + Y$ is in $\mathbb{DL}(P)$. Thus $\mathbb{DL}(P)$ is always an m-dimensional space of pencils associated with P. A basis for $\mathbb{DL}(P)$ corresponding to the standard basis $v = e_i$, $i = 1\colon m$, for $\mathbb{C}^m$ is derived in [7, Sect. 3.3]. In this work we focus on linearizations in $\mathbb{DL}(P)$.

Just as for $\mathbb{L}_1$ and $\mathbb{L}_2$, almost all pencils in $\mathbb{DL}(P)$ are linearizations [13, Thm. 6.8]. In fact, there is a beautiful characterization of the subset of pencils $L \in \mathbb{DL}(P)$ that are linearizations [13, Thm. 6.7]: they are those for which no eigenvalue of P is a root of the polynomial $\mathsf{p}(\lambda; v) := v^T \Lambda = \sum_{i=1}^m v_i \lambda^{m-i}$, where when $v_1 = 0$ we define ∞ to be a root of $\mathsf{p}(\lambda; v)$. Throughout this work we assume that the pencils $L \in \mathbb{DL}(P)$ under consideration are linearizations.

The polynomials with $m > 1$ of greatest practical importance are the quadratics. For later use we note that for $m = 2$ and $Q(\lambda) = \lambda^2 A + \lambda B + C$,

$$\mathbb{DL}(Q) = \left\{ L(\lambda) = \lambda \begin{bmatrix} v_1 A & v_2 A \\ v_2 A & v_2 B - v_1 C \end{bmatrix} + \begin{bmatrix} v_1 B - v_2 A & v_1 C \\ v_1 C & v_2 C \end{bmatrix} : \; v \in \mathbb{C}^2 \right\}, \tag{1.8}$$

which can be deduced directly from the definition of $\mathbb{DL}$ in (1.7).

We now summarize the organization of the paper. In section 2 we define and describe properties of a relative condition number for a simple eigenvalue of P and a condition number of Dedieu and Tisseur for the problem in homogeneous form. Although there is no direct connection between the two condition numbers, both are of interest, and all results in the paper are stated for both. In section 3 we obtain for a linearization in $\mathbb{DL}(P)$ expressions for the condition numbers that separate the dependence on P from that of the vector v that defines the linearization. These expressions are then used in section 4 to approximately minimize the condition numbers over all v. The pencils with $v = e_1$ and $v = e_m$, which are linearizations when A_0 and A_m, respectively, are nonsingular, are shown always to be almost optimal within $\mathbb{DL}(P)$ for eigenvalues of modulus greater than or less than 1, respectively, provided that the measure $\rho = \max_i \|A_i\|_2 / \min(\|A_0\|_2, \|A_m\|_2)$ of the scaling of the problem is of order 1. This result generalizes and strengthens earlier results of Tisseur [16] for the quadratic case. Under the same scaling assumption these two linearizations are shown to be about as well conditioned as the original polynomial. How to extend the results to sets of eigenvalues, and the situation where we know only a region containing the eigenvalues, is discussed in section 4.1.

In section 5 we turn to quadratic polynomials and show that a simple scaling converts the problem to one that is well scaled, provided the quadratic is not too heavily damped. In section 6 we prove the equality of the condition number of an eigenvalue of a linearization in $\mathbb{DL}(P)$ with the condition number of the corresponding reciprocal eigenvalue of a linearization of the "reversal" of the polynomial. In section 7 we show that the ratio of the condition number of the first companion linearization to that of P at a given λ depends on the product of a rational function of $|\lambda|$ and the

norms $\|A_i\|$ with the ratio of the norms of the left eigenvectors of the pencil and P. This result, and its analogue for the second companion form, reveals and gives insight into potential instability of the companion forms.

Finally, the numerical experiments in section 8 show the ability of our analysis to predict well the accuracy of eigenvalues computed via different linearizations.

2. Eigenvalue condition number. Let λ be a simple, finite, nonzero eigenvalue of P in (1.1) with corresponding right eigenvector x and left eigenvector y. A normwise condition number of λ can be defined by

$$\kappa_P(\lambda) = \lim_{\epsilon\to 0} \sup \left\{ \frac{|\Delta\lambda|}{\epsilon|\lambda|} : \big(P(\lambda+\Delta\lambda) + \Delta P(\lambda+\Delta\lambda)\big)(x+\Delta x) = 0, \right.$$

$$\left. \|\Delta A_i\|_2 \le \epsilon\omega_i,\ i = 0\colon m \right\}, \tag{2.1}$$

where $\Delta P(\lambda) = \sum_{i=0}^m \lambda^i \Delta A_i$. The ω_i are nonnegative weights that allow flexibility in how the perturbations are measured; in particular, ΔA_i can be forced to zero by setting $\omega_i = 0$. An explicit formula for this condition number is given in the following result.

THEOREM 2.1 (Tisseur [16, Thm. 5]). *The normwise condition number $\kappa_P(\lambda)$ is given by*

$$\kappa_P(\lambda) = \frac{\big(\sum_{i=0}^m |\lambda|^i \omega_i\big)\|y\|_2\|x\|_2}{|\lambda|\,|y^* P'(\lambda) x|}. \tag{2.2}$$

The condition number $\kappa_P(\lambda)$ has the disadvantage that it is not defined for zero or infinite eigenvalues. In order to give a unified treatment for all λ, we rewrite the polynomial in the homogeneous form

$$P(\alpha,\beta) = \sum_{i=0}^m \alpha^i \beta^{m-i} A_i$$

and consider eigenvalues as pairs $(\alpha,\beta) \ne (0,0)$ that are solutions of the scalar equation $\det P(\alpha,\beta) = 0$; thus $\lambda \equiv \alpha/\beta$. More precisely, since $P(\alpha,\beta)$ is homogeneous in α and β, we define an eigenvalue as any line through the origin in $\mathbb{C}^2$ of solutions of $\det P(\alpha,\beta) = 0$. Let $T_{(\alpha,\beta)}\mathbb{P}_1$ denote the tangent space at (α,β) to $\mathbb{P}_1$, the projective space of lines through the origin in $\mathbb{C}^2$. Dedieu and Tisseur [3] define a condition operator $K(\alpha,\beta) : (\mathbb{C}^{n\times n})^{m+1} \to T_{(\alpha,\beta)}\mathbb{P}_1$ for the eigenvalue (α,β) as the differential of the map from the $(m+1)$-tuple $(A_0,\dots,A_m)$ to (α,β) in projective space. The significance of the condition operator is shown by the following result, which is an extension of a result of Dedieu [2, Thm. 6.1]. Here and below, we sometimes write a representative of an eigenvalue (α,β) as a row vector $[\alpha,\beta] \in \mathbb{C}^{1\times 2}$.

THEOREM 2.2. *Let (α,β) be a simple eigenvalue of $P(\alpha,\beta)$ with representative $[\alpha,\beta]$ normalized so that $\|[\alpha,\beta]\|_2 = 1$. For sufficiently small $(m+1)$-tuples*

$$\Delta A \equiv (\Delta A_0,\dots,\Delta A_m),$$

the perturbed polynomial $\widetilde{P}(\alpha,\beta) = \sum_{i=0}^m \alpha^i\beta^{m-i}(A_i + \Delta A_i)$ has a simple eigenvalue $(\widetilde{\alpha},\widetilde{\beta})$ for which, with the normalization $[\alpha,\beta][\widetilde{\alpha},\widetilde{\beta}]^* = 1$,

$$[\widetilde{\alpha},\widetilde{\beta}] = [\alpha,\beta] + K(\alpha,\beta)\Delta A + o(\|\Delta A\|).$$

A condition number $\kappa_P(\alpha,\beta)$ can be defined as a norm of the condition operator:

$$\kappa_P(\alpha,\beta) = \max_{\|\Delta A\|\le 1} \frac{\|K(\alpha,\beta)\Delta A\|_2}{\|[\alpha,\beta]\|_2},$$

where the norm on ΔA is arbitrary. Note that this condition number is well defined, since the right-hand side is independent of the choice of representative of the eigenvalue (α,β). Let $\theta\big((\mu,\nu),(\widetilde{\mu},\widetilde{\nu})\big)$ be the angle between the two lines (μ,ν) and $(\widetilde{\mu},\widetilde{\nu})$. Then for θ small enough,

$$|\theta\big((\mu,\nu),(\widetilde{\mu},\widetilde{\nu})\big)| \le \big|\tan\big(\theta\big((\mu,\nu),(\widetilde{\mu},\widetilde{\nu})\big)\big)\big| = \left\| [\widetilde{\mu},\widetilde{\nu}] \frac{\|[\mu,\nu]\|_2}{[\widetilde{\mu},\widetilde{\nu}][\mu,\nu]^*} - \frac{[\mu,\nu]}{\|[\mu,\nu]\|_2} \right\|_2.$$

Inserting the particular representatives $[\alpha,\beta]$ and $[\widetilde{\alpha},\widetilde{\beta}]$ of the original and perturbed eigenvalues, normalized as in Theorem 2.2, gives

$$\big|\theta\big((\alpha,\beta),(\widetilde{\alpha},\widetilde{\beta})\big)\big| \le \|[\alpha,\beta]-[\widetilde{\alpha},\widetilde{\beta}]\|_2 = \|K(\alpha,\beta)\Delta A\|_2 + o(\|\Delta A\|).$$

Hence, the angle between the original and perturbed eigenvalues satisfies

$$\big|\theta\big((\alpha,\beta),(\widetilde{\alpha},\widetilde{\beta})\big)\big| \le \kappa_P(\alpha,\beta)\|\Delta A\| + o(\|\Delta A\|). \tag{2.3}$$

By taking the sine of both sides we obtain a perturbation bound in terms of $\sin|\theta|$, which is the chordal distance between (α,β) and $(\widetilde{\alpha},\widetilde{\beta})$ as used by Stewart and Sun [15, Chap. 6]. Of course, $\sin|\theta| \le |\theta|$ and asymptotically these two measures of distance are equal.

We will take for the norm on $(\mathbb{C}^{n\times n})^{m+1}$ the ω-weighted Frobenius norm

$$\|A\| = \|(A_0,\dots,A_m)\| = \|[\omega_0^{-1}A_0,\dots,\omega_m^{-1}A_m]\|_F, \tag{2.4}$$

where the ω_i are nonnegative weights that are analogous to those in (2.1). Define the operators $\mathcal{D}_\alpha \equiv \frac{\partial}{\partial\alpha}$ and $\mathcal{D}_\beta \equiv \frac{\partial}{\partial\beta}$. The following result is a trivial extension of a result of Dedieu and Tisseur [3, Thm. 4.2] that treats the unweighted Frobenius norm.

THEOREM 2.3. *The normwise condition number $\kappa_P(\alpha,\beta)$ of a simple eigenvalue (α,β) is given by*

$$\kappa_P(\alpha,\beta) = \left(\sum_{i=0}^{m} |\alpha|^{2i}|\beta|^{2(m-i)}\omega_i^2\right)^{1/2} \frac{\|y\|_2\|x\|_2}{\big|y^*(\bar{\beta}\mathcal{D}_\alpha P - \bar{\alpha}\mathcal{D}_\beta P)|_{(\alpha,\beta)}x\big|}. \tag{2.5}$$

As a check, we note that the expression (2.5) is independent of the choice of representative of (α,β) and of the scaling of x and y. Note also that for a simple eigenvalue the denominator terms $y^*P'(\lambda)x$ in (2.2) and $y^*(\bar{\beta}\mathcal{D}_\alpha - \bar{\alpha}\mathcal{D}_\beta)P|_{(\alpha,\beta)}x$ in (2.5) are both nonzero, as shown in [1, Thm. 3.2] for the former and [3, Thm. 3.3(iii)] for the latter.

To summarize, $\kappa_P(\lambda)$ and $\kappa_P(\alpha,\beta)$ are two different measures of the sensitivity of a simple eigenvalue. The advantage of $\kappa_P(\lambda)$ is that it is an immediate generalization of the well-known Wilkinson condition number for the standard eigenproblem [18, p. 69] and it measures the relative change in an eigenvalue, which is a concept readily understood by users of numerical methods. In favor of $\kappa_P(\alpha,\beta)$ is that it elegantly treats all eigenvalues, including those at zero and infinity; moreover, it provides the bound (2.3) for the angular error, which is an alternative to the relative error bound

that $\kappa_P(\lambda)$ provides. Both condition numbers are therefore of interest and we will treat both in the next section.

We note that in MATLAB 7.0 (R14) the function `polyeig` that solves the polynomial eigenvalue problem returns the condition number $\kappa_P(\alpha,\beta)$ as an optional output argument.

3. Eigenvalue conditioning of linearizations. We now focus on the condition numbers $\kappa_L(\lambda)$ and $\kappa_L(\alpha,\beta)$ of a simple eigenvalue of a linearization $L(\lambda) = \lambda X + Y \in \mathbb{DL}(P)$. Our aim is to obtain expressions for these condition numbers with two properties: they separate the dependence on P from that of the vector v and they have minimal explicit dependence on X and Y. In the next section we will consider how to minimize these expressions over all v. Note the distinction between the condition numbers κ_L of the pencil and κ_P of the original polynomial. Note also that a simple eigenvalue of L is necessarily a simple eigenvalue of P, and vice versa, in view of (1.2).

We first carry out the analysis for $\kappa_L(\alpha,\beta)$. Let x and y denote right and left eigenvectors of P, and z and w denote right and left eigenvectors of L, all corresponding to the eigenvalue (α,β). Recalling that $\lambda = \alpha/\beta$, define

$$\begin{aligned} L(\alpha,\beta) &= \alpha X + \beta Y = \beta L(\lambda), \\ \Lambda_{\alpha,\beta} &= [\alpha^{m-1}, \alpha^{m-2}\beta, \ldots, \beta^{m-1}]^T = \beta^{m-1}\Lambda. \end{aligned}$$

In view of the relations in section 1 we can take

$$w = \overline{\Lambda}_{\alpha,\beta} \otimes y, \qquad z = \Lambda_{\alpha,\beta} \otimes x. \tag{3.1}$$

(These expressions are valid for both finite and infinite eigenvalues.)

The condition number that we wish to evaluate is obtained by applying Theorem 2.3 to L:

$$\kappa_L(\alpha,\beta) = \sqrt{|\alpha|^2\omega_X^2 + |\beta|^2\omega_Y^2}\ \frac{\|w\|_2\|z\|_2}{\left|w^*(\bar\beta\mathcal{D}_\alpha L - \bar\alpha\mathcal{D}_\beta L)|_{(\alpha,\beta)}z\right|}, \tag{3.2}$$

where an obvious notation has been used for the weights in (2.4).

We can rewrite the condition in (1.5) that characterizes a member of $\mathbb{L}_1$ as

$$L(\alpha,\beta)(\Lambda_{\alpha,\beta} \otimes I_n) = v \otimes P(\alpha,\beta), \tag{3.3}$$

where for the moment α and β denote variables. Differentiating with respect to α gives

$$\mathcal{D}_\alpha L(\alpha,\beta)(\Lambda_{\alpha,\beta} \otimes I_n) + L(\alpha,\beta)(\mathcal{D}_\alpha\Lambda_{\alpha,\beta} \otimes I_n) = v \otimes \mathcal{D}_\alpha P(\alpha,\beta). \tag{3.4}$$

Now evaluate this equation at an eigenvalue[2] (α,β). Multiplying on the left by w^* and on the right by $1 \otimes x$, and using (3.1), we obtain

$$\begin{aligned} w^*(\mathcal{D}_\alpha L)|_{(\alpha,\beta)}z &= \Lambda_{\alpha,\beta}^T v \otimes y^*(\mathcal{D}_\alpha P)|_{(\alpha,\beta)}x \\ &= \Lambda_{\alpha,\beta}^T v \cdot y^*(\mathcal{D}_\alpha P)|_{(\alpha,\beta)}x. \end{aligned} \tag{3.5}$$

Exactly the same argument leads to

$$w^*(\mathcal{D}_\beta L)|_{(\alpha,\beta)}z = \Lambda_{\alpha,\beta}^T v \cdot y^*(\mathcal{D}_\beta P)|_{(\alpha,\beta)}x. \tag{3.6}$$

[2]Strictly speaking, here and later we are evaluating at a representative of an eigenvalue. All the condition number formulae are independent of the choice of representative.

Hence, from (3.5) and (3.6),

$$w^*(\bar\beta\mathcal{D}_\alpha L - \bar\alpha\mathcal{D}_\beta L)|_{(\alpha,\beta)}z = \Lambda_{\alpha,\beta}^T v \cdot y^*(\bar\beta\mathcal{D}_\alpha P - \bar\alpha\mathcal{D}_\beta P)|_{(\alpha,\beta)}x.$$

The first factor on the right can be viewed as a homogeneous scalar polynomial in α and β, so we introduce the notation

$$\mathsf{p}(\alpha,\beta;v) := v^T\Lambda_{\alpha,\beta} = \sum_{i=1}^m v_i\alpha^{m-i}\beta^{i-1} = \Lambda_{\alpha,\beta}^T v. \tag{3.7}$$

Noting, from (3.1), that $\|w\|_2 = \|\Lambda_{\alpha,\beta}\|_2\|y\|_2$ and $\|z\|_2 = \|\Lambda_{\alpha,\beta}\|_2\|x\|_2$, we obtain an alternative form of (3.2) that clearly separates the dependence on P from that on the vector v that defines the linearization. Now we write $\kappa_L(\alpha,\beta;v)$ to indicate the dependence of κ_L on the vector $v\in\mathbb{C}^m$ that defines the linearization.

THEOREM 3.1. *Let (α,β) be a simple eigenvalue of P with right and left eigenvectors x and y, respectively. Then, for any pencil $L(\alpha,\beta) = \alpha X + \beta Y \in \mathbb{DL}(P)$ that is a linearization of P,*

$$\kappa_L(\alpha,\beta;v) = \frac{\sqrt{|\alpha|^2\omega_X^2 + |\beta|^2\omega_Y^2}}{|\mathsf{p}(\alpha,\beta;v)|} \cdot \frac{\|\Lambda_{\alpha,\beta}\|_2^2\,\|y\|_2\|x\|_2}{\left|y^*(\bar\beta\mathcal{D}_\alpha P - \bar\alpha\mathcal{D}_\beta P)|_{(\alpha,\beta)}x\right|}, \tag{3.8}$$

where v is the vector in (3.3).

Now we give a similar analysis for the condition number $\kappa_L(\lambda)$ of a finite, nonzero λ. In view of (2.2), our aim is to obtain an expression for $|w^*L'(\lambda)z|$. Since $L\in\mathbb{L}_1$,

$$L(\lambda)(\Lambda\otimes I_n) = v\otimes P(\lambda). \tag{3.9}$$

Differentiating (3.9) with respect to λ gives

$$L'(\lambda)(\Lambda\otimes I_n) + L(\lambda)(\Lambda'\otimes I_n) = v\otimes P'(\lambda). \tag{3.10}$$

Evaluating at an eigenvalue λ, premultiplying by $w^* = \Lambda^T\otimes y^*$, postmultiplying by $1\otimes x$, and using (3.1), gives

$$w^*L'(\lambda)z = \Lambda^T v\otimes y^*P'(\lambda)x = \Lambda^T v\cdot y^*P'(\lambda)x.$$

Analogously to (3.7), we write

$$\mathsf{p}(\lambda;v) := v^T\Lambda = \sum_{i=1}^m v_i\lambda^{m-i} = \Lambda^T v$$

for the polynomial defined by v with variable λ.

THEOREM 3.2. *Let λ be a simple, finite, nonzero eigenvalue of P with right and left eigenvectors x and y, respectively. Then, for any pencil $L(\lambda) = \lambda X + Y \in \mathbb{DL}(P)$ that is a linearization of P,*

$$\kappa_L(\lambda;v) = \frac{(|\lambda|\omega_X + \omega_Y)}{|\mathsf{p}(\lambda;v)|} \cdot \frac{\|\Lambda\|_2^2\|y\|_2\|x\|_2}{|\lambda|\;|y^*P'(\lambda)x|}, \tag{3.11}$$

where v is the vector in (3.9).

The expression (3.8) shows that $\kappa_L(\alpha,\beta)$ is finite if and only if (α,β) is not a zero of $\mathsf{p}(\alpha,\beta;v)$, and (3.11) gives essentially the same information for $\lambda\neq 0,\infty$. This is consistent with the theory in [13] which shows, as noted in section 1, that $L(\lambda)$ is a linearization for $P(\lambda)$ if and only if no eigenvalue of P (including ∞) is a root of $\mathsf{p}(\lambda;v)$.

4. Minimizing the condition numbers $\kappa_L(\alpha,\beta)$ and $\kappa_L(\lambda)$. A pencil $L(\lambda) \in \mathbb{DL}(P)$ is uniquely defined by the vector v in (1.5). Our aim is to minimize the condition numbers $\kappa_L(\lambda)$ and $\kappa_L(\alpha,\beta)$ over all $v \in \mathbb{C}^m$, thereby identifying a best conditioned linearization for a particular eigenvalue.

A technical subtlety is that the minimum of $\kappa_L(\alpha,\beta)$ over v could potentially occur at a v for which L is not a linearization, since we are minimizing for a particular eigenvalue, whereas the property of being a linearization is a property concerning all the eigenvalues. In this case formulas (3.8) and (3.11) are not valid. However, such "bad" v form a closed, nowhere dense set of measure zero [13, Thm. 6.8] and an arbitrarily small perturbation to v can make L a linearization.

Expressions (3.8) and (3.11) have similar forms, with dependence on v confined to the $\mathsf{p}(\cdot)$ terms in the denominator and the ω terms in the numerator. For most of this section we work with the condition number (3.8) for the pencil in homogeneous form; we return to $\kappa_L(\lambda)$ at the end of the section.

For the weights we will take the natural choice

$$\omega_X = \|X\|_2, \qquad \omega_Y = \|Y\|_2. \tag{4.1}$$

Since the entries of X and Y are linear combinations of the entries of v [13, Thm. 5.3], this choice makes the condition numbers independent of the scaling of v.

We consider first the v-dependence of $\|X\|_2$ and $\|Y\|_2$.

LEMMA 4.1. *For $L(\lambda) = \lambda X + Y \in \mathbb{DL}(P)$ defined by $v \in \mathbb{C}^m$ we have*

$$\|v\|_2\|A_m\|_2 \le \|X\|_2 \le m\, r^{1/2} \max_i \|A_i\|_2 \|v\|_2, \tag{4.2}$$

$$\|v\|_2\|A_0\|_2 \le \|Y\|_2 \le m\, r^{1/2} \max_i \|A_i\|_2 \|v\|_2, \tag{4.3}$$

where r is the number of nonzero entries in v.

Proof. Partition X and Y as block $m \times m$ matrices with $n \times n$ blocks. From [7, Sect. 3.3] or [13, Thm. 3.5] we know that the first block column of X is $v \otimes A_m$ and the last block column of Y is $v \otimes A_0$. The lower bounds are therefore immediate. From [7, Sect. 3.3] or [13, Thm. 5.3] it can be seen that each block of X has the form

$$X_{ij} = \sum_{k=1}^{m} s_k v_k A_{\ell_k}, \tag{4.4}$$

where $s_k \in \{-1, 0, 1\}$ and the indices ℓ_k are distinct. Hence

$$\|X_{ij}\|_2 \le \max_k \|A_k\|_2 \sum_{k=1}^{m} |v_k| = \max_k \|A_k\|_2 \|v\|_1 \le r^{1/2} \max_k \|A_k\|_2 \|v\|_2.$$

The upper bound on $\|X\|_2$ follows on using

$$\|X\|_2 \le m \max_{i,j} \|X_{ij}\|_2, \tag{4.5}$$

which holds for any block $m \times m$ matrix. An identical argument gives the upper bound for $\|Y\|_2$. ∎

Hence, provided the $\|A_i\|_2$ values vary little in magnitude with i, the numerator of (3.8) varies little in magnitude with v if $\|v\|_2$ is fixed. Under this proviso, we will approximately minimize the condition number $\kappa_L(\alpha,\beta)$ if we maximize the $\mathsf{p}(\alpha,\beta;v)$

term. We therefore restrict our attention to the denominator of the expression (3.8) for κ_L and maximize $|\mathsf{p}(\alpha,\beta;v)| = |\Lambda_{\alpha,\beta}^T v|$ subject to $\|v\|_2 = 1$, for a given eigenvalue (α,β). By the Cauchy–Schwarz inequality, the maximizing v, and the corresponding value of the polynomial, are

$$v_* = \frac{\overline{\Lambda}_{\alpha,\beta}}{\|\Lambda_{\alpha,\beta}\|_2}, \qquad |\mathsf{p}(\alpha,\beta;v_*)| = \|\Lambda_{\alpha,\beta}\|_2. \tag{4.6}$$

Two special cases that play an important role in the rest of this paper are worth noting:

$$\begin{aligned} (\alpha,\beta) = (1,0), \quad \lambda = \infty \quad &\Rightarrow \quad v_* = e_1, \\ (\alpha,\beta) = (0,1), \quad \lambda = 0 \quad &\Rightarrow \quad v_* = e_m. \end{aligned}$$

The next theorem compares the condition numbers for $v = e_1$ and $v = e_m$ with the optimal condition number. Define

$$\rho = \frac{\max_i \|A_i\|_2}{\min(\|A_0\|_2, \|A_m\|_2)} \geq 1. \tag{4.7}$$

When we write $\inf_v \kappa_L(\alpha,\beta;v)$ the infimum is understood to be taken over v for which L is a linearization.

THEOREM 4.2. *Let (α,β) be a simple eigenvalue of P and consider pencils $L \in \mathbb{DL}(P)$. Take the weights* (4.1) *for κ_L. Then*

$$\kappa_L(\alpha,\beta;e_1) \leq \rho m^{3/2} \inf_v \kappa_L(\alpha,\beta;v) \quad \textit{if } A_0 \textit{ is nonsingular and } |\alpha| \geq |\beta|, \tag{4.8}$$

$$\kappa_L(\alpha,\beta;e_m) \leq \rho m^{3/2} \inf_v \kappa_L(\alpha,\beta;v) \quad \textit{if } A_m \textit{ is nonsingular and } |\alpha| \leq |\beta|. \tag{4.9}$$

Proof. Note first that the conditions that A_0 and A_m are nonsingular ensure that 0 and ∞, respectively, are not eigenvalues of P, and hence that $v = e_1$ and $v = e_m$, respectively, yield linearizations.

Since $\kappa_L(\alpha,\beta;v)$ is invariant under scaling of v, we can set $\|v\|_2 = 1$. In view of the bounds in Lemma 4.1, the v-dependent term $\sqrt{|\alpha|^2\omega_X^2 + |\beta|^2\omega_Y^2}$ in the numerator of (3.8) is bounded below by $\min(\|A_0\|_2, \|A_m\|_2)\sqrt{|\alpha|^2 + |\beta|^2}$ for any v, and bounded above by $m \max_i \|A_i\|_2 \sqrt{|\alpha|^2 + |\beta|^2}$ when $v = e_i$ for some i. Hence to prove (4.8) it suffices to show that

$$\max_{\|v\|_2=1} |\mathsf{p}(\alpha,\beta;v)| \leq \sqrt{m}\, |\mathsf{p}(\alpha,\beta;e_1)| \quad \text{for } |\alpha| \geq |\beta|. \tag{4.10}$$

This inequality is trivial for $\beta = 0$, so we can assume $\beta \neq 0$ and divide through by β^{m-1} to rewrite the desired inequality as

$$\max_{\|v\|_2=1} |\mathsf{p}(\lambda;v)| \leq \sqrt{m}\, |\mathsf{p}(\lambda;e_1)| \quad \text{for } |\lambda| \geq 1.$$

But this inequality follows from

$$|\mathsf{p}(\lambda;v)| = |\Lambda^T v| \leq \|\Lambda\|_2 \leq \sqrt{m}\, |\lambda^{m-1}| = \sqrt{m}\, |\mathsf{p}(\lambda;e_1)|.$$

The proof of (4.9) is entirely analogous. ∎

Theorem 4.2 says that for matrix polynomials with coefficient matrices of roughly equal norm, one of the two pencils with $v = e_1$ and $v = e_m$ will always give a near

optimal condition number κ_L for a given eigenvalue; moreover, which pencil is nearly optimal depends only on whether that eigenvalue is greater than or less than 1 in modulus. Note, however, that taking the wrong choice of $v = e_1$ or $v = e_m$ can be disastrous:

$$\kappa_L(0, \beta; e_1) = \infty, \qquad \kappa_L(\alpha, 0; e_m) = \infty \tag{4.11}$$

(and in these situations the pencils are not linearizations); see the final example in section 8.

For the quadratic polynomial $Q(\lambda) = \lambda^2 A + \lambda B + C$, the pencils corresponding to $v = e_1$ and $v = e_m (= e_2)$ are, respectively (from (1.8)),

$$L_1(\lambda) = \lambda \begin{bmatrix} A & 0 \\ 0 & -C \end{bmatrix} + \begin{bmatrix} B & C \\ C & 0 \end{bmatrix}, \qquad L_2(\lambda) = \lambda \begin{bmatrix} 0 & A \\ A & B \end{bmatrix} + \begin{bmatrix} -A & 0 \\ 0 & C \end{bmatrix}. \tag{4.12}$$

These pencils were analyzed by Tisseur [16], along with a companion form linearization (which belongs to $\mathbb{L}_1$ but not $\mathbb{DL}$). She showed that if $\|A\|_2 = \|B\|_2 = \|C\|_2 = 1$ then $\kappa_{L_1}(\lambda) \le \kappa_{L_2}(\lambda)$ for $|\lambda| \ge \sqrt{2}$ and $\kappa_{L_1}(\lambda) \ge \kappa_{L_2}(\lambda)$ for $|\lambda| \le 2^{-1/2}$. Our analysis in Theorem 4.2 implies that analogous inequalities hold for arbitrary degrees m and arbitrary ρ. In fact, working directly from Lemma 4.1 we can show that

$$\begin{aligned} \kappa_L(\alpha, \beta; e_1) &\le \kappa_L(\alpha, \beta; e_m) \quad \text{if } |\alpha| \ge (\rho m)^{\frac{1}{m-1}} |\beta|, \\ \kappa_L(\alpha, \beta; e_m) &\le \kappa_L(\alpha, \beta; e_1) \quad \text{if } |\beta| \ge (\rho m)^{\frac{1}{m-1}} |\alpha|, \end{aligned}$$

with entirely analogous inequalities holding for $\kappa_L(\lambda)$.

Now we compare the optimal $\kappa_L(\alpha, \beta; v)$ with $\kappa_P(\alpha, \beta)$, the condition number of the eigenvalue for the original polynomial.

THEOREM 4.3. *Let (α, β) be a simple eigenvalue of P. Then*

$$\frac{1}{\rho} \le \frac{\inf_v \kappa_L(\alpha, \beta; v)}{\kappa_P(\alpha, \beta)} \le m^2 \rho,$$

where the weights are chosen as $\omega_i \equiv \|A_i\|_2$ for κ_P and as in (4.1) for κ_L.

Proof. From Theorem 2.3,

$$\kappa_P(\alpha, \beta) = \frac{\left(\sum_{i=0}^m |\alpha|^{2i} |\beta|^{2(m-i)} \|A_i\|_2^2\right)^{1/2} \|y\|_2 \|x\|_2}{\left| y^* (\bar{\beta} \mathcal{D}_\alpha P - \bar{\alpha} \mathcal{D}_\beta P)|_{(\alpha,\beta)} x \right|}.$$

On the other hand, for $v = v_*$ in (4.6) we have, from Theorem 3.1,

$$\kappa_L(\alpha, \beta; v_*) = \frac{\sqrt{|\alpha|^2 \|X\|_2^2 + |\beta|^2 \|Y\|_2^2}\, \|\Lambda_{\alpha,\beta}\|_2 \|y\|_2 \|x\|_2}{\left| y^* (\bar{\beta} \mathcal{D}_\alpha P - \bar{\alpha} \mathcal{D}_\beta P)|_{(\alpha,\beta)} x \right|}. \tag{4.13}$$

If L is not a linearization for $v = v_*$ then we need to interpret v_* as an arbitrarily small perturbation of v_* for which L is a linearization. Using (4.2) and (4.3) and $\sum_{i=0}^m |\alpha|^{2i} |\beta|^{2(m-i)} \|A_i\|_2^2 \ge (|\alpha|^{2m} + |\beta|^{2m}) \min(\|A_0\|_2, \|A_m\|_2)^2$, it is easy to see that

$$\frac{\kappa_L(\alpha, \beta; v_*)}{\kappa_P(\alpha, \beta)} \le \rho m^{3/2} f(\alpha, \beta),$$

where

$$f(\alpha,\beta) = \frac{\sqrt{|\alpha|^2+|\beta|^2}\left(\sum_{i=1}^m |\alpha|^{2(i-1)}|\beta|^{2(m-i)}\right)^{1/2}}{\sqrt{|\alpha|^{2m}+|\beta|^{2m}}}.$$

From (A.1) in Lemma A.1 we have $f(\alpha,\beta) \le \sqrt{m}$. The upper bound follows since $\inf_v \kappa_L(\alpha,\beta;v) \le \kappa_L(\alpha,\beta;v_*)$. For the lower bound we have, for any v with $\|v\|_2 = 1$,

$$\begin{aligned}\frac{\kappa_L(\alpha,\beta;v)}{\kappa_P(\alpha,\beta)} &= \frac{\sqrt{|\alpha|^2\|X\|_2^2+|\beta|^2\|Y\|_2^2}\,\|\Lambda_{\alpha,\beta}\|_2^2}{\left(\sum_{i=0}^m |\alpha|^{2i}|\beta|^{2(m-i)}\|A_i\|_2^2\right)^{1/2}|\mathsf{p}(\alpha,\beta;v)|}\\ &\ge \frac{\sqrt{|\alpha|^2+|\beta|^2}\min(\|A_0\|_2,\|A_m\|_2)\|\Lambda_{\alpha,\beta}\|_2}{\left(\sum_{i=0}^m |\alpha|^{2i}|\beta|^{2(m-i)}\right)^{1/2}\max_i\|A_i\|_2}\\ &\ge \frac{1}{\rho}\frac{\sqrt{|\alpha|^2+|\beta|^2}\left(\sum_{i=1}^m |\alpha|^{2(i-1)}|\beta|^{2(m-i)}\right)^{1/2}}{\left(\sum_{i=0}^m |\alpha|^{2i}|\beta|^{2(m-i)}\right)^{1/2}} =: \frac{1}{\rho}g(\alpha,\beta),\end{aligned}$$

since $|\mathsf{p}(\alpha,\beta;v)| \le \|\Lambda_{\alpha,\beta}\|_2$ by the Cauchy–Schwarz inequality. From (A.2), $g(\alpha,\beta) \ge 1$, and the lower bound follows. ∎

Now we state the analogues of Theorem 4.2 and 4.3 for $\kappa_L(\lambda)$. Recall that ρ is defined in (4.7).

THEOREM 4.4. *Let λ be a simple, finite, nonzero eigenvalue of P and consider pencils $L \in \mathbb{DL}(P)$. Take the weights* (4.1) *for κ_L. Then*

$$\kappa_L(\lambda;e_1) \le \rho m^{3/2}\inf_v \kappa_L(\lambda;v) \quad \text{if } A_0 \text{ is nonsingular and } |\lambda| \ge 1, \tag{4.14}$$

$$\kappa_L(\lambda;e_m) \le \rho m^{3/2}\inf_v \kappa_L(\lambda;v) \quad \text{if } A_m \text{ is nonsingular and } |\lambda| \le 1. \tag{4.15}$$

Proof. The proof is entirely analogous to that of Theorem 4.2. ∎

THEOREM 4.5. *Let λ be a simple, finite, nonzero eigenvalue of P. Then*

$$\left(\frac{2\sqrt{m}}{m+1}\right)\frac{1}{\rho} \le \frac{\inf_v \kappa_L(\lambda;v)}{\kappa_P(\lambda)} \le m^2\rho,$$

where the weights are chosen as $\omega_i \equiv \|A_i\|_2$ for κ_P and as in (4.1) *for L.*

Proof. The proof is very similar to that of Theorem 4.3, but with slightly different f and g having the form of f_3 and f_4 in Lemma A.1. ∎

Theorems 4.3 and 4.5 show that for polynomials whose coefficient matrices do not vary too much in norm, the best conditioned linearization in $\mathbb{DL}(P)$ for a particular eigenvalue is about as well conditioned as P itself for that eigenvalue, to within a small constant factor. This is quite a surprising result, because the condition numbers $\kappa_L(\alpha,\beta)$ and $\kappa_L(\lambda)$ permit arbitrary perturbations in $L(\lambda) = \lambda X + Y$ that do not respect the zero and repeated block structure of X and Y (as exhibited for two particular instances for $m = 2$ in (4.12)). Under the same assumptions on the $\|A_i\|_2$, by combining Theorems 4.2 and 4.3, or Theorems 4.4 and 4.5, we can conclude that, for any given eigenvalue, one of the two pencils with $v = e_1$ and $v = e_m$ will be about as well conditioned as P itself for that eigenvalue.

4.1. Several eigenvalues. Suppose now that several eigenvalues $(\alpha_1, \beta_1), \ldots, (\alpha_r, \beta_r)$ are of interest and that neither $|\alpha_i| \geq |\beta_i|$ for all i nor $|\alpha_i| \leq |\beta_i|$ for all i. A reasonable way to define a single pencil that is best for all these eigenvalues is by maximizing the 2-norm of the r-vector of the reciprocals of the eigenvalue condition numbers for the pencil. This vector can be written, using Theorem 3.1, as

$$\begin{aligned}&\operatorname{diag}\big((|\alpha_i|^2\omega_X^2 + |\beta_i|^2\omega_Y^2)^{1/2}\,\|\Lambda_{\alpha_i,\beta_i}\|_2^2\|y_i\|_2\|x_i\|_2\big)^{-1}\\ &\times \operatorname{diag}(|y_i^*(\bar\beta_i\mathcal{D}_\alpha P - \bar\alpha_i\mathcal{D}_\beta P)|_{(\alpha_i,\beta_i)}x_i|)\begin{bmatrix}\Lambda_{\alpha_1,\beta_1}^T\\ \vdots\\ \Lambda_{\alpha_r,\beta_r}^T\end{bmatrix} v =: Bv.\end{aligned}$$

Assume that $\rho = O(1)$, so that ω_X and ω_Y in (4.1) are roughly constant in $\|v\|_2$. Then we can set $\omega_X = \omega_Y = 1$ and define the optimal v as the right singular vector corresponding to the largest singular value of B. This approach requires knowledge of the eigenvectors x_i and y_i as well as the λ_i. If the eigenvectors are not known then we can simplify B further to

$$\operatorname{diag}\Big((|\alpha_i|^2 + |\beta_i|^2)^{1/2}\,\|\Lambda_{\alpha_i,\beta_i}\|_2^2\Big)^{-1}\begin{bmatrix}\Lambda_{\alpha_1,\beta_1}^T\\ \vdots\\ \Lambda_{\alpha_r,\beta_r}^T\end{bmatrix}.$$

So far we have implicitly assumed that we have a good estimate of the eigenvalues of interest. Suppose, instead, that we know only a region S of the complex plane in which the eigenvalues of interest lie. In this case a natural approach is to try to minimize the v-dependent part of the eigenvalue condition number over S. Continuing to assume $\rho = O(1)$, and working now with $\kappa_{_L}(\lambda; v)$, the problem becomes to find the v that achieves the maximum in the problem

$$\max_{\|v\|_2=1}\ \min_{\lambda\in S} |\mathsf{p}(\lambda; v)|.$$

This uniform (or Chebyshev) complex approximation problem can be expressed as a semi-infinite programming problem and solved by numerical methods for such problems [14, sect. 2.3].

5. Quadratic polynomials. We now concentrate our attention on quadratic polynomials, $Q(\lambda) = \lambda^2 A + \lambda B + C$, as these are in practice the most important polynomials of degree 2 or higher. Write

$$a = \|A\|_2, \quad b = \|B\|_2, \quad c = \|C\|_2. \tag{5.1}$$

The quantity ρ in Theorems 4.2–4.5 is now

$$\rho = \frac{\max(a, b, c)}{\min(a, c)}.$$

Clearly, ρ is of order 1 if

$$b \lesssim \max(a, c) \quad \text{and} \quad a \approx c.$$

If these conditions are not satisfied then we can consider scaling Q. Write $\lambda = \mu\gamma$, $\gamma \in \mathbb{R}$ and

$$Q(\lambda) = \lambda^2 A + \lambda B + C = \mu^2(\gamma^2 A) + \mu(\gamma B) + C =: \mu^2\widetilde{A} + \mu\widetilde{B} + \widetilde{C} =: \widetilde{Q}(\mu). \tag{5.2}$$

The γ that minimizes $\max(\|\widetilde{A}\|_2/\|\widetilde{B}\|_2, \|\widetilde{C}\|_2/\|\widetilde{B}\|_2) = \max(\gamma a/b, c/(\gamma b))$ is easily seen to be

$$\gamma = \sqrt{c/a}, \tag{5.3}$$

and it yields

$$\|\widetilde{A}\|_2 = c, \quad \|\widetilde{B}\|_2 = b\sqrt{c/a}, \quad \|\widetilde{C}\|_2 = c.$$

Hence, for the scaled problem,

$$\rho = \max(1, b/\sqrt{ac}).$$

This scaling is intended to improve the conditioning of the linearizations, but what does it do to the conditioning of the quadratic itself? It is easy to see that $\kappa_P(\lambda)$ is invariant under scaling when $\omega_i = \|A_i\|_2$, but that $\kappa_P(\alpha, \beta)$ is scale-dependent. We note that the scaling (5.2) and (5.3) is used by Fan, Lin, and Van Dooren [4]; see section 7.

With these observations we can combine and specialize Theorems 4.4 and 4.5 as follows.

THEOREM 5.1. *Let λ denote a simple eigenvalue of $Q(\lambda) = \lambda^2 A + \lambda B + C$ or of the scaled quadratic $\widetilde{Q}$ defined by* (5.2) *and* (5.3). *Take the weights* (4.1) *for $\kappa_L(\lambda)$. With the notation* (5.1)*, assume that either*

- *$b \lesssim \max(a, c)$ and $a \approx c$, in which case let $P = Q$ and $L \in \mathbb{DL}(Q)$, or*
- *$b \lesssim \sqrt{ac}$, in which case let $P = \widetilde{Q}$ and $L \in \mathbb{DL}(\widetilde{Q})$.*

Then if C is nonsingular and $|\lambda| \geq 1$, the linearization with $v = e_1$ has $\kappa_L(\lambda; e_1) \approx \kappa_P(\lambda)$, while if A is nonsingular and $|\lambda| \leq 1$, the linearization with $v = e_2$ has $\kappa_L(\lambda; e_2) \approx \kappa_P(\lambda)$.

If we think of Q as representing a mechanical system with damping, then the near-optimality of the $v = e_1$ and $v = e_2$ linearizations holds for Q that are not too heavily damped. One class of Q for which $b \lesssim \sqrt{ac}$ automatically holds is the elliptic Q [8], [11]: those for which A is Hermitian positive definite, B and C are Hermitian, and $(x^*Bx)^2 < 4(x^*Ax)(x^*Cx)$ for all nonzero $x \in \mathbb{C}^n$.

An analogue of Theorem 5.1 for $\kappa_L(\alpha, \beta)$ can be obtained from Theorems 4.2 and 4.3.

6. Connection with linearization of reversal of P. Consider the quadratic $Q(\lambda) = \lambda^2 A + \lambda B + C$ and the "reversed" quadratic $\mathrm{rev}Q(\lambda) = \lambda^2 C + \lambda B + A$, whose eigenvalues are the reciprocals of those of Q. Tisseur [16, Lem. 10] shows that if λ is a simple, finite, nonzero eigenvalue of Q and $\mu = 1/\lambda$ the corresponding simple eigenvalue of $\mathrm{rev}Q$ then, with the weights (4.1), $\kappa_{\widetilde{L}_1}(\mu) = \kappa_{L_2}(\lambda)$ and $\kappa_{\widetilde{L}_2}(\mu) = \kappa_{L_1}(\lambda)$, where L_1 and L_2 are the pencils corresponding to $v = e_1$ and $v = e_2$ given in (4.12) and $\widetilde{L}_1$ and $\widetilde{L}_2$ are the corresponding pencils for $\mathrm{rev}Q$. In essence this result says that one cannot improve the condition of an eigenvalue of a linearization by regarding it as the reciprocal of an eigenvalue of the reversed quadratic. In this section we generalize this result in three respects: to any vector v (not just $v = e_1$ or e_2), to arbitrary degree polynomials, and to zero and infinite eigenvalues.

Define

$$\mathrm{rev}P(\lambda) = \lambda^m P(1/\lambda),$$

where P has degree m, which is the polynomial obtained by reversing the order of the coefficient matrices of P. Let $L(\lambda) = \lambda X + Y \in \mathbb{DL}(P)$ with vector v and $\widetilde{L}(\lambda) = \lambda\widetilde{X} + \widetilde{Y} \in \mathbb{DL}(\mathrm{rev}P)$ with vector Rv, where

$$R = \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix} \in \mathbb{R}^{m\times m}.$$

LEMMA 6.1. *L is a linearization for P if and only if $\widetilde{L}$ is a linearization for* rev$P$.

Proof. The roots of $\mathsf{p}(\lambda; Rv)$ are the reciprocals of the roots of $\mathsf{p}(\lambda; v)$, while the eigenvalues of rev$P$ are the reciprocals of the eigenvalues of P. The result now follows from [13, Thm. 6.7]. ∎

We now work with the condition number $\kappa_L(\alpha, \beta)$. Note that (α, β) is an eigenvalue of P with right and left eigenvectors x and y if and only if (β, α) is an eigenvalue of rev$P$ with right and left eigenvectors x and y. Also note that in homogeneous variables $\mathrm{rev}P(\alpha, \beta) = P(\beta, \alpha)$.

LEMMA 6.2. *If the weights ω_X and ω_Y for L and weights $\omega_{\widetilde{X}}$ and $\omega_{\widetilde{Y}}$ for $\widetilde{L}$ satisfy $\omega_X = \omega_{\widetilde{Y}}$ and $\omega_Y = \omega_{\widetilde{X}}$ then $\kappa_L(\alpha, \beta) = \kappa_{\widetilde{L}}(\beta, \alpha)$.*

Proof. We have, from (3.8),

$$\kappa_L(\alpha,\beta) = \frac{\sqrt{|\alpha|^2\omega_X^2 + |\beta|^2\omega_Y^2}}{|\mathsf{p}(\alpha,\beta;v)|} \cdot \frac{\|\Lambda_{\alpha,\beta}\|_2^2\,\|y\|_2\|x\|_2}{\left|y^*(\bar\beta\mathcal{D}_\alpha P - \bar\alpha\mathcal{D}_\beta P)|_{(\alpha,\beta)}x\right|},$$

$$\kappa_{\widetilde{L}}(\beta,\alpha) = \frac{\sqrt{|\beta|^2\omega_{\widetilde{X}}^2 + |\alpha|^2\omega_{\widetilde{Y}}^2}}{|\mathsf{p}(\beta,\alpha;Rv)|} \cdot \frac{\|\Lambda_{\beta,\alpha}\|_2^2\,\|y\|_2\|x\|_2}{\left|y^*(\bar\alpha\mathcal{D}_\alpha\mathrm{rev}P - \bar\beta\mathcal{D}_\beta\mathrm{rev}P)|_{(\beta,\alpha)}x\right|}.$$

We show that each of the four terms in the first expression equals the corresponding term in the second expression. The assumptions on the weights clearly imply equality of the square root terms. Next, $\Lambda_{\beta,\alpha} = R\Lambda_{\alpha,\beta}$, so $\Lambda_{\beta,\alpha}$ and $\Lambda_{\alpha,\beta}$ have the same 2-norm, while $\mathsf{p}(\alpha, \beta; v) \equiv \mathsf{p}(\beta, \alpha; Rv)$. Finally,

$$\begin{aligned}(\bar\alpha\mathcal{D}_\alpha\mathrm{rev}P - \bar\beta\mathcal{D}_\beta\mathrm{rev}P)|_{(\beta,\alpha)} &= \bar\alpha(\mathcal{D}_\alpha\mathrm{rev}P)|_{(\beta,\alpha)} - \bar\beta(\mathcal{D}_\beta\mathrm{rev}P)|_{(\beta,\alpha)}\\ &= \bar\alpha(\mathcal{D}_\beta P)|_{(\alpha,\beta)} - \bar\beta(\mathcal{D}_\alpha P)|_{(\alpha,\beta)}\\ &= -(\bar\beta\mathcal{D}_\alpha P - \bar\alpha\mathcal{D}_\beta P)|_{(\alpha,\beta)},\end{aligned}$$

which implies the equality of the final two denominator terms. ∎

Do the conditions $\omega_X = \omega_{\widetilde{Y}}$ and $\omega_Y = \omega_{\widetilde{X}}$ hold for the natural choice of weights $\omega_X \equiv \|X\|_2$, $\omega_Y \equiv \|Y\|_2$? The next lemma shows that they do, and shows an even stronger relationship between L and $\widetilde{L}$.

LEMMA 6.3. *We have*

$$\widetilde{L}(\lambda) = (R \otimes I_n)\mathrm{rev}L(\lambda)(R \otimes I_n), \tag{6.1}$$

and so $\widetilde{X} = (R \otimes I_n)Y(R \otimes I_n)$ and $\widetilde{Y} = (R \otimes I_n)X(R \otimes I_n)$. Hence $\|\widetilde{X}\| = \|Y\|$ and $\|\widetilde{Y}\| = \|X\|$ for any unitarily invariant norm.

Proof. $\widetilde{L}$ is defined as the unique pencil in $\mathbb{DL}(\mathrm{rev}P) = \mathbb{L}_1(\mathrm{rev}P) \cap \mathbb{L}_2(\mathrm{rev}P)$ corresponding to the vector Rv. Therefore to establish (6.1) it suffices to show that the pencil $(R \otimes I_n)\mathrm{rev}L(\lambda)(R \otimes I_n)$ belongs to both $\mathbb{L}_1(\mathrm{rev}P)$ and $\mathbb{L}_2(\mathrm{rev}P)$ with

vector Rv, that is, it satisfies the appropriate versions of properties (1.5) and (1.6). The other results then follow.

Recall that $\mathrm{rev}P(\lambda) = \lambda^m P(1/\lambda)$ and note that $\lambda^{m-1}\Lambda(1/\lambda) = R\Lambda$, where $\Lambda(\lambda) \equiv \Lambda$ is defined in (1.4). If $L \in \mathbb{L}_1(P)$ with vector v then

$$\begin{aligned}
& L(\lambda)\cdot(\Lambda \otimes I_n) = v \otimes P(\lambda) \\
\Rightarrow\quad & L(1/\lambda)\cdot(\Lambda(1/\lambda) \otimes I_n) = v \otimes P(1/\lambda) \\
\Rightarrow\quad & \lambda L(1/\lambda)\cdot(\lambda^{m-1}\Lambda(1/\lambda) \otimes I_n) = v \otimes \lambda^m P(1/\lambda) \\
\Rightarrow\quad & \mathrm{rev}L(\lambda)\cdot(R\Lambda \otimes I_n) = v \otimes \mathrm{rev}P(\lambda) \\
\Rightarrow\quad & (R \otimes I_n)\mathrm{rev}L(\lambda)(R \otimes I_n)\cdot(\Lambda \otimes I_n) = (R \otimes I_n)(v \otimes \mathrm{rev}P(\lambda)) = Rv \otimes \mathrm{rev}P(\lambda),
\end{aligned}$$

which means that $(R \otimes I_n)\mathrm{rev}L(\lambda)(R \otimes I_n) \in \mathbb{L}_1(\mathrm{rev}P)$ with vector Rv.

Similarly, it can be shown that $L \in \mathbb{L}_2(P)$ with vector v implies that $(R \otimes I_n)\mathrm{rev}L(\lambda)(R \otimes I_n) \in \mathbb{L}_2(\mathrm{rev}P)$ with vector Rv. ∎

Combining the previous three lemmas we obtain the following generalization of Tisseur [16, Lem. 10].

THEOREM 6.4. *Let (α, β) be a simple eigenvalue of P, so that (β, α) is a simple eigenvalue of* rev$P$. *Suppose $L \in \mathbb{DL}(P)$ with vector v is a linearization of P. Then $\widetilde{L} \in \mathbb{DL}(\mathrm{rev}P)$ with vector Rv is a linearization of* rev$P$ *and, if the weights are chosen as in* (4.1), $\kappa_L(\alpha, \beta) = \kappa_{\widetilde{L}}(\beta, \alpha)$.

An analogue of Theorem 6.4 stating that $\kappa_L(\lambda) = \kappa_{\widetilde{L}}(1/\lambda)$ for finite, nonzero λ can also be derived.

7. Companion linearizations. Associated with P are two companion form pencils, $C_1(\lambda) = \lambda X_1 + Y_1$ and $C_2(\lambda) = \lambda X_2 + Y_2$, called the first and second companion forms [12, sect. 14.1], respectively, where

$$X_1 = X_2 = \mathrm{diag}(A_m, I_n, \ldots, I_n),$$

$$Y_1 = \begin{bmatrix} A_{m-1} & A_{m-2} & \ldots & A_0 \\ -I_n & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & -I_n & 0 \end{bmatrix}, \qquad Y_2 = \begin{bmatrix} A_{m-1} & -I_n & \ldots & 0 \\ A_{m-2} & 0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & -I_n \\ A_0 & 0 & \ldots & 0 \end{bmatrix}.$$

The pencil C_1 belongs to $\mathbb{L}_1(P)$ with $v = e_1$ in (1.5), while C_2 belongs to $\mathbb{L}_2(P)$ with $w = e_1$ in (1.6). Neither pencil is in $\mathbb{DL}(P)$, but both are always linearizations [12, sect. 14.1].

We wish to compare the conditioning of C_1 and C_2 with that of P and of an appropriate $\mathbb{DL}(P)$ linearization. Our first result shows that it suffices to analyze the conditioning of C_1, because any results about the conditioning of C_1 translate to C_2 simply by transposing the coefficient matrices A_i.

LEMMA 7.1. *Let λ, or (α, β) in homogeneous form, be a simple eigenvalue of P, and take $\omega_i = \|A_i\|_2$. Then*

$$\kappa_P(\alpha, \beta) = \kappa_{P^T}(\alpha, \beta), \qquad \kappa_P(\lambda) = \kappa_{P^T}(\lambda).$$

Moreover,

$$\kappa_{C_2(P)}(\alpha, \beta) = \kappa_{C_1(P^T)}(\alpha, \beta), \qquad \kappa_{C_2(P)}(\lambda) = \kappa_{C_1(P^T)}(\lambda),$$

where $C_i(P)$, $i = 1, 2$, denotes the ith companion linearization for P, and P^T denotes the polynomial obtained by transposing each coefficient matrix A_i.

Proof. If (λ, x, y) is an eigentriple for P then $(\lambda, \overline{y}, \overline{x})$ is an eigentriple for P^T. The first two equalities follow by considering the formulae (2.2) and (2.5). It is easy to see that $C_2(P) = C_1(P^T)^T$. The second pair of equalities are therefore special cases of the first. ∎

For the rest of the section we work with λ and $\kappa(\lambda)$; for (α, β) and $\kappa(\alpha, \beta)$ analogous results hold. We first obtain a formula for left eigenvectors w^* of C_1.

LEMMA 7.2. *The vector $y \in \mathbb{C}^n$ is a left eigenvector of P corresponding to a simple, finite, nonzero eigenvalue λ if and only if*

$$w = \begin{bmatrix} I \\ (\lambda A_m + A_{m-1})^* \\ \vdots \\ (\lambda^{m-1} A_m + \lambda^{m-2} A_{m-1} + \cdots + A_1)^* \end{bmatrix} y \tag{7.1}$$

is a left eigenvector of C_1 corresponding to λ.

Proof. Since C_1 is a linearization of P, λ is a simple eigenvalue of C_1. The proof therefore consists of a direct verification that $w^* C_1(\lambda) = 0$. ∎

Lemma 7.2 shows that, even though $C_1 \notin \mathbb{L}_2(P)$, a left eigenvector of P can be recovered from one of C_1—simply by reading off the leading n components.

Since $C_1 \in \mathbb{L}_1(P)$, we know that the right eigenvectors z and x of C_1 and P are related by $z = \Lambda \otimes x$. Evaluating (3.10) (which holds for any member of $\mathbb{L}_1$) with $L = C_1$ at an eigenvalue λ, then multiplying on the left by w^* and on the right by $1 \otimes x = x$, we obtain

$$w^* C_1'(\lambda) z = w^* (v \otimes P'(\lambda) x).$$

Using the formula (7.1) for w and the fact that $v = e_1$ gives

$$w^* C_1'(\lambda) z = y^* P'(\lambda) x.$$

By applying Theorem 2.1 to C_1 we obtain the following analogue of Theorem 3.2.

THEOREM 7.3. *Let λ be a simple, finite, nonzero eigenvalue of P with right and left eigenvectors x and y, respectively. Then, for the first companion linearization $C_1(\lambda) = \lambda X_1 + Y_1$,*

$$\kappa_{C_1}(\lambda) = \frac{(|\lambda| \omega_{X_1} + \omega_{Y_1}) \, \|w\|_2 \|\Lambda\|_2 \|x\|_2}{|\lambda| \, |y^* P'(\lambda) x|},$$

where w is given by (7.1).

Now we can compare the condition number of the first companion form with that of P. We have

$$\frac{\kappa_{C_1}(\lambda)}{\kappa_P(\lambda)} = \frac{\|w\|_2}{\|y\|_2} \cdot \frac{(|\lambda| \omega_{X_1} + \omega_{Y_1}) \, \|\Lambda\|_2}{\sum_{i=0}^m |\lambda|^i \omega_i}.$$

We choose the weights $\omega_{X_1} = \|X_1\|_2$, $\omega_{Y_1} = \|Y_1\|_2$, and $\omega_i = \|A_i\|_2$ in (2.2). We therefore need bounds on the norms of X_1 and Y_1. These are provided by the next lemma, which is similar to Lemma 4.1.

LEMMA 7.4. *For $C_1(\lambda) = \lambda X_1 + Y_1$ we have $\|X_1\|_2 = \max(\|A_m\|_2, 1)$ and*

$$\max\Big(1, \max_{i=0:m-1} \|A_i\|_2\Big) \le \|Y_1\|_2 \le m \max\Big(1, \max_{i=0:m-1} \|A_i\|_2\Big). \tag{7.2}$$

Proof. The proof is straightforward, using (4.5). ∎

For notational simplicity we will now concentrate on the quadratic case, $m = 2$. With the notation (5.1), we have

$$\frac{\psi}{2^{1/2}}\frac{\|w\|_2}{\|y\|_2} \le \frac{\kappa_{C_1}(\lambda)}{\kappa_P(\lambda)} \le 2\psi\frac{\|w\|_2}{\|y\|_2}, \tag{7.3}$$

where

$$\psi = \frac{(1+|\lambda|)\big(\max(a,1)|\lambda| + \max(b,c,1)\big)}{a|\lambda|^2 + b|\lambda| + c} \ge 1$$

and

$$\frac{\|w\|_2}{\|y\|_2} = \frac{\left\|\begin{bmatrix} I \\ (\lambda A + B)^* \end{bmatrix} y\right\|_2}{\|y\|_2} = \frac{\left\|\begin{bmatrix} I \\ (\lambda^{-1}C)^* \end{bmatrix} y\right\|_2}{\|y\|_2} \tag{7.4}$$

satisfies

$$1 \le \frac{\|w\|_2}{\|y\|_2} \le \min\big((1 + (|\lambda|a+b)^2)^{1/2}, (1 + c^2/|\lambda|^2)^{1/2}\big).$$

Therefore $\kappa_{C_1}(\lambda)$ will be of the same order of magnitude as $\kappa_P(\lambda)$ only if both ψ and $\|w\|_2/\|y\|_2$ are of order 1. It is difficult to characterize when these conditions hold. However, it is clear that, unlike for the $\mathbb{DL}(P)$ linearizations, the condition of C_1 is affected by scaling $A_i \leftarrow \gamma A_i$, $i = 0\colon m$, as might be expected in view of the mixture of identity matrices and A_i that make up the blocks of X_1 and Y_1. Indeed if $a, b, c \ll 1$, then $\psi \gg 1$, while if $a, b, c \gg |\lambda| \ge 1$, then $\|w\|_2/\|y\|_2 \gg 1$, unless y is nearly a null vector for $(\lambda A + B)^*$ and C^*. The only straightforward conditions that guarantee $\kappa_{C_1}(\lambda) \approx \kappa_P(\lambda)$ are $a \approx b \approx c \approx 1$: then $\psi \approx 1$ and one of the two expressions for $\|w\|_2/\|y\|_2$ in (7.4) is clearly of order 1 (the first if $|\lambda| \le 1$, otherwise the second). The predilection of the first companion form for coefficient matrices of unit 2-norm was shown from a different viewpoint by Tisseur [16, Thm. 7]: she proves that when $a = b = c = 1$, applying a backward stable solver to the companion pencil is backward stable for the original quadratic.

It is natural to scale the problem to try to bring the 2-norms of A, B, and C close to 1. The scaling of Fan, Lin, and Van Dooren [4], which was motivated by backward error considerations, has precisely this aim. It converts $Q(\lambda) = \lambda^2 A + \lambda B + C$ to $\widetilde{Q}(\mu) = \mu^2\widetilde{A} + \mu\widetilde{B} + \widetilde{C}$, where

$$\lambda = \gamma\mu, \qquad Q(\lambda)\delta = \mu^2(\gamma^2\delta A) + \mu(\gamma\delta B) + \delta C \equiv \widetilde{Q}(\mu), \tag{7.5a}$$

$$\gamma = \sqrt{c/a}, \qquad \delta = 2/(c + b\gamma). \tag{7.5b}$$

This is the scaling γ we used in section 5 combined with the multiplication of each coefficient matrix by δ.

Now we compare $\kappa_{C_1}(\lambda)$ with $\kappa_L(\lambda; v_*)$, where v_* for λ is defined analogously to v_* for (α, β) in (4.6) by

$$v_* = \frac{\overline{\Lambda}}{\|\Lambda\|_2}, \qquad |\mathsf{p}(\lambda; v_*)| = \|\Lambda\|_2.$$

We have, from (3.11),

$$\kappa_L(\lambda; v_*) = \frac{(|\lambda|\omega_X + \omega_Y)\|\Lambda\|_2\|y\|_2\|x\|_2}{|\lambda|\,|y^*P'(\lambda)x|},$$

and so

$$\frac{\kappa_{C_1}(\lambda)}{\kappa_L(\lambda; v_*)} = \frac{\|w\|_2}{\|y\|_2} \cdot \frac{|\lambda|\omega_{X_1} + \omega_{Y_1}}{|\lambda|\omega_X + \omega_Y}.$$

Again, specializing to $m = 2$, and using Lemmas 4.1 and 7.4, we have

$$\frac{\|w\|_2}{\|y\|_2} \cdot \frac{\big(\max(a,1)|\lambda| + \max(b,c,1)\big)}{2^{3/2}\max(a,b,c)(1+|\lambda|)} \le \frac{\kappa_{C_1}(\lambda)}{\kappa_L(\lambda; v_*)} \le \frac{\|w\|_2}{\|y\|_2} \cdot \frac{\big(\max(a,1)|\lambda| + 2\max(b,c,1)\big)}{a|\lambda| + c}. \tag{7.6}$$

If $a \approx b \approx c \approx 1$ then we can conclude that $\kappa_{C_1}(\lambda) \approx \kappa_L(\lambda; v_*)$. However, $\kappa_{C_1}(\lambda) \gg \kappa_L(\lambda; v_*)$ if $\|w\|_2/\|y\|_2 \gg 1$ or if (for example) $a, b, c \ll 1$.

Our results for the companion forms are not as neat as those in section 4 for the $\mathbb{DL}(P)$ linearizations, which focus attention on a single, easily computed or estimated, scalar parameter ρ. The conditioning of the companion forms relative to P and to the class $\mathbb{DL}(P)$ depends on both (a) the ratios of norms of left eigenvectors of C_1 and P, and (b) rational functions of the coefficient matrix norms and λ. It does not seem possible to bound the norm ratio in a useful way a priori. Therefore the only easily checkable conditions that we can identify under which the companion forms can be guaranteed to be optimally conditioned are $\|A_i\|_2 \approx 1$, $i = 0\colon m$ (our proof of this fact for $m = 2$ is easily seen to generalize to arbitrary m).

Finally, we note that the bounds (7.3) and (7.6) remain true when "λ" is replaced by "α, β," with just minor changes to the constants.

8. Numerical experiments. We illustrate the theory on four quadratic eigenvalue problems. Our experiments were performed in MATLAB 7, for which the unit roundoff is $2^{-53} \approx 1.1 \times 10^{-16}$. To obtain the angular error $\theta((\alpha,\beta),(\widetilde{\alpha},\widetilde{\beta}))$ for a computed eigenvalue $(\widetilde{\alpha},\widetilde{\beta})$, we took as exact eigenvalue (α,β) the value computed in MATLAB's VPA arithmetic at 40 digit precision. In our figures, the x-axis is the eigenvalue index and the eigenvalues are sorted in increasing order of absolute value. We compare the condition numbers of the quadratic Q, the first companion form, and the $\mathbb{DL}(Q)$ linearizations with $v = e_1$ and $v = e_2$. All our problems have (real) symmetric coefficient matrices so we know from Lemma 7.1 that the second companion form has exactly the same condition numbers as the first companion form. In two of the problems we apply the scaling given by (7.5). Table 8.1 reports the problem sizes, the coefficient matrix norms, and the values of ρ in (4.7) before and after scaling.

Our first problem shows the benefits of scaling. It comes from applying the Galerkin method to a PDE describing the wave motion of a vibrating string with clamped ends in a spatially inhomogeneous environment [5], [8]. The quadratic Q is elliptic; the eigenvalues are nonreal and have absolute values in the interval $[1, 25]$. Figure 8.1 shows the condition numbers $\kappa_L(\alpha,\beta)$ for the $\mathbb{DL}(Q)$ linearization with $v = e_1$ and the first companion linearization, the condition number $\kappa_P(\alpha,\beta)$ for Q, and the angular errors in the eigenvalues computed by applying the QZ algorithm to the two linearizations. Figure 8.2 shows the corresponding information for the

TABLE 8.1
Problem statistics.

Problem	Wave		Nuclear		Mass-spring	Acoustics	
n	25		8		50	107	
	Unscaled	Scaled	Unscaled	Scaled	Unscaled	Unscaled	Scaled
$\|A\|_2$	1.57e0	1.85e0	2.35e8	1.18e0	1.00e0	1.00e0	2.00e0
$\|B\|_2$	3.16e0	1.49e-1	4.35e10	8.21e-1	3.20e2	5.74e-2	3.64e-5
$\|C\|_2$	9.82e2	1.85e0	1.66e13	1.18e0	5.00e0	9.95e6	2.00e0
ρ	6.25e2	1.00e0	7.06e4	1.00e0	3.20e2	9.95e6	1.00e0

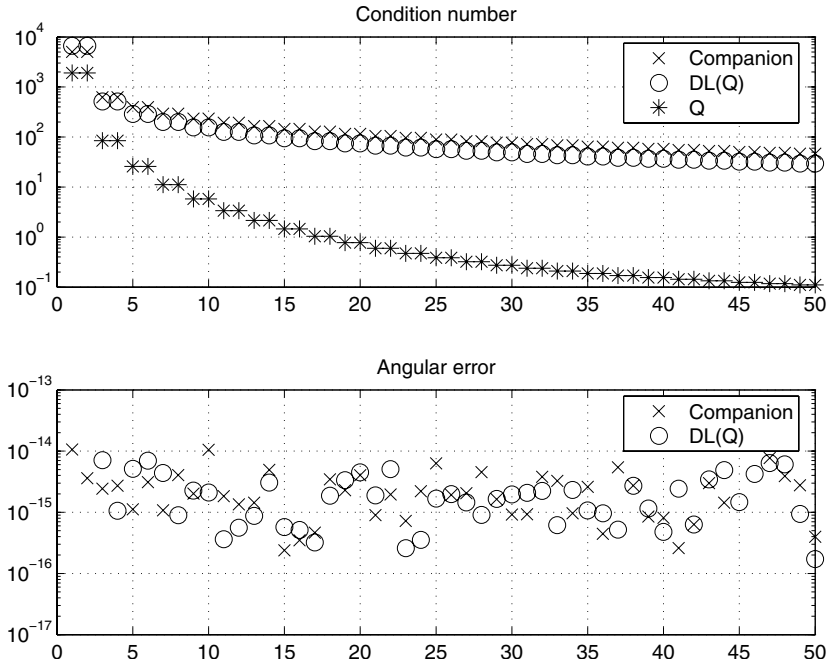


FIG. 8.1. *Wave problem, unscaled;* $v = e_1$, $\rho = 625$.

scaled problem. Since the eigenvalues are all of modulus at least 1, we know from Theorem 4.3 that for every eigenvalue, the $\mathbb{DL}(Q)$ linearization with $v = e_1$ has condition number within a factor $4\rho = 2500$ of the condition number for Q. The actual ratios are between 3.5 and 266. Since this problem is elliptic, we know from Theorem 5.1 that for the scaled problem, whose eigenvalues lie between 0.04 and 1 in modulus, the $\mathbb{DL}(\widetilde{Q})$ linearization with $v = e_2$ will have condition number similar to that of $\widetilde{Q}$ for every eigenvalue. This is confirmed by Figure 8.2; the maximum ratio of condition numbers is 3.3. The benefit of the smaller condition numbers after scaling is clear from the figures: the angular error of the computed eigenvalues is smaller by a factor roughly equal to the reduction in condition number. The behavior of the companion linearization is very similar to that of the $\mathbb{DL}(Q)$ linearizations, and this is predicted by our theory since the term $\psi\|w\|_2/\|y\|_2$ in (7.3) varies from 3.7 to 511 without scaling and only 1.0 to 4.5 with scaling.

The next problem is a simplified model of a nuclear power plant [9], [17]. There are 2 real and 14 nonreal eigenvalues, with absolute values in the interval $(17, 362)$. Since $\rho = 7 \times 10^4$, it is not surprising that the $\mathbb{DL}(Q)$ linearization with $v = e_1$ has

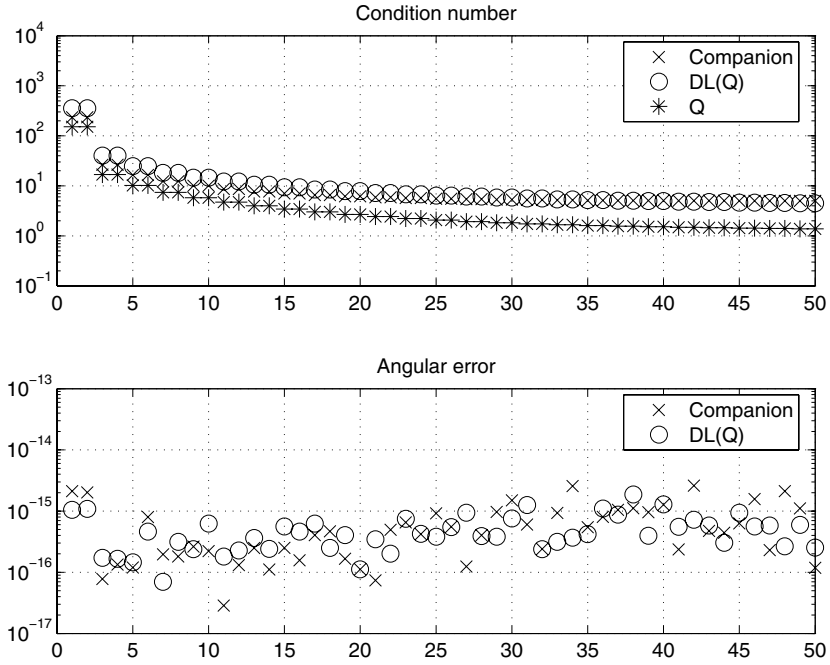


FIG. 8.2. *Wave problem, scaled;* $v = e_2$, $\rho = 1$.

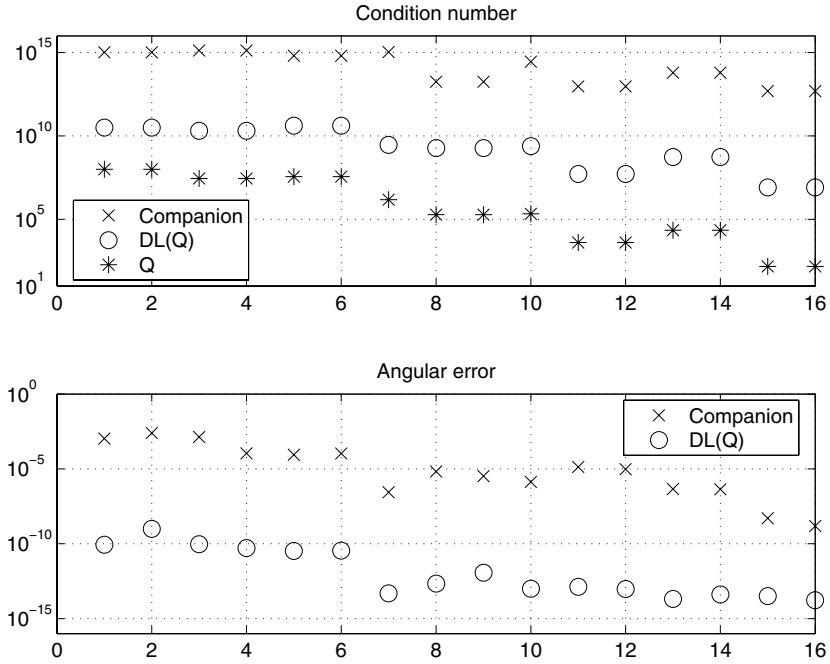


FIG. 8.3. *Nuclear power plant problem, unscaled;* $v = e_1$, $\rho = 7 \times 10^4$.

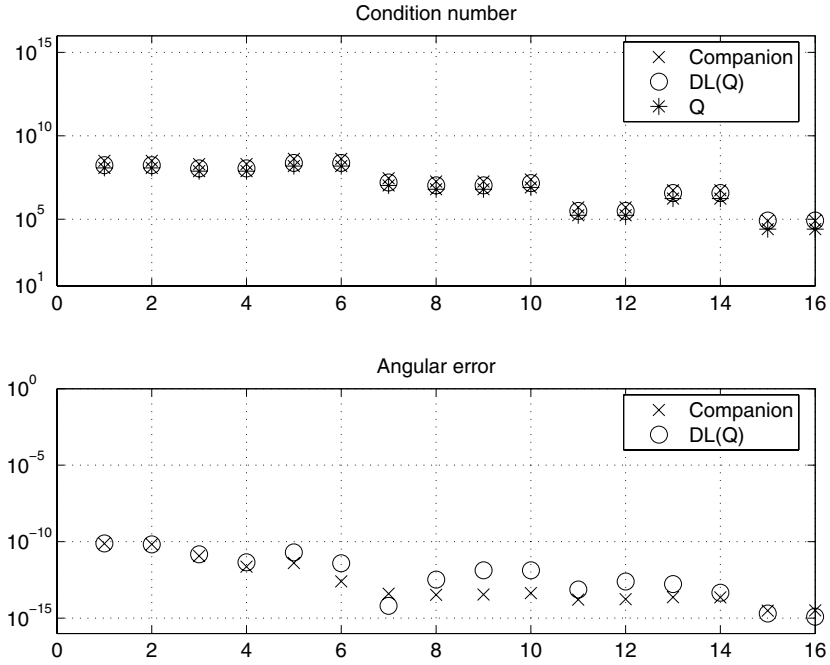


FIG. 8.4. *Nuclear power plant problem, scaled;* $v = e_2$, $\rho = 1$.

eigenvalue condition numbers up to 369 times as large as those of Q, as Figure 8.3 indicates. Although the problem is not elliptic, $\|B\|_2 \leq \sqrt{\|A\|_2\|C\|_2}$, and so our theory says that scaling will make the $\mathbb{DL}(Q)$ linearization with $v = e_2$ (since the scaled eigenvalues have modulus at most 1) optimally conditioned. This prediction is confirmed in Figure 8.4. Scaling also brings a dramatic improvement in the conditioning and accuracy of the companion linearization; again, this is predicted by our theory since the scaled problem has coefficient matrices of norm approximately 1, and the magnitude of the reduction is explained by the term $\psi\|w\|_2/\|y\|_2$ in (7.3), which has a maximum of 2×10^{10} without scaling and 1.5 with scaling. Scaling results in an increase in the condition numbers $\kappa_P(\alpha, \beta)$ by factors ranging from 1.2 to 173.

Our third problem is a standard damped mass-spring system, as described in [17, sect. 3.9]. The matrix $A = I$, B is tridiagonal with super- and subdiagonal elements all -64 and diagonal $128, 192, 192, \ldots, 192$, and C is tridiagonal with super- and subdiagonal elements all -1 and diagonal $2, 3, \ldots, 3$. Here, $\rho = 320$. The eigenvalues are all negative, with 50 eigenvalues of large modulus ranging from -320 to -6.4 and 50 small modulus eigenvalues approximately -1.5×10^{-2}. Figures 8.5 and 8.6 show the results for $v = e_1$ and $v = e_2$, respectively. Our theory suggests that for the eigenvalues of large modulus the linearization with $v = e_1$ will have nearly optimal conditioning, while for eigenvalues of small modulus the linearization with $v = e_2$ will be nearly optimal. This behavior is seen very clearly in the figures, with a sharp change in condition number at the three order of magnitude jump in the eigenvalues. This example also clearly displays nonoptimal conditioning of the first companion linearization for small eigenvalues: for the 50 eigenvalues of small modulus, $\kappa_{C_1}(\alpha, \beta)$ exceeds $\kappa_P(\alpha, \beta)$ and $\kappa_L(\alpha, \beta; e_2)$ by a factor at least 10^3, and again this is accurately reflected in the bounds (7.3). For this problem, scaling has essentially no effect on

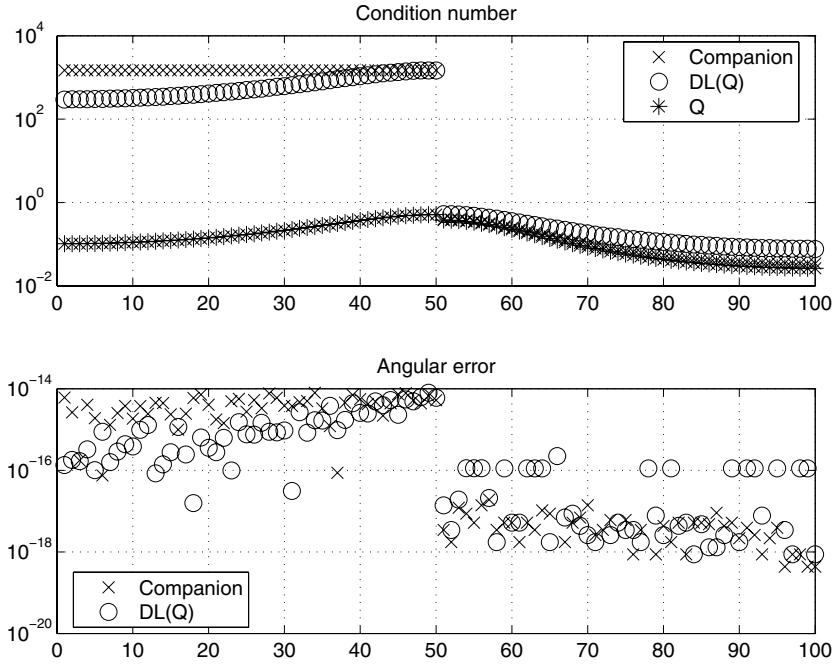


FIG. 8.5. *Damped mass-spring system, unscaled;* $v = e_1$, $\rho = 320$.

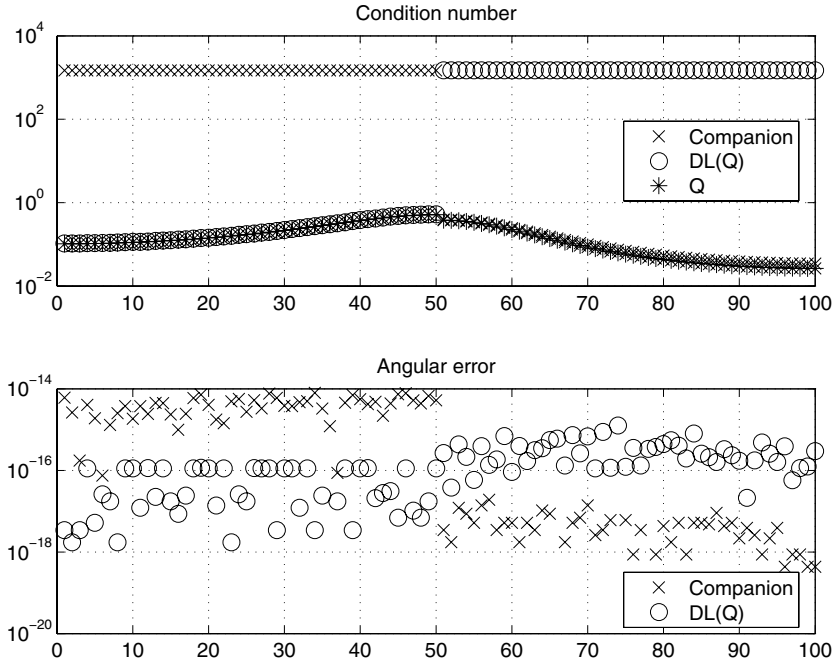


FIG. 8.6. *Damped mass-spring system, unscaled;* $v = e_2$, $\rho = 320$.

the two $\mathbb{DL}(Q)$ linearizations, but for the companion linearization it increases the condition number for the large eigenvalues and decreases it for the small eigenvalues, with the result that all the condition numbers lie between 3.6 and 13.

Finally, we describe an example that emphasizes the importance in our analysis of the condition that the pencil $L \in \mathbb{DL}(P)$ is a linearization of P. The problem is a quadratic of dimension 107 arising from an acoustical model of a speaker box [10]. After scaling, $\rho = 1$. The computed eigenvalues from the companion form have moduli of order 1, except for two eigenvalues with moduli of order 10^{-5}. We found the pencil with $v = e_2$ to have eigenvalue condition numbers of the same order of magnitude as those of Q (namely from 10^6 to 10^{13})—as predicted by the theory. But for $v = e_1$ the conditioning of L was orders of magnitude worse than that of Q for every eigenvalue, which at first sight appears to contradict the theory. The explanation is that this problem has a singular A_0 and hence a zero eigenvalue; L is therefore not a linearization for $v = e_1$, as we noted earlier: see the first sentence of the proof of Theorem 4.2 and (4.11). In fact, since $L \in \mathbb{DL}(P)$ is not a linearization for $v = e_1$ it is a nonregular pencil [13, Thm. 4.3]. This example is therefore entirely consistent with the theory.

Appendix A.

The following lemma is needed in the proofs of Theorems 4.3 and 4.5. We omit the proof.

LEMMA A.1. *Consider the functions*

$$f_1(x) = \frac{(1+x^2)(1+x^2+x^4+\cdots+x^{2(m-1)})}{1+x^{2m}},$$

$$f_2(x) = \frac{(1+x^2)(1+x^2+x^4+\cdots+x^{2(m-1)})}{1+x^2+x^4+\cdots+x^{2m}},$$

$$f_3(x) = \frac{(1+x)^2(1+x^2+x^4+\cdots+x^{2(m-1)})}{(1+x^m)^2},$$

$$f_4(x) = \frac{(1+x)^2(1+x^2+x^4+\cdots+x^{2(m-1)})}{(1+x+x^2+\cdots+x^m)^2}.$$

The functions f_1, f_2, f_3, and f_4 are all unimodal on $[0,\infty)$, with a unique interior extreme point at $x = 1$ and another extreme point at $x = 0$. In particular, we have the following sharp bounds:

$$1 \le f_1(x) \le m, \tag{A.1}$$

$$1 \le f_2(x) \le \frac{2m}{m+1}, \tag{A.2}$$

$$1 \le f_3(x) \le m, \tag{A.3}$$

$$\frac{4m}{(m+1)^2} \le f_4(x) \le 1. \tag{A.4}$$

Acknowledgments. We thank Niloufer Mackey for assistance in the preparation of this paper.

REFERENCES

[1] A. L. Andrew, K.-W. E. Chu, and P. Lancaster, *Derivatives of eigenvalues and eigenvectors of matrix functions*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 903–926.

[2] J.-P. Dedieu, *Condition operators, condition numbers, and condition number theorem for the generalized eigenvalue problem*, Linear Algebra Appl., 263 (1997), pp. 1–24.

[3] J.-P. Dedieu and F. Tisseur, *Perturbation theory for homogeneous polynomial eigenvalue problems*, Linear Algebra Appl., 358 (2003), pp. 71–94.

[4] H.-Y. Fan, W.-W. Lin, and P. Van Dooren, *Normwise scaling of second order polynomial matrices*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 252–256.

[5] P. Freitas, M. Grinfield, and P. Knight, *Stability of finite-dimensional systems with indefinite damping*, Adv. Math. Sci. Appl., 17 (1997), pp. 435–446.

[6] I. Gohberg, P. Lancaster, and L. Rodman, *Matrix Polynomials*, Academic Press, New York, 1982.

[7] N. J. Higham, D. S. Mackey, N. Mackey, and F. Tisseur, *Symmetric Linearizations for Matrix Polynomials*, SIAM J. Matrix Anal. Appl., to appear.

[8] N. J. Higham, F. Tisseur, and P. M. Van Dooren, *Detecting a definite Hermitian pair and a hyperbolic or elliptic quadratic eigenvalue problem, and associated nearness problems*, Linear Algebra Appl., 351–352 (2002), pp. 455–474.

[9] T. Itoh, *Damped vibration mode superposition method for dynamic response analysis*, Earthquake Engrg. Struct. Dyn., 2 (1973), pp. 47–57.

[10] T. R. Kowalski, *Extracting a Few Eigenpairs of Symmetric Indefinite Matrix Pencils*, Ph.D. thesis, Department of Mathematics, University of Kentucky, Lexington, KY, 2000.

[11] P. Lancaster, *Quadratic eigenvalue problems*, Linear Algebra Appl., 150 (1991), pp. 499–506.

[12] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, 2nd ed., Academic Press, London, 1985.

[13] 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.

[14] R. Reemtsen and S. Görner, *Numerical methods for semi-infinite programming: A survey*, in Semi-Infinite Programming, R. Reemtsen and J.-J. Rückmann, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 195–275.

[15] G. W. Stewart and J. Sun, *Matrix Perturbation Theory*, Academic Press, London, 1990.

[16] F. Tisseur, *Backward error and condition of polynomial eigenvalue problems*, Linear Algebra Appl., 309 (2000), pp. 339–361.

[17] F. Tisseur and K. Meerbergen, *The quadratic eigenvalue problem*, SIAM Rev., 43 (2001), pp. 235–286.

[18] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, UK, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 4, pp. 1029–1051

STRUCTURED POLYNOMIAL EIGENVALUE PROBLEMS: GOOD VIBRATIONS FROM GOOD LINEARIZATIONS*

D. STEVEN MACKEY†, NILOUFER MACKEY‡, CHRISTIAN MEHL§, AND VOLKER MEHRMANN§

Abstract. Many applications give rise to nonlinear eigenvalue problems with an underlying structured matrix polynomial. In this paper several useful classes of structured polynomials (e.g., palindromic, even, odd) are identified and the relationships between them explored. A special class of linearizations which reflect the structure of these polynomials, and therefore preserve symmetries in their spectra, is introduced and investigated. We analyze the existence and uniqueness of such linearizations and show how they may be systematically constructed.

Key words. nonlinear eigenvalue problem, palindromic matrix polynomial, even matrix polynomial, odd matrix polynomial, Cayley transformation, structured linearization, preservation of eigenvalue symmetry

AMS subject classifications. 65F15, 15A18, 15A57, 93B60

DOI. 10.1137/050628362

1. Introduction.

We consider $n \times n$ matrix polynomials of the form

$$P(\lambda) = \sum_{i=0}^{k} \lambda^i A_i, \quad A_0, \ldots, A_k \in \mathbb{F}^{n\times n},\ A_k \neq 0, \tag{1.1}$$

where $\mathbb{F}$ denotes the field $\mathbb{R}$ or $\mathbb{C}$. The numerical solution of the associated polynomial eigenvalue problem $P(\lambda)x = 0$ is one of the most important tasks in the vibration analysis of buildings, machines, and vehicles [11], [22], [35]. In many applications, several of which are summarized in [27], the coefficient matrices have a further structure which reflects the properties of the underlying physical model, and it is important that numerical methods respect this structure.

Our main motivation stems from a project with the company SFE GmbH in Berlin which investigates rail traffic noise caused by high speed trains [16], [17]. The eigenvalue problem that arises in this project from the vibration analysis of rail tracks has the form

$$(\lambda^2 A + \lambda B + A^T)x = 0, \tag{1.2}$$

where A, B are complex square matrices with B complex symmetric and A singular. The impact of the theory developed in this paper on the solution of this particular eigenvalue problem will be discussed further in section 4. (See also the article in [20].)

*Received by the editors April 1, 2005; accepted for publication (in revised form) by I. C. F. Ipsen April 20, 2006; published electronically December 18, 2006. This work was supported by Deutsche Forschungsgemeinschaft through MATHEON, the DFG Research Center *Mathematics for key technologies* in Berlin.

http://www.siam.org/journals/simax/28-4/62836.html

†School of Mathematics, The University of Manchester, Sackville Street, Manchester, M60 1QD, UK (smackey@ma.man.ac.uk). The work of this author was supported by Engineering and Physical Sciences Research Council grant GR/S31693.

‡Department of Mathematics, Western Michigan University, Kalamazoo, MI 49008 (nil.mackey@wmich.edu, http://homepages.wmich.edu/~mackey/).

§Technische Universität Berlin, Institut für Mathematik, Sekretariat MA 4-5, D-10623 Berlin, Germany (mehl@math.tu-berlin.de, mehrmann@math.tu-berlin.de).

Observe that the matrix polynomial in (1.2) has the property that reversing the order of the coefficient matrices, followed by taking their transpose, leads back to the original matrix polynomial. By analogy with linguistic palindromes, of which

"sex at noon taxes"

is perhaps a less well-known example,[1] we say such matrix polynomials are *T-palindromic*.

Quadratic real and complex T-palindromic eigenvalue problems also arise in the mathematical modeling and numerical simulation of the behavior of periodic surface acoustic wave (SAW) filters [36], whereas the computation of the Crawford number [15] associated with the perturbation analysis of symmetric generalized eigenvalue problems produces a quadratic $*$-palindromic eigenvalue problem, where $*$ stands for conjugate transpose. Higher order matrix polynomials with a $*$-palindromic structure arise in the optimal control of higher order difference equations [27].

A related class of structured eigenvalue problems arises in the study of corner singularities in anisotropic elastic materials [3], [4], [25], [33] and gyroscopic systems [35]. Here the problem is of the form

$$P(\lambda)v = (\lambda^2 M + \lambda G + K)v = 0, \tag{1.3}$$

with large and sparse coefficients $M = M^T$, $G = -G^T$, $K = K^T$ in $\mathbb{R}^{n\times n}$. The matrix polynomial in (1.3) is reminiscent of an even function: replacing λ with $-\lambda$, followed by taking the transpose, leads back to the original matrix polynomial. We therefore say such matrix polynomials are *T-even*. Higher order $*$-even eigenvalue problems arise in the linear quadratic optimal control problem for higher order systems of ordinary differential equations [34]. Under different nomenclature, even matrix polynomials have recently received much attention [3], [5], [33], [34].

The classical approach to investigating or numerically solving polynomial eigenvalue problems is *linearization*, in which the given polynomial (1.1) is transformed into a $kn \times kn$ matrix pencil $L(\lambda) = \lambda X + Y$ that satisfies

$$E(\lambda)L(\lambda)F(\lambda) = \begin{bmatrix} P(\lambda) & 0 \\ 0 & I_{(k-1)n} \end{bmatrix}, \tag{1.4}$$

where $E(\lambda)$ and $F(\lambda)$ are *unimodular matrix polynomials* [11]. (A matrix polynomial is *unimodular* if its determinant is a nonzero constant, independent of λ.) Standard methods for linear eigenvalue problems as in [2], [26], [30] can then be applied.

The companion forms [11] provide the standard examples of linearizations for a matrix polynomial $P(\lambda)$ as in (1.1). Let $X_1 = X_2 = \operatorname{diag}(A_k, I_n, \dots, I_n)$,

$$Y_1 = \begin{bmatrix} A_{k-1} & A_{k-2} & \dots & A_0 \\ -I_n & 0 & \dots & 0 \\ & \ddots & \ddots & \vdots \\ 0 & & -I_n & 0 \end{bmatrix}, \quad \text{and} \quad Y_2 = \begin{bmatrix} A_{k-1} & -I_n & & 0 \\ A_{k-2} & 0 & \ddots & \\ \vdots & \vdots & \ddots & -I_n \\ A_0 & 0 & \dots & 0 \end{bmatrix}.$$

Then $C_1(\lambda) = \lambda X_1 + Y_1$ and $C_2(\lambda) = \lambda X_2 + Y_2$ are, respectively, the *first* and *second companion forms* for $P(\lambda)$. Unfortunately, since these companion linearizations do not reflect the structure present in palindromic or even matrix polynomials, the

[1] Invented by the mathematician Peter Hilton in 1947 for his thesis advisor J.H.C. Whitehead. It is probable, Hilton says, that this palindrome may have been known before 1947. When Whitehead lamented its brevity, Hilton responded by crafting the palindromic masterpiece "Doc, note, I dissent. A fast never prevents a fatness. I diet on cod" [18], [19].

corresponding linearized pencil can usually only be treated with methods for general matrix pencils. In a finite precision environment, a numerical method that ignores the structure may produce physically meaningless results [35], e.g., may lose symmetries in the spectrum. Therefore, it is important to construct linearizations that reflect the structure of the given matrix polynomial, and then develop numerical methods for the corresponding linear eigenvalue problem that properly address these structures as well. The latter topic has been an important recent area of research; see, e.g., [5], [6], [7], [30], [32], and the references therein.

In this paper we show that the pencil spaces $\mathbb{L}_1(P)$ and $\mathbb{L}_2(P)$, developed in [28] by generalizing the first and second companion forms, are rich enough to include subspaces of pencils that reflect palindromic, even, or odd structure of a matrix polynomial P. Extending the notion of *Cayley transformation* to matrix polynomials, we show in section 2.2 how this transformation connects (anti-)palindromic and odd/even structures. Section 3 is devoted to the introduction and analysis of structured linearizations for the various structured matrix polynomials under consideration. The general linearization approach of [28] is summarized and then exploited to obtain the main results of this paper: identification of structured pencils in $\mathbb{L}_1(P)$, a constructive method for generating them, and necessary and sufficient conditions for these pencils to be linearizations, thereby correctly retaining information on eigenvalues and eigenvectors of the original matrix polynomial. These results are then used to identify situations when existence of structure-preserving linearizations is not guaranteed.

Finally, in section 4 we elucidate the subtitle "good vibrations from good linearizations" by discussing the impact of the theory developed in this paper on the palindromic eigenvalue problem (1.2) arising in the vibration analysis of rail tracks.

2. Basic structures, spectral properties, and Cayley transformations. In this section we formally define the structured polynomials that are studied in this paper, show how the structure of a polynomial is reflected in its spectra, and establish connections between the various classes of structured polynomials by extending the classical definition of Cayley transformations to matrix polynomials. For conciseness, the symbol $\star$ is used as an abbreviation for transpose T in the real case and for either T or conjugate transpose $*$ in the complex case.

DEFINITION 2.1. *Let $Q(\lambda) = \sum_{i=0}^{k} \lambda^i B_i$, where $B_0, \ldots, B_k \in \mathbb{F}^{m\times n}$, be a matrix polynomial of degree k, that is, $B_k \neq 0$. Then we define the* adjoint $Q^\star(\lambda)$ *and the* reversal $\operatorname{rev} Q(\lambda)$ *of $Q(\lambda)$, respectively, by*

$$Q^\star(\lambda) := \sum_{i=0}^{k} \lambda^i B_i^\star \quad \textit{and} \quad \operatorname{rev} Q(\lambda) := \lambda^k Q(1/\lambda) = \sum_{i=0}^{k} \lambda^{k-i} B_i. \tag{2.1}$$

If $\deg(Q(\lambda))$ denotes the *degree* of the matrix polynomial $Q(\lambda)$, then, in general, $\deg(\operatorname{rev} Q(\lambda)) \leq \deg(Q(\lambda))$ and $\operatorname{rev}\big(Q_1(\lambda)\cdot Q_2(\lambda)\big) = \operatorname{rev} Q_1(\lambda)\cdot \operatorname{rev} Q_2(\lambda)$, whenever the product $Q_1(\lambda)\cdot Q_2(\lambda)$ is defined. Using (2.1), the various structured matrix polynomials under consideration are defined in Table 2.1.

For a scalar polynomial $p(x)$, T-palindromic is the same as palindromic (i.e., $\operatorname{rev} p(x) = p(x)$), while $*$-palindromic is equivalent to conjugate-palindromic (i.e., $\operatorname{rev} \overline{p}(x) = p(x)$). Analogous simplifications occur in the scalar polynomial case for all other structures defined in Table 2.1.

Two matrices that play an important role in our investigation are the $k\times k$ reverse identity R_k in the context of palindromic structures, and the $k \times k$ diagonal matrix Σ_k of alternating signs in the context of even/odd structures (the subscript k will be

TABLE 2.1
Definitions of basic structures.

palindromic	$\operatorname{rev} P(\lambda) = P(\lambda)$	anti-palindromic	$\operatorname{rev} P(\lambda) = -P(\lambda)$
$\star$-palindromic	$\operatorname{rev} P^\star(\lambda) = P(\lambda)$	$\star$-anti-palindromic	$\operatorname{rev} P^\star(\lambda) = -P(\lambda)$
even	$P(-\lambda) = P(\lambda)$	odd	$P(-\lambda) = -P(\lambda)$
$\star$-even	$P^\star(-\lambda) = P(\lambda)$	$\star$-odd	$P^\star(-\lambda) = -P(\lambda)$

suppressed whenever it is clear from the context):

$$R := R_k := \begin{bmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{bmatrix}_{k\times k} \quad \text{and} \quad \Sigma := \Sigma_k := \begin{bmatrix} (-1)^{k-1} & & 0 \\ & \ddots & \\ 0 & & (-1)^0 \end{bmatrix}. \tag{2.2}$$

2.1. Spectral symmetry. A distinguishing feature of the structured matrix polynomials in Table 2.1 is the special symmetry properties of their spectra, described in the following result.

THEOREM 2.2. *Let $P(\lambda) = \sum_{i=0}^{k} \lambda^i A_i$, $A_k \neq 0$, be a regular matrix polynomial that has one of the structures listed in Table* 2.1*. Then the spectrum of $P(\lambda)$ has the pairing depicted in Table* 2.2*. Moreover, the algebraic, geometric, and partial multiplicities of the two eigenvalues in each such pair are equal. (Here, we allow $\lambda = 0$ and interpret $1/\lambda$ as the eigenvalue ∞.)*

TABLE 2.2
Spectral symmetries.

Structure of $P(\lambda)$	Eigenvalue pairing
(anti-)palindromic, T-(anti-)palindromic	$(\lambda, 1/\lambda)$
$*$-palindromic, $*$-anti-palindromic	$(\lambda, 1/\overline{\lambda})$
even, odd, T-even, T-odd	$(\lambda, -\lambda)$
$*$-even, $*$-odd	$(\lambda, -\overline{\lambda})$

Proof. We first recall some well-known facts [8], [10], [11] about the companion forms $C_1(\lambda)$ and $C_2(\lambda)$ of a regular matrix polynomial $P(\lambda)$:

- $P(\lambda)$ and $C_1(\lambda)$ have the same eigenvalues (including ∞), with the same algebraic, geometric, and partial multiplicities.
- $C_1(\lambda)$ and $C_2(\lambda)$ are always *strictly equivalent* , i.e., there exist nonsingular *constant* matrices E and F such that $C_1(\lambda) = E \cdot C_2(\lambda) \cdot F$.
- Strictly equivalent pencils have the same eigenvalues (including ∞), with the same algebraic, geometric, and partial multiplicities.

With these facts in hand, we first consider the case when $P(\lambda)$ is $\star$-palindromic or $\star$-anti-palindromic, so that $\operatorname{rev} P^\star(\lambda) = \chi_P P(\lambda)$ for $\chi_P = \pm 1$. Our strategy is to show that $C_1(\lambda)$ is strictly equivalent to $\operatorname{rev} C_1^\star(\lambda)$, from which the desired eigenvalue pairing and equality of multiplicities then follow. Using the nonsingular matrix

$$T := \begin{bmatrix} \chi_P I & \chi_P A_{k-1} & \cdots & \chi_P A_1 \\ 0 & 0 & & -I \\ \vdots & & \iddots & \\ 0 & -I & & 0 \end{bmatrix},$$

we first show that $C_1(\lambda)$ is strictly equivalent to $\operatorname{rev} C_2^\star(\lambda)$:

$$\begin{aligned}
T\cdot C_1(\lambda)\cdot(R_k\otimes I_n) &= T\cdot\left(\lambda\begin{bmatrix}0 & & & A_k\\ & & I & \\ & \iddots & & \\ I & & & 0\end{bmatrix}+\begin{bmatrix}A_0 & A_1 & \dots & A_{k-1}\\ 0 & 0 & & -I\\ \vdots & & \iddots & \\ 0 & -I & & 0\end{bmatrix}\right)\\
&= \lambda\begin{bmatrix}\chi_P A_1 & \dots & \chi_P A_{k-1} & \chi_P A_k\\ -I & & 0 & 0\\ & \ddots & & \vdots\\ 0 & & -I & 0\end{bmatrix}+\begin{bmatrix}\chi_P A_0 & & & 0\\ & I & & \\ & & \ddots & \\ 0 & & & I\end{bmatrix}\\
&= \lambda\begin{bmatrix}A_{k-1} & -I & & 0\\ \vdots & & \ddots & \\ A_1 & 0 & & -I\\ A_0 & 0 & \dots & 0\end{bmatrix}^\star+\begin{bmatrix}A_k & & & 0\\ & I & & \\ & & \ddots & \\ 0 & & & I\end{bmatrix}^\star\\
&= \operatorname{rev} C_2^\star(\lambda).
\end{aligned}$$

But $\operatorname{rev} C_2^\star(\lambda)$ is always strictly equivalent to $\operatorname{rev} C_1^\star(\lambda)$, since $C_1(\lambda)$ and $C_2(\lambda)$ are. This completes the proof for this case.

For the case of palindromic or anti-palindromic matrix polynomials, i.e., polynomials $P(\lambda)$ satisfying $\operatorname{rev} P(\lambda)=\chi_P P(\lambda)$, an analogous computation shows that

$$T\cdot C_1(\lambda)\cdot(R_k\otimes I_n)=\operatorname{rev} C_1(\lambda),$$

i.e., $C_1(\lambda)$ is strictly equivalent to $\operatorname{rev} C_1(\lambda)$, which again implies the desired eigenvalue pairing and equality of multiplicities.

Next assume that $P(\lambda)$ is $\star$-even or $\star$-odd, so $P^\star(-\lambda)=\varepsilon_P P(\lambda)$ for $\varepsilon_P=\pm 1$. We show that $C_1(\lambda)$ is strictly equivalent to $C_1^\star(-\lambda)$, from which the desired pairing of eigenvalues and equality of multiplicities follows. We begin by observing that $C_1(\lambda)$ is strictly equivalent to $C_2^\star(-\lambda)$:

$$\begin{aligned}
&\big(\operatorname{diag}(\varepsilon_P,-\Sigma_{k-1})\otimes I_n\big)\cdot C_1(\lambda)\cdot(\Sigma_k\otimes I_n)\\
&=\lambda\begin{bmatrix}\varepsilon_P(-1)^{k-1}A_k & & & 0\\ & -I & & \\ & & \ddots & \\ 0 & & & -I\end{bmatrix}+\begin{bmatrix}\varepsilon_P(-1)^{k-1}A_{k-1} & \dots & \varepsilon_P(-1)^1A_1 & \varepsilon_P A_0\\ -I & & 0 & 0\\ & \ddots & & \vdots\\ 0 & & -I & 0\end{bmatrix}\\
&=-\lambda\begin{bmatrix}A_k & & & 0\\ & I & & \\ & & \ddots & \\ 0 & & & I\end{bmatrix}^\star+\begin{bmatrix}A_{k-1} & -I & & 0\\ \vdots & & \ddots & \\ A_1 & 0 & & -I\\ A_0 & 0 & \dots & 0\end{bmatrix}^\star=C_2^\star(-\lambda).
\end{aligned}$$

The strict equivalence of $C_2^\star(-\lambda)$ and $C_1^\star(-\lambda)$ now follows from that of $C_2(\lambda)$ and $C_1(\lambda)$, and the proof for this case is complete.

For even or odd polynomials, that is, when $P(-\lambda)=\varepsilon_P P(\lambda)$, an analogous computation

$$\big(\operatorname{diag}(\varepsilon_P,-\Sigma_{k-1})\otimes I_n\big)\cdot C_1(\lambda)\cdot(\Sigma_k\otimes I_n)=C_1(-\lambda)$$

shows that $C_1(\lambda)$ is strictly equivalent to $C_1(-\lambda)$, which implies the desired eigenvalue pairing and equality of multiplicities. ∎

If the coefficient matrices of P are real, then the eigenvalues of a $\star$-even or $\star$-odd matrix polynomial occur in quadruples $(\lambda, \overline{\lambda}, -\lambda, -\overline{\lambda})$. This property has sometimes been referred to as "Hamiltonian spectral symmetry" because the eigenvalues of real Hamiltonian matrices have such symmetry [31], [34]. However, this feature is not confined to Hamiltonian matrices, but is shared by matrices in Lie algebras associated with any real scalar product. Similarly, the eigenvalues of real $\star$-palindromic and $\star$-anti-palindromic matrix polynomials occur in quadruples $(\lambda, \overline{\lambda}, 1/\lambda, 1/\overline{\lambda})$, a property sometimes referred to as "symplectic spectral symmetry" because real symplectic matrices exhibit this behavior. But once again, this type of eigenvalue symmetry is an instance of a more general phenomenon associated with matrices in the Lie group of any real scalar product, such as the real pseudo-orthogonal (Lorentz) groups. See [1], [7], [24], [31] for detailed coverage of Hamiltonian and symplectic matrices, and see [12], [29] for properties of matrices in the Lie algebra or Lie group of more general scalar products.

Remark 2.3. In Definition 2.1 we could have defined the adjoint of an $n \times n$ matrix polynomial with respect to the adjoint of a more general scalar product, rather than restricting $\star$ to being just transpose or conjugate transpose. For example, with any nonsingular matrix M we can define a bilinear scalar product $\langle x, y\rangle := x^T My$ and denote the adjoint of a matrix $A \in \mathbb{F}^{n\times n}$ with respect to this scalar product by $A^\star = M^{-1}A^T M$. (For a sesquilinear scalar product $\langle x, y\rangle := x^* My$, its corresponding adjoint is $A^\star = M^{-1}A^* M$.) Then for an $n \times n$ matrix polynomial $P(\lambda)$, the definition of the corresponding adjoint $P^\star(\lambda)$ is formally identical to Definition 2.1; the structures in Table 2.1 also make sense as written with $\star$ denoting the adjoint of a general scalar product. Well-known examples of this are the skew-Hamiltonian/Hamiltonian pencils [33], which are $\star$-odd with respect to the symplectic form defined by $M = J = \left[\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}\right]$.

However, if the matrix M defining a bilinear scalar product satisfies $M^T = \varepsilon M$ for $\varepsilon = \pm 1$ (or $M^* = \varepsilon M$, $|\varepsilon| = 1$, $\varepsilon \in \mathbb{C}$, in the sesquilinear case), then not much is gained by this apparent extra generality. Note this includes all the standard examples, which are either symmetric or skew-symmetric bilinear forms or Hermitian sesquilinear forms. In the bilinear case, we have

$$\begin{aligned} P(\lambda) \text{ is } \star\text{-palindromic} &\Leftrightarrow \operatorname{rev} P^\star(\lambda) = \operatorname{rev}\left(M^{-1}P^T(\lambda)M\right) = P(\lambda) \\ &\Leftrightarrow \operatorname{rev}\left(MP(\lambda)\right)^T = \operatorname{rev}\left(P^T(\lambda)M^T\right) = \varepsilon MP(\lambda) \\ &\Leftrightarrow MP(\lambda) \text{ is } T\text{-palindromic or } T\text{-anti-palindromic,} \end{aligned}$$

depending on the sign of ε. A similar argument shows that $\star$-evenness or $\star$-oddness of $P(\lambda)$ is equivalent to T-evenness or T-oddness of $MP(\lambda)$. Analogous results also hold in the sesquilinear case when $M^* = \varepsilon M$. Thus for any of the standard scalar products with adjoint $\star$, the $\star$-structures in Table 2.1 can be reduced to either the $\star = T$ or $\star = *$ case; in particular this implies that the eigenvalue pairing results of Theorem 2.2 extend to these more general $\star$-structures. Note this reduction shows that the skew-Hamiltonian/Hamiltonian pencils mentioned above are equivalent to T-even or $*$-even pencils.

2.2. Cayley transformations of matrix polynomials. It is well known that the *Cayley transformation* and its generalization to matrix pencils [24], [32] relates Hamiltonian structure to symplectic structure for both matrices and pencils. By extending the classical definition of this transformation to matrix polynomials, we now develop analogous relationships between (anti-)palindromic and odd/even matrix polynomials, and their $\star$-variants.

Our choice of definition is motivated by the following observation: the only Möbius transformations of the complex plane that map reciprocal pairs $(\mu, 1/\mu)$ to plus/minus pairs $(\lambda, -\lambda)$ are $\alpha\big(\frac{\mu-1}{\mu+1}\big)$ and $\beta\big(\frac{1+\mu}{1-\mu}\big)$, where α, $\beta \in \mathbb{C}$ are nonzero constants. When $\alpha = \beta = 1$, these transformations also map conjugate reciprocal pairs $(\mu, 1/\overline{\mu})$ to conjugate plus/minus pairs $(\lambda, -\overline{\lambda})$. Combining this with Theorem 2.2, we see that the Möbius transformations $\frac{\mu-1}{\mu+1}$, $\frac{1+\mu}{1-\mu}$ translate the spectral symmetries of (anti-)palindromic matrix polynomials and their $\star$-variants to those of odd/even matrix polynomials and their $\star$-variants. Consequently, it is reasonable to anticipate that Cayley transformations modeled on these particular Möbius transformations might have an analogous effect on structure at the level of matrix polynomials. These observations therefore lead us to adopt the following definition as the natural extension, given our context, of the Cayley transformation to matrix polynomials.

DEFINITION 2.4. *Let $P(\lambda)$ be a matrix polynomial of degree k as in* (1.1)*. Then the* Cayley transformation of $P(\lambda)$ with pole at -1 *or* $+1$*, respectively, is the matrix polynomial*

$$\text{(2.3)}\quad \mathcal{C}_{-1}(P)(\mu) := (\mu+1)^k P\left(\frac{\mu-1}{\mu+1}\right), \quad \textit{resp.}, \quad \mathcal{C}_{+1}(P)(\mu) := (1-\mu)^k P\left(\frac{1+\mu}{1-\mu}\right).$$

When viewed as maps on the space of $n \times n$ matrix polynomials of degree $k \geq 1$, the Cayley transformations in (2.3) can be shown by a direct calculation to be inverses of each other, up to a scaling factor.

PROPOSITION 2.5. *For any $n \times n$ matrix polynomial P of degree $k \geq 1$, we have $\mathcal{C}_{+1}(\mathcal{C}_{-1}(P)) = \mathcal{C}_{-1}(\mathcal{C}_{+1}(P)) = 2^k \cdot P$.*

The next lemma gives some straightforward observations that are helpful in relating the structure in a matrix polynomial to that in its Cayley transformations.

LEMMA 2.6. *Let P be a matrix polynomial of degree $k \geq 1$. Then*

$$\text{(2.4)}\qquad \big(\mathcal{C}_{-1}(P)\big)^\star(\mu) = \mathcal{C}_{-1}(P^\star)(\mu), \qquad \big(\mathcal{C}_{+1}(P)\big)^\star(\mu) = \mathcal{C}_{+1}(P^\star)(\mu),$$

$$\text{(2.5a)}\qquad \operatorname{rev}\big(\mathcal{C}_{-1}(P)\big)^\star(\mu) = (\mu+1)^k P^\star\left(-\frac{\mu-1}{\mu+1}\right), \quad \mu \neq -1,$$

$$\text{(2.5b)}\qquad \operatorname{rev}\big(\mathcal{C}_{+1}(P)\big)^\star(\mu) = (-1)^k(1-\mu)^k P^\star\left(-\frac{1+\mu}{1-\mu}\right), \quad \mu \neq 1.$$

Proof. The proof of (2.4) is straightforward. We only prove (2.5b); the proof of (2.5a) is similar. Since $\mathcal{C}_{+1}(P)$, and hence $\mathcal{C}_{+1}(P)^\star$, are matrix polynomials of degree k,

$$\begin{aligned}\operatorname{rev}\big(\mathcal{C}_{+1}(P)\big)^\star(\mu) &= \mu^k\big(\mathcal{C}_{+1}(P)\big)^\star\left(\frac{1}{\mu}\right) = \mu^k\mathcal{C}_{+1}(P^\star)\left(\frac{1}{\mu}\right) && \text{by (2.4), (2.1)}\\ &= \mu^k(1-1/\mu)^k P^\star\left(\frac{1+1/\mu}{1-1/\mu}\right) && \text{by (2.3)}\end{aligned}$$

and (2.5b) is now immediate. ∎

THEOREM 2.7. *Let $P(\lambda)$ be a matrix polynomial of degree $k \geq 1$. Then the correspondence between structure in $P(\lambda)$ and in its Cayley transformations is as stated in Table* 2.3*.*

Proof. Since the proofs of the equivalences are similar, we establish only one of them. We show that $P(\lambda)$ is $\star$-even if and only if $\mathcal{C}_{+1}(P)(\mu)$ is $\star$-palindromic when k is even and $\star$-anti-palindromic when k is odd. Now $P(\lambda)$ being $\star$-even is equivalent,

TABLE 2.3
Cayley transformations of structured matrix polynomials.

	$\mathcal{C}_{-1}(P)(\mu)$		$\mathcal{C}_{+1}(P)(\mu)$	
$P(\lambda)$	k even	k odd	k even	k odd
palindromic $\star$-palindromic	even $\star$-even	odd $\star$-odd	even $\star$-even	
anti-palindromic $\star$-anti-palindromic	odd $\star$-odd	even $\star$-even	odd $\star$-odd	
even $\star$-even	palindromic $\star$-palindromic		palindromic $\star$-palindromic	anti-palindromic $\star$-anti-palindromic
odd $\star$-odd	anti-palindromic $\star$-anti-palindromic		anti-palindromic $\star$-anti-palindromic	palindromic $\star$-palindromic

by definition, to $P^\star(-\lambda) = P(\lambda)$ for all λ. Setting $\lambda = \frac{1+\mu}{1-\mu}$ and multiplying by $(1-\mu)^k$ yields

$$\begin{aligned} P(\lambda) \text{ is } \star\text{-even} &\iff (1-\mu)^k P^\star\left(-\frac{1+\mu}{1-\mu}\right) = (1-\mu)^k P\left(\frac{1+\mu}{1-\mu}\right) \quad \text{for all } \mu \neq 1 \\ &\iff (-1)^k \mathrm{rev}\,(\mathcal{C}_{+1}(P))^\star(\mu) = \mathcal{C}_{+1}(P)(\mu) \quad \text{by Lemma 2.6,} \end{aligned}$$

from which the desired result follows. ∎

Observe that the results in Table 2.3 are consistent with $\mathcal{C}_{-1}(P)$ and $\mathcal{C}_{+1}(P)$ being essentially inverses of each other (Proposition 2.5).

Theorem 2.7 establishes a relationship between $\star$-(anti-)palindromic and $\star$-even/odd matrix polynomials via the Cayley transformation. Since $\star$-even/odd matrix polynomials can be interpreted as generalizations of Hamiltonian matrices [33], [34], and since it is well known that Hamiltonian matrices and symplectic matrices are related via the Cayley transformation [31], $\star$-(anti-)palindromic matrix polynomials can be thought of as generalizations of symplectic matrices.

3. Structured linearizations. As sources of structured linearizations for the structured polynomials listed in Table 2.1, we consider the vector spaces $\mathbb{L}_1(P)$ and $\mathbb{L}_2(P)$, introduced in [28]. We establish the existence of structured pencils in these spaces, show how they can be explicitly constructed, and give necessary and sufficient conditions for them to be linearizations of the given matrix polynomial P.

3.1. Vector spaces of potential linearizations. The vector spaces $\mathbb{L}_1(P)$ and $\mathbb{L}_2(P)$ consist of pencils that generalize the first and second companion forms $C_1(\lambda)$ and $C_2(\lambda)$ of $P(\lambda)$, respectively:

$$\mathbb{L}_1(P) := \left\{ L(\lambda) = \lambda X + Y \ : \ L(\lambda)\cdot(\Lambda \otimes I_n) = v \otimes P(\lambda),\, v \in \mathbb{F}^k \right\}, \tag{3.1}$$

$$\mathbb{L}_2(P) := \left\{ L(\lambda) = \lambda X + Y \ : \ \left(\Lambda^T \otimes I_n\right)\cdot L(\lambda) = w^T \otimes P(\lambda),\, w \in \mathbb{F}^k \right\}, \tag{3.2}$$

where $\Lambda = \begin{bmatrix} \lambda^{k-1} & \lambda^{k-2} & \dots & \lambda & 1 \end{bmatrix}^T$, and $\otimes$ denotes the Kronecker product. A direct calculation shows that

$$C_1(\lambda)\cdot(\Lambda \otimes I_n) = e_1 \otimes P(\lambda) \quad \text{and} \quad \left(\Lambda^T \otimes I_n\right)\cdot C_2(\lambda) = e_1^T \otimes P(\lambda),$$

so $C_1(\lambda) \in \mathbb{L}_1(P)$ and $C_2(\lambda) \in \mathbb{L}_2(P)$ for any $P(\lambda)$. The vector v in (3.1) is called the *right ansatz vector* of $L(\lambda) \in \mathbb{L}_1(P)$ because $L(\lambda)$ is multiplied on the right by $\Lambda \otimes I_n$ to give $v \otimes P(\lambda)$. Analogously, the vector w in (3.2) is called the *left ansatz vector* of $L(\lambda) \in \mathbb{L}_2(P)$.

The pencil spaces $\mathbb{L}_j(P)$ were designed with the aim of providing an arena of potential linearizations that is fertile enough to contain those that reflect additional structures in P, but small enough that these linearizations still share salient features of the companion forms $C_j(\lambda)$. First, when $P(\lambda)$ is regular, the mild hypothesis of a pencil in $\mathbb{L}_j(P)$ being regular is sufficient to guarantee that it is indeed a linearization for P. In fact, as shown in [28], regularity makes these pencils strong linearizations for $P(\lambda)$, i.e., $\operatorname{rev} L(\lambda)$ is also a linearization for $\operatorname{rev} P(\lambda)$. This ensures that the Jordan structures of both the finite and infinite eigenvalues of P are always faithfully reflected in L, just as is done by the companion forms. Without this extra property of being a strong linearization, any Jordan structure compatible with the algebraic multiplicity of the infinite eigenvalue of $P(\lambda)$ can be realized by some linearization [23]. Second, eigenvectors of $P(\lambda)$ are easily recoverable from those of $L(\lambda)$. Indeed, the definition of $\mathbb{L}_1(P)$ implies that $L(\lambda) \cdot (\Lambda \otimes x) = v \otimes (P(\lambda)x)$ for all $x \in \mathbb{F}^n$. Thus, whenever x is a right eigenvector of $P(\lambda)$ associated with an eigenvalue λ, then $\Lambda \otimes x$ is a right eigenvector of $L(\lambda)$ associated with λ. Similar observations hold for $L(\lambda) \in \mathbb{L}_2(P)$ and left eigenvectors. Finally, when $P(\lambda)$ is regular, almost all pencils in $\mathbb{L}_j(\lambda)$ are regular, and thus strong linearizations for $P(\lambda)$—the ones that are not from a closed nowhere dense set of measure zero [28].

3.1.1. Shifted sums. The *column-shifted sum* and *row-shifted sum* are convenient tools that readily allow one to construct pencils in $\mathbb{L}_1(P)$ and $\mathbb{L}_2(P)$, respectively. They also enable one to easily test when a given pencil is in $\mathbb{L}_j(P)$.

DEFINITION 3.1 (shifted sums). *Let $X = (X_{ij})$ and $Y = (Y_{ij})$ be block $k \times k$ matrices in $\mathbb{F}^{kn \times kn}$ with blocks $X_{ij}, Y_{ij} \in \mathbb{F}^{n \times n}$. Then the* column shifted sum $X \boxplus\!\!\rightarrow Y$ *and* row shifted sum $X \underset{\downarrow}{\boxplus} Y$ *are defined to be*

$$X \boxplus\!\!\rightarrow Y := \begin{bmatrix} X_{11} & \dots & X_{1k} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ X_{k1} & \dots & X_{kk} & 0 \end{bmatrix} + \begin{bmatrix} 0 & Y_{11} & \dots & Y_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & Y_{k1} & \dots & Y_{kk} \end{bmatrix} \in \mathbb{F}^{kn \times k(n+1)},$$

$$X \underset{\downarrow}{\boxplus} Y := \begin{bmatrix} X_{11} & \cdots & X_{1m} \\ \vdots & \ddots & \vdots \\ X_{m1} & \dots & X_{mm} \\ 0 & \dots & 0 \end{bmatrix} + \begin{bmatrix} 0 & \dots & 0 \\ Y_{11} & \dots & Y_{1m} \\ \vdots & \ddots & \vdots \\ Y_{m1} & \dots & Y_{mm} \end{bmatrix} \in \mathbb{F}^{k(n+1) \times kn}.$$

With $P(\lambda) = \sum_{i=0}^{k} \lambda^i A_i$, and $L(\lambda) = \lambda X + Y$, a straightforward calculation with the shifted sums now reveals the equivalences

$$\begin{array}{c} L(\lambda) \in \mathbb{L}_1(P) \\ \text{with right ansatz vector } v \end{array} \iff X \boxplus\!\!\rightarrow Y = v \otimes [A_k \; A_{k-1} \; \dots \; A_0], \tag{3.3}$$

$$\begin{array}{c} L(\lambda) \in \mathbb{L}_2(P) \\ \text{with left ansatz vector } w \end{array} \iff X \underset{\downarrow}{\boxplus} Y = w^T \otimes \begin{bmatrix} A_k \\ \vdots \\ A_0 \end{bmatrix}. \tag{3.4}$$

3.2. Building T-palindromic pencils in $\mathbb{L}_1(P)$. For the moment, let us focus our attention on $\mathbb{L}_1(P)$ and try to construct a T-palindromic pencil in $\mathbb{L}_1(P)$ for a

matrix polynomial $P(\lambda)$ that is T-palindromic. We begin with the simplest nontrivial example.

Example 3.2. Consider the T-palindromic matrix polynomial $\lambda^2 A + \lambda B + A^T$, where $B = B^T$ and $A \neq 0$. We try to construct a T-palindromic pencil $L(\lambda) \in \mathbb{L}_1(P)$ with a nonzero right ansatz vector $v = [\, v_1, v_2 \,]^T \in \mathbb{F}^2$. This means that $L(\lambda)$ must be of the form

$$L(\lambda) = \lambda Z + Z^T =: \lambda \begin{bmatrix} D & E \\ F & G \end{bmatrix} + \begin{bmatrix} D^T & F^T \\ E^T & G^T \end{bmatrix}, \quad D, E, F, G \in \mathbb{F}^{n \times n}.$$

Since $L(\lambda) \in \mathbb{L}_1(P)$, the equivalence given by (3.3) implies that

$$Z \boxplus Z^T = \begin{bmatrix} D & E + D^T & F^T \\ F & G + E^T & G^T \end{bmatrix} = \begin{bmatrix} v_1 A & v_1 B & v_1 A^T \\ v_2 A & v_2 B & v_2 A^T \end{bmatrix}.$$

Equating corresponding blocks in the first and last columns, we obtain $D = v_1 A$, $F = v_2 A = v_1 A$, and $G = v_2 A$. This forces $v_1 = v_2$, since $A \neq 0$ by assumption. From either block of the middle column, we see that $E = v_1(B - A^T)$; with this choice for E all the equations are consistent, thus yielding

$$L(\lambda) = \lambda Z + Z^T = v_1 \left(\lambda \begin{bmatrix} A & B - A^T \\ A & A \end{bmatrix} + \begin{bmatrix} A^T & A^T \\ B - A & A^T \end{bmatrix} \right). \tag{3.5}$$

This example illustrates three important properties: (1) the choice of right ansatz vectors v for which $L(\lambda) \in \mathbb{L}_1(P)$ is T-palindromic is restricted; (2) once one of these restricted right ansatz vectors v is chosen, a T-palindromic pencil $L(\lambda) \in \mathbb{L}_1(P)$ is uniquely determined; (3) interchanging the first and second block rows of $L(\lambda)$, i.e., premultiplying by $R_2 \otimes I$, yields the pencil

$$(R_2 \otimes I) L(\lambda) = v_1 \left(\lambda \begin{bmatrix} A & A \\ A & B - A^T \end{bmatrix} + \begin{bmatrix} B - A & A^T \\ A^T & A^T \end{bmatrix} \right),$$

which the column- and row-shifted sums easily confirm to be a pencil in the *double ansatz space* $\mathbb{DL}(P) := \mathbb{L}_1(P) \cap \mathbb{L}_2(P)$ with left and right ansatz vector $v = [v_1, v_1]^T$. These three observations turn out to be true in general for T-palindromic matrix polynomials P and T-palindromic pencils in $\mathbb{L}_1(P)$.

THEOREM 3.3. *Let $P(\lambda)$ be a T-palindromic matrix polynomial and suppose $L(\lambda) \in \mathbb{L}_1(P)$ with right ansatz vector v. Then the pencil $L(\lambda)$ is T-palindromic if and only if $Rv = v$ and $(R \otimes I)L(\lambda) \in \mathbb{DL}(P)$ with right and left ansatz vector Rv, where R is the reverse identity as in* (2.2). *Moreover, for* any *$v \in \mathbb{F}^k$ satisfying $Rv = v$ there exists a unique pencil $L(\lambda) \in \mathbb{L}_1(P)$ with right ansatz vector v and T-palindromic structure.*

The proof of this theorem is deferred to the next section, where it is subsumed under the even more general result stated in Theorem 3.5.

The double ansatz space $\mathbb{DL}(P)$ was introduced in [28] as a natural space in which to look for pencils that enjoy both the right and left eigenvector recovery properties. This feature was successfully exploited in [14] to find linearizations with optimally conditioned eigenvalues. Now Example 3.2 suggests that $\mathbb{DL}(P)$ could also play an important role in the search for structured linearizations.

3.3. Existence of structured pencils in $\mathbb{L}_1(P)$. For a $\star$-(anti-)palindromic or $\star$-even/odd polynomial it is natural to seek a linearization with the same structure

as P. From the point of view of numerical analysis, however, one of the most important reasons for using a structure-preserving method is to preserve spectral symmetries. But we see in Table 2.2 that for each structure under consideration there is also an "anti" version of that structure with the same spectral symmetry. Thus it makes sense to try to linearize a structured polynomial with an "anti-structured" pencil as well as with a structured one; so in this section we also characterize the pencils in $\mathbb{L}_1(P)$ having the "anti-structure" of P.

Before turning to the main results of this section, we draw the reader's attention to two key properties of $\mathbb{DL}(P)$ that will be systematically used in their proofs. Recall that the left and right ansatz vectors of the double ansatz pencil $(R_2 \otimes I)L(\lambda)$ in Example 3.2 coincide. This is, in fact, a property shared by all pencils in $\mathbb{DL}(P)$, thus leading to the notion of a single *ansatz vector* instead of separate left/right ansatz vectors for these pencils. Furthermore, every pencil in $\mathbb{DL}(P)$ is uniquely determined by its ansatz vector.

THEOREM 3.4 (see [13], [28]). *Let $P(\lambda) = \sum_{i=0}^{k} \lambda^i A_i$ be a (not necessarily regular) matrix polynomial with coefficients in $\mathbb{F}^{n\times n}$ and $A_k \neq 0$. Then for vectors $v, w \in \mathbb{F}^k$ there exists a $kn \times kn$ matrix pencil $L(\lambda) \in \mathbb{DL}(P)$ with right ansatz vector w and left ansatz vector v if and only if $v = w$. Moreover, the pencil $L(\lambda) \in \mathbb{DL}(P)$ is uniquely determined by v.*

We now extend the result of Theorem 3.3 to $\star$-(anti-)palindromic structures, showing that there is only a restricted class of admissible right ansatz vectors v that can support a structured or "anti-structured" pencil in $\mathbb{L}_1(P)$. In each case the restrictions on the vector v can be concisely described using the reverse identity $R = R_k$ as defined in (2.2).

THEOREM 3.5. *Suppose the matrix polynomial $P(\lambda)$ is $\star$-palindromic or $\star$-anti-palindromic. Then for pencils $L(\lambda) \in \mathbb{L}_1(P)$ with right ansatz vector v, conditions* (i) *and* (ii) *in Table* 3.1 *are equivalent. Moreover, for* any *$v \in \mathbb{F}^k$ satisfying one of the admissibility conditions for v in* (ii)*, there exists a unique pencil $L(\lambda) \in \mathbb{L}_1(P)$ with right ansatz vector v and the corresponding structure in* (i).

TABLE 3.1

Structure of $P(\lambda)$	Equivalent conditions	
	(i) $L(\lambda)$ is	(ii) $(R \otimes I)L(\lambda) \in \mathbb{DL}(P)$ with ansatz vector Rv and
T-palindromic	T-palindromic	$Rv = v$
	T-anti-palindromic	$Rv = -v$
T-anti-palindromic	T-palindromic	$Rv = -v$
	T-anti-palindromic	$Rv = v$
$*$-palindromic	$*$-palindromic	$Rv = \overline{v}$
	$*$-anti-palindromic	$Rv = -\overline{v}$
$*$-anti-palindromic	$*$-palindromic	$Rv = -\overline{v}$
	$*$-anti-palindromic	$Rv = \overline{v}$

Proof. We consider all eight cases simultaneously. Let $P(\lambda)$ be $\star$-palindromic or $\star$-anti-palindromic, so that $\operatorname{rev} P^\star(\lambda) = \chi_{_P} P(\lambda)$ for $\chi_{_P} = \pm 1$.

(i) $\Rightarrow$ (ii): By (i), $\operatorname{rev} L^\star(\lambda) = \chi_{_L} L(\lambda)$ for $\chi_{_L} = \pm 1$. Since $L(\lambda) \in \mathbb{L}_1(P)$, we have

$$L(\lambda)(\Lambda \otimes I) = v \otimes P(\lambda). \tag{3.6}$$

Taking the reversal of both sides of (3.6), and noting that $R\Lambda = \operatorname{rev} \Lambda$, we have

$$\operatorname{rev} L(\lambda)(R \otimes I)(\Lambda \otimes I) = \operatorname{rev} L(\lambda)\big((\operatorname{rev} \Lambda) \otimes I\big) = v \otimes \operatorname{rev} P(\lambda).$$

Now applying the adjoint $\star$ to both sides, we obtain

$$(\Lambda^\star \otimes I)(R \otimes I) \operatorname{rev} L^\star(\lambda^\star) = v^\star \otimes \operatorname{rev} P^\star(\lambda^\star)\,,$$

or equivalently

$$\big(\Lambda^\star \otimes I\big)(R \otimes I)L(\lambda^\star) = (\chi_{_P}\chi_{_L}\, v^\star) \otimes P(\lambda^\star)\,, \tag{3.7}$$

since $L(\lambda)$ and $P(\lambda)$ are either $\star$-palindromic or $\star$-anti-palindromic. Then using the fact that (3.7) is an identity, we replace $\lambda^\star$ with λ to obtain

$$(\Lambda^T \otimes I)(R \otimes I)L(\lambda) = (\chi_{_P}\chi_{_L}\, v^\star) \otimes P(\lambda)\,,$$

thus showing $(R \otimes I)L(\lambda)$ to be in $\mathbb{L}_2(P)$ with left ansatz vector $w = \chi_{_P}\chi_{_L}\,(v^\star)^T$. On the other hand, multiplying (3.6) on the left by $R \otimes I$ yields

$$(R \otimes I)L(\lambda)(\Lambda \otimes I) = (Rv) \otimes P(\lambda)\,,$$

so $(R \otimes I)L(\lambda)$ is also in $\mathbb{L}_1(P)$ with right ansatz vector Rv. Thus $(R \otimes I)L(\lambda)$ is in $\mathbb{DL}(P) = \mathbb{L}_1(P) \cap \mathbb{L}_2(P)$, and from Theorem 3.4 the equality of right and left ansatz vectors implies that $Rv = \chi_{_P}\chi_{_L}\,(v^\star)^T$. All eight variants of condition (ii) now follow by noting that $(v^*)^T = \overline{v}$ and $(v^T)^T = v$.

(ii) $\Rightarrow$ (i): Since $(R \otimes I)L(\lambda)$ is in $\mathbb{DL}(P)$ with ansatz vector Rv, we have

$$(R \otimes I)L(\lambda)(\Lambda \otimes I) \;=\; (Rv) \otimes P(\lambda)\,, \tag{3.8}$$

$$\big((\Lambda^T R) \otimes I\big)L(\lambda) \;=\; (\Lambda^T \otimes I)(R \otimes I)L(\lambda) \;=\; (Rv)^T \otimes P(\lambda)\,. \tag{3.9}$$

Applying the adjoint $\star$ to both ends of (3.9) gives

$$L^\star(\lambda^\star)\big((R(\Lambda^T)^\star) \otimes I\big) = R(v^T)^\star \otimes P^\star(\lambda^\star),$$

or equivalently

$$L^\star(\lambda)\big((R\Lambda) \otimes I\big) = R(v^T)^\star \otimes P^\star(\lambda). \tag{3.10}$$

Note that all cases of condition (ii) may be expressed in the form $R(v^T)^\star = \varepsilon\chi_{_P}\, v$, where $\varepsilon = \pm 1$. Then taking the reversal of both sides in (3.10) and using $R\Lambda = \operatorname{rev} \Lambda$, we obtain

$$\operatorname{rev} L^\star(\lambda)(\Lambda \otimes I) = (\varepsilon\chi_{_P}\, v) \otimes \operatorname{rev} P^\star(\lambda) = (\varepsilon v) \otimes P(\lambda)\,,$$

and after multiplying by $\varepsilon(R \otimes I)$,

$$\varepsilon(R \otimes I) \operatorname{rev} L^\star(\lambda)(\Lambda \otimes I) = (Rv) \otimes P(\lambda).$$

Thus we see that the pencil $\varepsilon(R \otimes I)\operatorname{rev} L^\star(\lambda)$ is in $\mathbb{L}_1(P)$ with right ansatz vector Rv. Starting over again from identity (3.8) and taking the adjoint $\star$ of both sides, we obtain by analogous reasoning that

$$\begin{aligned}
& (R \otimes I)L(\lambda)(\Lambda \otimes I) = (Rv) \otimes P(\lambda) \\
\iff\ & (\Lambda^T \otimes I)L^\star(\lambda)(R \otimes I) = (v^\star R) \otimes P^\star(\lambda) = \big(v^\star \otimes P^\star(\lambda)\big)(R \otimes I) \\
\iff\ & (\Lambda^T \otimes I)L^\star(\lambda) = v^\star \otimes P^\star(\lambda) \\
\iff\ & (\operatorname{rev}\Lambda^T \otimes I)\operatorname{rev} L^\star(\lambda) = v^\star \otimes \operatorname{rev} P^\star(\lambda) \\
\iff\ & (\Lambda^T R \otimes I)\operatorname{rev} L^\star(\lambda) = (\varepsilon\chi_{_P} Rv)^T \otimes \operatorname{rev} P^\star(\lambda) = (\varepsilon Rv)^T \otimes P(\lambda) \\
\iff\ & (\Lambda^T \otimes I)\left(\varepsilon(R \otimes I)\operatorname{rev} L^\star(\lambda)\right) = (Rv)^T \otimes P(\lambda)\,.
\end{aligned}$$

Thus the pencil $\varepsilon(R \otimes I)\operatorname{rev} L^\star(\lambda)$ is also in $\mathbb{L}_2(P)$ with left ansatz vector Rv, and hence in $\mathbb{DL}(P)$ with ansatz vector Rv. But $(R \otimes I)L(\lambda) \in \mathbb{DL}(P)$ with exactly the same ansatz vector Rv, and so the uniqueness property of Theorem 3.4 implies that

$$\varepsilon(R \otimes I)\operatorname{rev} L^\star(\lambda) = (R \otimes I)L(\lambda)\,,$$

or equivalently $\varepsilon \operatorname{rev} L^\star(\lambda) = L(\lambda)$. Hence $L(\lambda)$ is $\star$-palindromic or $\star$-anti-palindromic, depending on the parameter ε, which implies all the variants of condition (i) in Table 3.2.

Finally, the existence and uniqueness of a structured pencil $L(\lambda)$ corresponding to any admissible right ansatz vector v follow directly from the existence and uniqueness in Theorem 3.4 of the $\mathbb{DL}(P)$-pencil $(R \otimes I)L(\lambda)$ for the ansatz vector Rv. ∎

We next present the analogue of Theorem 3.5 for $\star$-even and $\star$-odd polynomials. Here $\Sigma = \Sigma_k$ as defined in (2.2) helps to concisely describe the restriction on the ansatz vector v.

THEOREM 3.6. *Suppose the matrix polynomial $P(\lambda)$ is $\star$-even or $\star$-odd. Then for pencils $L(\lambda) \in \mathbb{L}_1(P)$ with right ansatz vector v, conditions* (i) *and* (ii) *in Table* 3.2 *are equivalent. Moreover, for* any *$v \in \mathbb{F}^k$ satisfying one of the admissibility conditions for v in* (ii)*, there exists a unique pencil $L(\lambda) \in \mathbb{L}_1(P)$ with right ansatz vector v and the corresponding structure in* (i).

TABLE 3.2

Structure of $P(\lambda)$	Equivalent conditions	
	(i) $L(\lambda)$ is	(ii) $(\Sigma \otimes I)L(\lambda) \in \mathbb{DL}(P)$ with ansatz vector Σv and
T-even	T-even	$\Sigma v = v$
	T-odd	$\Sigma v = -v$
T-odd	T-even	$\Sigma v = -v$
	T-odd	$\Sigma v = v$
$*$-even	$*$-even	$\Sigma v = \overline{v}$
	$*$-odd	$\Sigma v = -\overline{v}$
$*$-odd	$*$-even	$\Sigma v = -\overline{v}$
	$*$-odd	$\Sigma v = \overline{v}$

Proof. Follow the strategy, mutatis mutandis, of the proof of Theorem 3.5: replace multiplications by $R \otimes I$ with multiplications by $\Sigma \otimes I$, and replace taking reversals with the substitution $\lambda \mapsto -\lambda$. Observe that for the vector Λ, this substitution is equivalent to premultiplication by Σ, since $\Sigma\Lambda = [(-\lambda)^{k-1}, \ldots, -\lambda, 1]^T$. ∎

Thus we see that the ansatz vectors of structured pencils closely reflect the structure of the pencil itself. This pleasing fact influences both the existence and the construction of structured linearizations, as we will see in the following sections.

3.4. Construction of structured pencils. As we have seen in Theorems 3.5 and 3.6, pencils in $\mathbb{L}_1(P)$ with one of the $\star$-structures listed in Table 2.1 are strongly related to pencils in $\mathbb{DL}(P)$. This observation leads to the following procedure for the construction of potential structured linearizations:

(1) Choose a right ansatz vector $v \in \mathbb{F}^k$ that is admissible for the desired type of $\star$-structure.
(2) Construct the unique $\widetilde{L}(\lambda) \in \mathbb{DL}(P)$ with ansatz vector $w = Rv$ or $w = \Sigma v$, according to the desired structure.
(3) Premultiply $\widetilde{L}(\lambda)$ by $R^{-1} \otimes I$ or $\Sigma^{-1} \otimes I$ to obtain a structured pencil in $\mathbb{L}_1(P)$ with right ansatz vector v.

All that remains is to show how to carry out step (2). This can be done concretely and explicitly using the following canonical basis for $\mathbb{DL}(P)$ derived in [13]. Given a matrix polynomial $P(\lambda) = \sum_{i=0}^{k} \lambda^i A_i$, consider for $j = 0, \ldots, k$ the block diagonal matrices $X_j = \operatorname{diag}(\mathcal{L}_j, -\mathcal{U}_{k-j})$, whose diagonal blocks are the block $j \times j$ block-Hankel matrices

$$\mathcal{L}_j = \begin{bmatrix} & & & A_k \\ & & \iddots & A_{k-1} \\ & \iddots & \iddots & \vdots \\ A_k & A_{k-1} & \cdots & A_{k-j+1} \end{bmatrix} \quad \text{and} \quad \mathcal{U}_j = \begin{bmatrix} A_{j-1} & \cdots & A_1 & A_0 \\ \vdots & \iddots & \iddots & \\ A_1 & \iddots & & \\ A_0 & & & \end{bmatrix}.$$

Observe that $\mathcal{L}_j, \mathcal{U}_j \in \mathbb{F}^{jn \times jn}$, with the convention that they are empty when $j = 0$. Thus each X_j is a block $k \times k$ matrix in $\mathbb{F}^{kn \times kn}$. As an illustration we give the complete list of matrices X_0, X_1, X_2, X_3 for $k = 3$:

$$\begin{bmatrix} -A_2 & -A_1 & -A_0 \\ -A_1 & -A_0 & 0 \\ -A_0 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} A_3 & 0 & 0 \\ 0 & -A_1 & -A_0 \\ 0 & -A_0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & A_3 & 0 \\ A_3 & A_2 & 0 \\ 0 & 0 & -A_0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 0 & A_3 \\ 0 & A_3 & A_2 \\ A_3 & A_2 & A_1 \end{bmatrix}.$$

Matrices of this type have appeared in the literature; see, e.g., [9], [22], and for the scalar ($n = 1$) case see [21]. One can easily compute the shifted sums

$$X_j \boxplus\!\!\rightarrow (-X_{j-1}) = e_j \otimes \begin{bmatrix} A_k & \cdots & A_0 \end{bmatrix} \quad \text{and} \quad X_j \boxplus\!\!\downarrow (-X_{j-1}) = e_j^T \otimes \begin{bmatrix} A_k \\ \vdots \\ A_0 \end{bmatrix},$$

thus verifying by (3.3) and (3.4) that the pencil $\lambda X_j - X_{j-1}$ is in $\mathbb{DL}(P)$ with ansatz vector e_j for $j = 1, \ldots, k$. Consequently the set $\{\lambda X_j - X_{j-1} : j = 1, \ldots, k\}$ constitutes the natural or canonical basis for $\mathbb{DL}(P)$. A pencil $\lambda X + Y$ in $\mathbb{DL}(P)$ with ansatz vector $w = [w_1, \ldots, w_k]^T$ can now be uniquely expressed as a linear combination

$$\lambda X + Y \;=\; \sum_{j=1}^{k} w_j \big(\lambda X_j - X_{j-1}\big) \;=\; \lambda \sum_{j=1}^{k} w_j X_j \;-\; \sum_{j=1}^{k} w_j X_{j-1}. \tag{3.11}$$

Note that there are alternative procedures for the construction of pencils from $\mathbb{DL}(P)$; an explicit formula, for example, is given in [28], while a recursive method using the shifted sum is presented in [27].

3.5. Which structured pencils are linearizations? Recall from section 3.1 that when $P(\lambda)$ is regular, then any regular pencil in $\mathbb{L}_1(P)$ is a (strong) linearization for P. Although there is a systematic procedure [28] for determining the regularity of a pencil $L(\lambda) \in \mathbb{L}_1(P)$, there is, in general, no connection between this regularity and the right ansatz vector of $L(\lambda)$. By contrast, for pencils in $\mathbb{DL}(P)$ there is a criterion that characterizes regularity directly in terms of their ansatz vectors, which gives these pencils an important advantage. Let $v = [v_1, v_2, \ldots, v_k]^T$ be the ansatz vector of $L(\lambda) \in \mathbb{DL}(P)$ and define the associated v-*polynomial* to be the scalar polynomial

$$\mathsf{p}(x\,;v) := v_1 x^{k-1} + v_2 x^{k-2} + \cdots + v_{k-1}x + v_k.$$

By convention, we say that ∞ is a root of $\mathsf{p}(x\,;v)$ if $v_1 = 0$. Then regularity of $L(\lambda) \in \mathbb{DL}(P)$ can be expressed in terms of the roots of this v-polynomial and the eigenvalues of P, as follows.

THEOREM 3.7 (eigenvalue exclusion theorem [28]). *Suppose that $P(\lambda)$ is a regular matrix polynomial and $L(\lambda)$ is in $\mathbb{DL}(P)$ with nonzero ansatz vector v. Then $L(\lambda)$ is regular and thus a (strong) linearization for $P(\lambda)$ if and only if no root of the v-polynomial $\mathsf{p}(x\,;v)$ is an eigenvalue of $P(\lambda)$.*

Note that in Theorem 3.7 we include ∞ as one of the possible roots of $\mathsf{p}(x\,;v)$ or eigenvalues of P. We can now quickly deduce the following theorem.

THEOREM 3.8 (structured linearization theorem). *Suppose the regular matrix polynomial $P(\lambda)$ and the nonzero pencil $L(\lambda) \in \mathbb{L}_1(P)$ have one of the 16 combinations of $\star$-structure considered in Tables 3.1 and 3.2. Let v be the nonzero right ansatz vector of $L(\lambda)$, and let*

$$w = \begin{cases} Rv & \text{if } P \text{ is } \star\text{-palindromic or } \star\text{-anti-palindromic}, \\ \Sigma v & \text{if } P \text{ is } \star\text{-even or } \star\text{-odd}. \end{cases}$$

Then $L(\lambda)$ is a (strong) linearization for $P(\lambda)$ if and only if no root of the v-polynomial $\mathsf{p}(x\,;w)$ is an eigenvalue of $P(\lambda)$.

Proof. For all eight structure combinations in Table 3.1, it was shown in Theorem 3.5 that $(R \otimes I)L(\lambda)$ is in $\mathbb{DL}(P)$ with ansatz vector Rv. Similarly, for the eight even/odd structure combinations in Table 3.2 it was shown in Theorem 3.6 that $(\Sigma \otimes I)L(\lambda)$ is in $\mathbb{DL}(P)$ with ansatz vector Σv. Since $L(\lambda)$ is a linearization for $P(\lambda)$ if and only if $(R \otimes I)L(\lambda)$ or $(\Sigma \otimes I)L(\lambda)$ is, the desired result follows immediately from the eigenvalue exclusion theorem. ∎

We illustrate the implications of Theorem 3.8 with an example.

Example 3.9. Suppose the T-palindromic polynomial $P(\lambda) = \lambda^2 A + \lambda B + A^T$ from Example 3.2 is regular. Theorem 3.3 restricts the admissible ansatz vectors $v \in \mathbb{F}^2$ of a T-palindromic pencil $L(\lambda) \in \mathbb{L}_1(P)$ to those that satisfy $Rv = v$, or equivalently, $v = (v_1, v_1)^T$. We see from Theorem 3.8 that such an $L(\lambda)$ will be a strong linearization for $P(\lambda)$ if and only if none of the roots of the v-polynomial $\mathsf{p}(x\,;Rv) = v_1 x + v_1$ are eigenvalues of $P(\lambda)$, that is, if and only if -1 is not an eigenvalue of $P(\lambda)$. On the other hand, a T-anti-palindromic pencil $\widetilde{L}(\lambda) \in \mathbb{L}_1(P)$ will be a linearization for P if and only if $\lambda = 1$ is not an eigenvalue of $P(\lambda)$. This is because every admissible ansatz vector for $\widetilde{L}(\lambda)$ is constrained by Theorem 3.5 to be of the form $\widetilde{v} = [\,v_1, -v_1\,]^T$, forcing $\mathsf{p}(x\,;R\widetilde{v}) = -v_1 x + v_1$, with only $+1$ as a root.

This example also illustrates another way in which structure influences the players in our story: when P is T-palindromic, any ansatz vector admissible for a T-(anti-)palindromic pencil in $\mathbb{L}_1(P)$ has components that read the same forwards or backwards (up to sign). This in turn forces the corresponding v-polynomial to be (anti-)palindromic. Theorems 3.5 and 3.6 imply that analogous parallels in structure hold for other combinations of $\star$-structures in P and L and the relevant v-polynomial $\mathsf{p}(x\,;Rv)$ or $\mathsf{p}(x\,;\Sigma v)$; for convenience these are listed together in Table 3.3.

TABLE 3.3
Parallelism of structures.

$P(\lambda)$	$L(\lambda) \in \mathbb{L}_1(P)$	v-polynomial
$\star$-palindromic	$\star$-palindromic	$\star$-palindromic
	$\star$-anti-palindromic	$\star$-anti-palindromic
$\star$-anti-palindromic	$\star$-palindromic	$\star$-anti-palindromic
	$\star$-anti-palindromic	$\star$-palindromic

$P(\lambda)$	$L(\lambda) \in \mathbb{L}_1(P)$	v-poly.
$\star$-even	$\star$-even	$\star$-even
	$\star$-odd	$\star$-odd
$\star$-odd	$\star$-even	$\star$-odd
	$\star$-odd	$\star$-even

3.6. When pairings degenerate. The parallel of structures between matrix polynomial, $\mathbb{L}_1(P)$-pencil, and v-polynomial (see Table 3.3) is aesthetically very pleasing: structure in a v-polynomial forces a pairing of its roots, as in Theorem 2.2, which is always of the *same qualitative type* as the eigenvalue pairing present in the original structured matrix polynomial. However, it turns out that this root pairing can sometimes be an obstruction to the existence of any structured linearization in $\mathbb{L}_1(P)$ at all.

Using an argument based mainly on the very simple form of admissible ansatz vectors when $k = 2$, we saw in Example 3.9 that a quadratic T-palindromic matrix polynomial having both 1 and -1 as eigenvalues cannot have a structured linearization in $\mathbb{L}_1(P)$: the presence of -1 in the spectrum precludes the existence of a T-palindromic linearization, while the eigenvalue 1 excludes T-anti-palindromic linearizations. We now show that this difficulty is actually a consequence of root pairing, and therefore can occur for higher degree polynomials.

When $P(\lambda)$ has even degree, all its ansatz vectors have even length, and hence the corresponding v-polynomials all have an odd number of roots (counting multiplicities and including ∞). Root pairing then forces at least one root of every v-polynomial to lie in a subset of $\mathbb{C}$ where this pairing "degenerates." This means that for any T-(anti-)palindromic matrix polynomial $P(\lambda)$ of even degree, *every* v-polynomial of a T-(anti-)palindromic pencil in $\mathbb{L}_1(P)$ has at least one root belonging to $\{-1, +1\}$. It follows that any such $P(\lambda)$ having both $+1$ and -1 as eigenvalues can have neither a T-palindromic nor a T-anti-palindromic linearization in $\mathbb{L}_1(P)$. For T-even/odd matrix polynomials $P(\lambda)$ of even degree, every relevant v-polynomial has a root belonging to $\{0, \infty\}$; thus if the spectrum of $P(\lambda)$ includes both 0 and ∞, then P cannot have a T-even or T-odd linearization in $\mathbb{L}_1(P)$.

When no structured linearization for $P(\lambda)$ exists in $\mathbb{L}_1(P)$, it is natural to ask whether $P(\lambda)$ has a structured linearization that is *not* in $\mathbb{L}_1(P)$, or perhaps has no structured linearizations at all. The next examples show that both alternatives may occur.

Example 3.10. Consider the T-palindromic polynomial $P(\lambda) = \lambda^2 + 2\lambda + 1$. Then the only eigenvalue of $P(\lambda)$ is -1, so by the observation in Example 3.9 we see that $P(\lambda)$ cannot have any T-palindromic linearization in $\mathbb{L}_1(P)$. But does $P(\lambda)$ have a

T-palindromic linearization $L(\lambda)$ which is not in $\mathbb{L}_1(P)$? Consider the general 2×2 T-palindromic pencil

$$L(\lambda) = \lambda Z + Z^T = \lambda \begin{bmatrix} w & x \\ y & z \end{bmatrix} + \begin{bmatrix} w & y \\ x & z \end{bmatrix} = \begin{bmatrix} w(\lambda+1) & \lambda x + y \\ \lambda y + x & z(\lambda+1) \end{bmatrix}, \tag{3.12}$$

and suppose it is a linearization for P. Since the only eigenvalue $\lambda = -1$ of $P(\lambda)$ has geometric multiplicity 1, the same must be true for $L(\lambda)$, that is, rank $L(-1) = 1$. But inserting $\lambda = -1$ in (3.12), we obtain a matrix that does not have rank 1 for any values of w, x, y, z. Thus $P(\lambda)$ does not have any T-palindromic linearization. However, $P(\lambda)$ does have a T-anti-palindromic linearization $\widetilde{L}(\lambda)$ in $\mathbb{L}_1(P)$ because it does not have the eigenvalue $+1$. Choosing $\widetilde{v} = (1, -1)^T$ as the right ansatz vector and following the procedure in section 3.4 yields the structured linearization

$$\widetilde{L}(\lambda) = \lambda \widetilde{Z} - \widetilde{Z}^T = \lambda \begin{bmatrix} 1 & 3 \\ -1 & 1 \end{bmatrix} - \begin{bmatrix} 1 & -1 \\ 3 & 1 \end{bmatrix} \in \mathbb{L}_1(P).$$

Example 3.11. Consider the T-palindromic matrix polynomial

$$P(\lambda) = \lambda^2 \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

Since $\det P(\lambda) = (\lambda^2 - 1)^2$, this polynomial $P(\lambda)$ has $+1$ and -1 as eigenvalues, each with algebraic multiplicity 2. Thus $P(\lambda)$ has neither a T-palindromic nor a T-anti-palindromic linearization in $\mathbb{L}_1(P)$. However, it is possible to construct a T-palindromic linearization for $P(\lambda)$ that is not in $\mathbb{L}_1(P)$. Starting with the first companion form $C_1(\lambda)$, one can verify that

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \cdot C_1(\lambda) \cdot \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \lambda \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

is a T-palindromic linearization for $P(\lambda)$. Using shifted sums, it can easily be verified that this linearization is in neither $\mathbb{L}_1(P)$ nor $\mathbb{L}_2(P)$.

Example 3.12. Consider the scalar matrix polynomial $P(\lambda) = \lambda^2 - 1$ which is T-anti-palindromic and has the roots ± 1. Again, the presence of these eigenvalues precludes the existence of either a T-palindromic or T-anti-palindromic linearization in $\mathbb{L}_1(P)$. But even more is true. It turns out that $P(\lambda)$ does not have any T-palindromic or T-anti-palindromic linearization at all. Indeed, suppose that $L_\varepsilon(\lambda) = \lambda Z + \varepsilon Z^T$ was a linearization for $P(\lambda)$, where $\varepsilon = \pm 1$; that is, $L_\varepsilon(\lambda)$ is T-palindromic if $\varepsilon = 1$ and T-anti-palindromic if $\varepsilon = -1$. Since $P(\lambda)$ does not have the eigenvalue ∞, neither does $L(\lambda)$, and so Z must be invertible. Thus $L_\varepsilon(\lambda)$ is strictly equivalent to the pencil $\lambda I + \varepsilon Z^{-1} Z^T$. But this being a linearization for $P(\lambda)$ forces the matrix $\varepsilon Z^{-1} Z^T$ to have the simple eigenvalues $+1$ and -1, and hence $\det \varepsilon Z^{-1} Z^T = -1$. However, we also see that

$$\det \varepsilon Z^{-1} Z^T = \varepsilon^2 \frac{1}{\det Z} \det Z = 1,$$

which is a contradiction. Hence $P(\lambda)$ has neither a T-palindromic linearization nor a T-anti-palindromic linearization.

One possibility for circumventing the difficulties associated with the eigenvalues ± 1 is to first deflate them in a structure-preserving manner, using a procedure that

works directly on the original matrix polynomial. Since the resulting matrix polynomial $P(\lambda)$ will not have these troublesome eigenvalues, a structured linearization in $\mathbb{L}_1(P)$ can then be constructed. Such structure-preserving deflation strategies are currently under investigation.

The situation is quite different for $*$-(anti-)palindromic and $*$-even/odd matrix polynomials, because now the set where pairing degenerates is the entire unit circle in $\mathbb{C}$, or the imaginary axis (including ∞), respectively. The contrast between having a continuum versus a finite set in which the root pairing degenerates makes a crucial difference in our ability to guarantee the existence of structured linearizations in $\mathbb{L}_1(P)$. Indeed, consider a regular $*$-palindromic matrix polynomial $P(\lambda)$ of degree k. Then the v-polynomial $\mathsf{p}(x\,;Rv)$ corresponding to an admissible ansatz vector is again $*$-palindromic, with $k-1$ roots occurring in pairs $(\lambda, 1/\overline{\lambda})$ by Theorem 2.2. Thus if k is even, at least one root of $\mathsf{p}(x\,;Rv)$ must lie on the unit circle. But since the spectrum of $P(\lambda)$ is a finite set, it is always possible to choose v so that all the roots of $\mathsf{p}(x\,;Rv)$ avoid the spectrum of $P(\lambda)$. Example 3.13 illustrates the case $k=2$.

Example 3.13. Consider a regular matrix polynomial $P(\lambda) = \lambda^2 A + \lambda B + A^*$ which is $*$-palindromic, that is, $B = B^*$. Choose ζ on the unit circle in $\mathbb{C}$ such that ζ is not an eigenvalue of $P(\lambda)$. Now choose $z \in \mathbb{C}$ so that $\zeta = -z/\overline{z}$. Then $v = (z, \overline{z})^T$ satisfies $Rv = \overline{v}$, and the associated v-polynomial $\mathsf{p}(x\,;Rv) = \overline{z}x + z$ has ζ as its only root. Therefore the $*$-palindromic pencil

$$L(\lambda) = \lambda \begin{bmatrix} zA & zB - \overline{z}A^* \\ \overline{z}A & zA \end{bmatrix} + \begin{bmatrix} \overline{z}A^* & zA^* \\ \overline{z}B - zA & \overline{z}A^* \end{bmatrix} \in \mathbb{L}_1(P)$$

with right ansatz vector v is a (strong) linearization for $P(\lambda)$ by Theorem 3.8.

The observations made in this section have parallels for $\star$-even/odd structures. A list of structured linearizations in $\mathbb{L}_1(P)$ for $\star$-(anti-)palindromic and $\star$-even/odd matrix polynomials of degree $k = 2, 3$ is compiled in Tables 3.4, 3.5, and 3.6.

3.7. The missing structures. So far in section 3 our attention has been focused on finding structured linearizations only for the eight $\star$-structures in Table 2.1. But what about "purely" palindromic, anti-palindromic, even, and odd matrix polynomials? Why have they been excluded from consideration? It turns out that these structures cannot be linearized in a structure-preserving way. For example, consider a regular palindromic polynomial $P(\lambda)$ of degree $k \geq 2$. By [11, Theorem 1.7] a pencil can be a linearization for $P(\lambda)$ only if the geometric multiplicity of each eigenvalue of the pencil is less than or equal to n. On the other hand, any palindromic linearization has the form $L(\lambda) = \lambda Z + Z$, and thus must have the eigenvalue -1 with geometric multiplicity kn. Analogous arguments exclude structure-preserving linearizations for anti-palindromic, even, and odd polynomials.

TABLE 3.4

Structured linearizations for $\lambda^2 A + \lambda B + C$. Except for the parameters $r \in \mathbb{R}$ and $z \in \mathbb{C}$, the linearizations are unique up to a (suitable) scalar factor. The last column lists the roots of the v*-polynomial* $\mathsf{p}(x\,; Mv)$ *corresponding to $M = R$ or $M = \Sigma$, respectively.*

Structure of $P(\lambda)$	Structure of $L(\lambda)$	v	$L(\lambda)$ with ansatz vector v	Root of $\mathsf{p}(x; Mv)$
T-palindromic	T-palindromic	$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$\lambda \begin{bmatrix} A & B-C \\ A & A \end{bmatrix} + \begin{bmatrix} C & C \\ B-A & C \end{bmatrix}$	-1
$B = B^T$ $C = A^T$	T-antipalindromic	$\begin{bmatrix} 1 \\ -1 \end{bmatrix}$	$\lambda \begin{bmatrix} A & B+C \\ -A & A \end{bmatrix} + \begin{bmatrix} -C & C \\ -B-A & -C \end{bmatrix}$	1
T-antipalindromic	T-palindromic	$\begin{bmatrix} 1 \\ -1 \end{bmatrix}$	$\lambda \begin{bmatrix} A & B+C \\ -A & A \end{bmatrix} + \begin{bmatrix} -C & C \\ -B-A & -C \end{bmatrix}$	1
$B = -B^T$ $C = -A^T$	T-antipalindromic	$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$\lambda \begin{bmatrix} A & B-C \\ A & A \end{bmatrix} + \begin{bmatrix} C & C \\ B-A & C \end{bmatrix}$	-1
$*$-palindromic	$*$-palindromic	$\begin{bmatrix} z \\ \bar{z} \end{bmatrix}$	$\lambda \begin{bmatrix} zA & zB-\bar{z}C \\ \bar{z}A & zA \end{bmatrix} + \begin{bmatrix} \bar{z}C & zC \\ \bar{z}B-zA & \bar{z}C \end{bmatrix}$	$-z/\bar{z}$
$B = B^*$ $C = A^*$	$*$-antipalindromic	$\begin{bmatrix} z \\ -\bar{z} \end{bmatrix}$	$\lambda \begin{bmatrix} zA & zB+\bar{z}C \\ -\bar{z}A & zA \end{bmatrix} + \begin{bmatrix} -\bar{z}C & zC \\ -\bar{z}B-zA & -\bar{z}C \end{bmatrix}$	$z/\bar{z}$
T-even	T-even	$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$	$\lambda \begin{bmatrix} 0 & -A \\ A & B \end{bmatrix} + \begin{bmatrix} A & 0 \\ 0 & C \end{bmatrix}$	∞
$A = A^T$ $B = -B^T$ $C = C^T$	T-odd	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	$\lambda \begin{bmatrix} A & 0 \\ 0 & C \end{bmatrix} + \begin{bmatrix} B & C \\ -C & 0 \end{bmatrix}$	0
T-odd	T-even	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	$\lambda \begin{bmatrix} A & 0 \\ 0 & C \end{bmatrix} + \begin{bmatrix} B & C \\ -C & 0 \end{bmatrix}$	0
$A = -A^T$ $B = B^T$ $C = -C^T$	T-odd	$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$	$\lambda \begin{bmatrix} 0 & -A \\ A & B \end{bmatrix} + \begin{bmatrix} A & 0 \\ 0 & C \end{bmatrix}$	∞
$*$-even	$*$-even	$\begin{bmatrix} i \\ r \end{bmatrix}$	$\lambda \begin{bmatrix} iA & -rA \\ rA & rB+iC \end{bmatrix} + \begin{bmatrix} rA+iB & iC \\ -iC & rC \end{bmatrix}$	$-ir$
$A = A^*$ $B = -B^*$ $C = C^*$	$*$-odd	$\begin{bmatrix} r \\ i \end{bmatrix}$	$\lambda \begin{bmatrix} rA & -iA \\ iA & iB+rC \end{bmatrix} + \begin{bmatrix} iA+rB & rC \\ -rC & iC \end{bmatrix}$	$\frac{i}{r}$

TABLE 3.5

$\star$-palindromic linearizations for the $\star$-palindromic matrix polynomial $\lambda^3 A + \lambda^2 B + \lambda B^\star + A^\star$. The last column lists the roots of the v-polynomial $\mathsf{p}(x\,;Rv)$ corresponding to Rv. All $\star$-palindromic linearizations in $\mathbb{L}_1(P)$ for this matrix polynomial are linear combinations of the first two linearizations in the case $\star = T$, and are real linear combinations of the first three linearizations in the case $\star = *$. A specific example is given by the fourth linearization.

v	$L(\lambda)$ with right ansatz vector v	Roots of $\mathsf{p}(x\,;Rv)$
$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$	$\lambda \begin{bmatrix} 0 & 0 & -A^\star \\ A & B & 0 \\ 0 & A & 0 \end{bmatrix} + \begin{bmatrix} 0 & A^\star & 0 \\ 0 & B^\star & A^\star \\ -A & 0 & 0 \end{bmatrix}$	$0, \infty$
$\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$	$\lambda \begin{bmatrix} A & B - A^\star & B^\star \\ 0 & A - B^\star & B - A^\star \\ A & 0 & A \end{bmatrix} + \begin{bmatrix} A^\star & 0 & A^\star \\ B^\star - A & A^\star - B & 0 \\ B & B^\star - A & A^\star \end{bmatrix}$	$i, -i$
$\begin{bmatrix} i \\ 0 \\ -i \end{bmatrix}$	$\lambda \begin{bmatrix} iA & iB + iA^* & iB^* \\ 0 & iA + iB^* & iB + iA^* \\ -iA & 0 & iA \end{bmatrix} + \begin{bmatrix} -iA^* & 0 & iA^* \\ -iB^* - iA & -iA^* - iB & 0 \\ -iB & -iB^* - iA & -iA^* \end{bmatrix}$	$1, -1$
$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$	$\lambda \begin{bmatrix} A & B - A^\star & B^\star - A^\star \\ A & B + A - B^\star & B - A^\star \\ A & A & A \end{bmatrix} + \begin{bmatrix} A^\star & A^\star & A^\star \\ B^\star - A & B^\star + A^\star - B & A^\star \\ B - A & B^\star - A & A^\star \end{bmatrix}$	$\frac{-1 \pm i\sqrt{3}}{2}$

TABLE 3.6

$\star$-even linearizations for the $\star$-even matrix polynomial $P(\lambda) = \lambda^3 A + \lambda^2 B + \lambda C + D$, where $A = -A^\star$, $B = B^\star$, $C = -C^\star$, $D = D^\star$. The last column lists the roots of the v-polynomial $\mathsf{p}(x\,;\Sigma v)$ corresponding to Σv. All $\star$-even linearizations in $\mathbb{L}_1(P)$ for this matrix polynomial are linear combinations of the first two linearizations in the case $\star = T$, and are real linear combinations of the first three linearizations in the case $\star = *$. A specific example is given by the fourth linearization.

v	$L(\lambda)$ with right ansatz vector v	Roots of $\mathsf{p}(x\,;\Sigma v)$
$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$	$\lambda \begin{bmatrix} 0 & 0 & A \\ 0 & -A & -B \\ A & B & C \end{bmatrix} + \begin{bmatrix} 0 & -A & 0 \\ A & B & 0 \\ 0 & 0 & D \end{bmatrix}$	∞
$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$	$\lambda \begin{bmatrix} A & 0 & 0 \\ 0 & C & D \\ 0 & -D & 0 \end{bmatrix} + \begin{bmatrix} B & C & D \\ -C & -D & 0 \\ D & 0 & 0 \end{bmatrix}$	0
$\begin{bmatrix} 0 \\ i \\ 0 \end{bmatrix}$	$\lambda \begin{bmatrix} 0 & -iA & 0 \\ iA & iB & 0 \\ 0 & 0 & iD \end{bmatrix} + \begin{bmatrix} iA & 0 & 0 \\ 0 & iC & iD \\ 0 & -iD & 0 \end{bmatrix}$	$0, \infty$
$\begin{bmatrix} 1 \\ 0 \\ 4 \end{bmatrix}$	$\lambda \begin{bmatrix} A & 0 & 4A \\ 0 & C - 4A & D - 4B \\ 4A & 4B - D & 4C \end{bmatrix} + \begin{bmatrix} B & C - 4A & D \\ 4A - C & 4B - D & 0 \\ D & 0 & 4D \end{bmatrix}$	$2i, -2i$

4. Good vibrations from good linearizations. As an illustration of the importance of structure preservation in practical problems, we indicate how the techniques developed in this paper have had a significant impact on computations in an eigenvalue problem occurring in the vibration analysis of rail tracks under excitation arising from high speed trains. This eigenvalue problem has the form

$$\left(\kappa A(\omega) + B(\omega) + \tfrac{1}{\kappa} A(\omega)^T\right) x = 0, \tag{4.1}$$

where A, B are large, sparse, parameter-dependent, complex square matrices with B complex symmetric and A highly singular. For details of the derivation of this model, see [16] and [17]. The parameter ω is the excitation frequency and the eigenvalue problem has to be solved over a wide frequency range of $\omega = 0$–$5{,}000\,\mathrm{Hz}$. Clearly, for any fixed value of ω, multiplying (4.1) by κ leads to the T-palindromic eigenvalue problem introduced in (1.2). In addition to the presence of a large number of zero and infinite eigenvalues caused by the rank deficiency of A, the finite nonzero eigenvalues cover a wide range of magnitudes that increases as the finite element discretization is made finer. The eigenvalues of the problem under consideration range from 10^{15} to 10^{-15}, thereby making this a very challenging numerical problem.

Attempts at solving this problem with the QZ-algorithm without respecting its structure resulted in computed eigenvalues with no correct digits even in quadruple precision arithmetic. Furthermore, the symmetry of the spectrum with respect to the unit circle was highly perturbed [16].

As an alternative, in [16], [17] a T-palindromic linearization for the eigenvalue problem (4.1) was used. Based on this linearization, the infinite and zero eigenvalues of the resulting T-palindromic pencil could be deflated in a structure-preserving way. The resulting smaller T-palindromic problem was then solved via different methods, resulting in eigenvalues with good accuracy in double precision arithmetic; i.e., the computed eigenvalues were accurate to within the range of the discretization error of the underlying finite element discretization. Thus physically useful eigenvalues were determined, with no modification in the mathematical model or in the discretization scheme. Only the numerical linear algebra was changed, to methods based on the new structure-preserving linearization techniques described in this paper.

Thus we see that the computation of "good vibrations" (i.e., accurate eigenvalues and eigenvectors) requires the use of "good linearizations" (i.e., linearizations that reflect the structure of the original polynomial).

5. Conclusions. The numerical solution of structured nonlinear eigenvalue problems is an important component of many applications. Building on the work in [28], we have developed a theory that provides criteria for the existence of strong linearizations that reflect $\star$-even/odd or $\star$-(anti-)palindromic structure of a matrix polynomial, and have presented a systematic method to construct such linearizations. As shown in [16], [17], numerical methods based on these structured linearizations are expected to be more effective in computing accurate eigenvalues in practical applications.

Acknowledgments. We thank the mathematics departments of the University of Manchester, Technische Universität Berlin, and Western Michigan University, and the Banff International Research Station for giving us the opportunity to carry out this joint research. We thank Françoise Tisseur for helpful comments on an earlier draft, and Ralph Byers for several enlightening discussions on this topic. Finally, we thank three referees for useful suggestions that helped us to significantly improve an earlier version of this paper.

REFERENCES

[1] H. Abou-Kandil, G. Freiling, V. Ionescu, and G. Jank, *Matrix Riccati Equations in Control and Systems Theory*, Birkhäuser, Basel, 2003.

[2] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, 3rd ed., SIAM, Philadelphia, 1999.

[3] T. Apel, V. Mehrmann, and D. Watkins, *Structured eigenvalue methods for the computation of corner singularities in* 3*D anisotropic elastic structures.*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 4459–4473.

[4] T. Apel, V. Mehrmann, and D. Watkins, *Numerical solution of large scale structured polynomial or rational eigenvalue problems*, in Foundations of Computational Mathematics, (Minneapolis 2002), F. Cucker and P. Olver, eds., London Math. Soc. Lecture Note Ser. 312, Cambridge University Press, Cambridge, UK, 2004, pp. 137–157.

[5] P. Benner, R. Byers, V. Mehrmann, and H. Xu, *Numerical computation of deflating subspaces of skew-Hamiltonian/Hamiltonian pencils*, SIAM J. Matrix Anal. Appl., 24 (2002), pp. 165–190.

[6] D. Chu, X. Liu, and V. Mehrmann, *A Numerically Strong Stable Method for Computing the Hamiltonian Schur Form*, Preprint 24-2004, Institut für Mathematik, TU Berlin, Berlin, 2004. Available online at http://www.math.tu-berlin.de/preprints/.

[7] H. Fassbender, *Symplectic Methods for the Symplectic Eigenvalue Problem*, Kluwer Academic, New York, 2000.

[8] F. R. Gantmacher, *Theory of Matrices*, Vol. 2, Chelsea, New York, 1959.

[9] S. Garvey, U. Prells, M. Friswell, and Z. Chen, *General isospectral flows for linear dynamic systems*, Linear Algebra Appl., 385 (2004), pp. 335–368.

[10] I. Gohberg, M. A. Kaashoek, and P. Lancaster, *General theory of regular matrix polynomials and band Toeplitz operators*, Integral Equations Operator Theory, 11 (1988), pp. 776–882.

[11] I. Gohberg, P. Lancaster, and L. Rodman, *Matrix Polynomials*, Academic Press, New York, 1982.

[12] I. Gohberg, P. Lancaster, and L. Rodman, *Matrices and Indefinite Scalar Products*, Birkhäuser, Basel, 1983.

[13] N. J. Higham, D. S. Mackey, N. Mackey, and F. Tisseur, *Symmetric linearizations for matrix polynomials*, SIAM J. Matrix Anal. Appl., to appear.

[14] N. J. Higham, D. S. Mackey, and F. Tisseur, *The conditioning of linearizations of matrix polynomials*, SIAM J. Matrix Anal. Appl., 28 (2006), 1005–1028.

[15] N. J. Higham, F. Tisseur, and P. M. V. Dooren, *Detecting a definite Hermitian pair and a hyperbolic or elliptic quadratic eigenvalue problem, and associated nearness problems*, Linear Algebra Appl., 351/352 (2002), pp. 455–474.

[16] A. Hilliges, *Numerische Lösung von quadratischen Eigenwertproblemen mit Anwendung in der Schienendynamik*, Diplomarbeit, TU Berlin, Institut für Mathematik, Berlin, 2004.

[17] A. Hilliges, C. Mehl, and V. Mehrmann, *On the solution of palindromic eigenvalue problems*, in Proceedings of the 4th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS), Jyväskylä, Finland, 2004. CD-ROM.

[18] P. Hilton, *Personal communication*, 2004.

[19] P. Hilton, *Inventing a palindrome*, College Math. J., 5 (1999), p. 422.

[20] I. C. F. Ipsen, *Accurate eigenvalues for fast trains*, SIAM News, 37(9), Nov. 2004, pp. 1–2.

[21] P. Lancaster, *Symmetric transformations of the companion matrix*, NABLA: Bull. Malayan Math. Soc., 8 (1961), pp. 146–148.

[22] P. Lancaster, *Lambda-matrices and Vibrating Systems*, Pergamon Press, Oxford, 1966.

[23] P. Lancaster and P. Psarrakos, *A Note on Weak and Strong Linearizations of Regular Matrix Polynomials*, Numerical Analysis Report 470, Manchester Centre for Computational Mathematics, Manchester, UK, 2005.

[24] P. Lancaster and L. Rodman, *The Algebraic Riccati Equation*, Oxford University Press, Oxford, UK, 1995.

[25] D. Leguillon, *Computation of* 3*d-singularities in elasticity*, in Boundary Value Problems and Integral Equations in Nonsmooth Domains, M. Costabel, M. Dauge, and S. Nicaise, eds., Lecture Notes in Pure Appl. Math. 167, Marcel Dekker, New York, 1995, pp. 161–170.

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

[27] D. S. Mackey, N. Mackey, C. Mehl, and V. Mehrmann, *Palindromic Polynomial Eigenvalue Problems: Good Vibrations from Good Linearizations*, Preprint 239, DFG Research Center Matheon, Mathematics for key technologies, TU Berlin, Berlin, 2005. Available online at http://www.matheon.de/.

[28] 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.

[29] D. S. Mackey, N. Mackey, and F. Tisseur, *Structured factorizations in scalar product spaces*, SIAM J. Matrix Anal. Appl., 27 (2006), pp. 821–850.

[30] *MATLAB, Version* 5. The MathWorks, Inc., Natick, MA, 1996.

[31] V. Mehrmann, *The Autonomous Linear Quadratic Control Problem, Theory and Numerical Solution*, Lecture Notes in Control and Inform. Sci. 163, Springer-Verlag, Heidelberg, 1991.

[32] V. Mehrmann, *A step toward a unified treatment of continuous and discrete time control problems*, Linear Algebra Appl., 241/243 (1996), pp. 749–779.

[33] 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.

[34] V. Mehrmann and D. Watkins, *Polynomial eigenvalue problems with Hamiltonian structure*, Electron. Trans. Numer. Anal., 13 (2002), pp. 106–118.

[35] F. Tisseur and K. Meerbergen, *The quadratic eigenvalue problem*, SIAM Rev., 43 (2001), pp. 235–286.

[36] S. Zaglmayr, *Eigenvalue Problems in SAW-Filter Simulations*, Diplomarbeit, Institute of Computational Mathematics, Johannes Kepler University Linz, Linz, Austria, 2002.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 4, pp. 1052–1068

STRUCTURED EIGENVALUE CONDITION NUMBERS*

MICHAEL KAROW†, DANIEL KRESSNER*, AND FRANÇOISE TISSEUR‡

Abstract. This paper investigates the effect of structure-preserving perturbations on the eigenvalues of linearly and nonlinearly structured eigenvalue problems. Particular attention is paid to structures that form Jordan algebras, Lie algebras, and automorphism groups of a scalar product. Bounds and computable expressions for structured eigenvalue condition numbers are derived for these classes of matrices, which include complex symmetric, pseudo-symmetric, persymmetric, skew-symmetric, Hamiltonian, symplectic, and orthogonal matrices. In particular we show that under reasonable assumptions on the scalar product, the structured and unstructured eigenvalue condition numbers are equal for structures in Jordan algebras. For Lie algebras, the effect on the condition number of incorporating structure varies greatly with the structure. We identify Lie algebras for which structure does not affect the eigenvalue condition number.

Key words. structured eigenvalue problem, condition number, Jordan algebra, Lie algebra, automorphism group, symplectic, perplectic, pseudo-orthogonal, pseudo-unitary, complex symmetric, persymmetric, perskew-symmetric, Hamiltonian, skew-Hamiltonian, structure preservation

AMS subject classifications. 65F15, 65F35

DOI. 10.1137/050628519

1. Introduction.

There is a growing interest in structured perturbation analysis due to the substantial development of algorithms for structured problems. When these algorithms preserve structure (see, for example, [2], [4], [13], and the literature cited therein) it is often appropriate to consider condition numbers that measure the sensitivity to structured perturbations. In this paper we investigate the effect of structure-preserving perturbations on linearly and nonlinearly structured eigenvalue problems.

Suppose that $\mathbb{S}$ is a class of structured matrices and define the (absolute) *structured condition number* of a simple eigenvalue λ of $A \in \mathbb{S}$ by

$$\kappa(A,\lambda;\mathbb{S}) = \lim_{\epsilon\to 0} \sup\Big\{ \frac{|\widehat{\lambda}-\lambda|}{\epsilon} : \widehat{\lambda} \in \mathrm{Sp}(A+E),\ A+E \in \mathbb{S},\ \|E\| \le \epsilon \Big\}, \tag{1.1}$$

where $\mathrm{Sp}(A+E)$ denotes the spectrum of $A+E$ and $\|\cdot\|$ is an arbitrary matrix norm. Let x and y be the normalized right and left eigenvectors associated with λ, i.e.,

$$Ax = \lambda x, \quad y^*A = \lambda y^*, \quad \|x\|_2 = \|y\|_2 = 1.$$

Moreover, let $\kappa(A,\lambda) \equiv \kappa(A,\lambda;\mathbb{C}^{n\times n})$ denote the standard unstructured eigenvalue condition number, where n is the dimension of A. Clearly,

$$\kappa(A,\lambda;\mathbb{S}) \le \kappa(A,\lambda).$$

*Received by the editors April 5, 2005; accepted for publication (in revised form) by K. Veselic May 23, 2006; published electronically December 18, 2006.
http://www.siam.org/journals/simax/28-4/62851.html

†Institut für Mathematik, MA 4-5, TU Berlin, Str. des 17. Juni 136, D-10623 Berlin, Germany (karow@math.tu-berlin.de, kressner@math.tu-berlin.de). The work of the second author was supported by the DFG Research Center MATHEON "Mathematics for key technologies" in Berlin.

‡School of Mathematics, The University of Manchester, Sackville Street, Manchester, M60 1QD, UK (ftisseur@ma.man.ac.uk, http://www.ma.man.ac.uk/~ftisseur). This work was supported by Engineering and Physical Sciences Research Council grant GR/S31693.

If this inequality is not always close to being attained, then $\kappa(A,\lambda)$ may severely overestimate the worst case effect of structured perturbations. Note that the standard eigenvalue condition number allows complex perturbations even if A is real. Our definition in (1.1) automatically forces the perturbations to be real when A is real and $\mathbb{S} \subset \mathbb{R}^{n\times n}$.

In this paper we consider the case where $\mathbb{S}$ is a smooth manifold. This covers linear structures and some nonlinear structures, such as orthogonal, unitary, and symplectic structures. We show that for such $\mathbb{S}$, the structured problem in (1.1) simplifies to a linearly constrained optimization problem. We obtain an explicit expression for $\kappa(A,\lambda;\mathbb{S})$, thereby extending Higham and Higham's work [11] for linear structures in $\mathbb{C}^{n\times n}$.

Associated with a scalar product in $\mathbb{R}^n$ or $\mathbb{C}^n$ are three important classes of structured matrices: an automorphism group, a Lie algebra, and a Jordan algebra. We specialize our results to each of these three classes, starting with the linear structures. We show that under mild assumptions on the scalar product, the structured and unstructured eigenvalue condition numbers are equal for structures in Jordan algebras. For example, this equality holds for real and complex symmetric matrices, pseudo-symmetric, persymmetric, Hermitian, and J-Hermitian matrices. For Lie algebras, the effect on the condition number of incorporating structure varies greatly with the structure. We identify Lie algebras for which structure does not affect the eigenvalue condition number, such as skew-Hermitian structures, and Lie algebras for which the ratio between the unstructured and structured eigenvalue condition number can be large, such as skew-symmetric or perskew-symmetric structures. Our treatment extends and unifies recent work on these classes of matrices by Graillat [9] and Rump [17].

Finally we show how to compute structured eigenvalue condition numbers when $\mathbb{S}$ is the automorphism group of a scalar product. This includes the classes of unitary, complex orthogonal, and symplectic matrices. We provide bounds for the ratio between the structured and unstructured condition number. In particular we show that for unitary matrices this ratio is always equal to 1. This latter result also holds for orthogonal matrices with one exception: when λ is real and simple, the structured eigenvalue condition number is zero.

Note that for $\lambda \neq 0$ a relative condition number, on both data and output spaces, can also be defined, which is just $\kappa(A,\lambda;\mathbb{S})\|A\|/|\lambda|$. Our results comparing structured and unstructured absolute condition numbers clearly apply to relative condition numbers without change.

The rest of this paper is organized as follows. Section 2 provides the definition and a computable expression for the structured eigenvalue condition number of a nonlinearly structured matrix. In section 3, we introduce the scalar products and the associated structures to be considered. We treat linear structures (Jordan and Lie algebras) in section 4 and investigate the corresponding structured condition numbers. Nonlinear structures (automorphism groups) are discussed in section 5.

2. Structured condition number. It is well known that simple eigenvalues $\lambda \in \mathrm{Sp}(A)$ depend analytically on the entries of A in a sufficiently small open neighborhood $\mathcal{B}_A$ of A [18]. To be more specific, there exists a uniquely defined analytic function $f_\lambda : \mathcal{B}_A \to \mathbb{C}$ so that $\lambda = f_\lambda(A)$ and $\widehat{\lambda} = f_\lambda(A+E)$ is an eigenvalue of $A+E$ for every $A+E \in \mathcal{B}_A$. Moreover, one has the expansion

$$\hat{\lambda} = \lambda + \frac{1}{|y^*x|} y^*Ex + O(\|E\|^2). \tag{2.1}$$

Combined with (1.1) this yields

$$\kappa(A,\lambda;\mathbb{S}) = \frac{1}{|y^*x|}\limsup_{\epsilon\to 0}\Big\{\frac{|y^*Ex|}{\epsilon} :\ A+E\in\mathbb{S},\ \|E\|\le\epsilon\Big\}. \tag{2.2}$$

The difficulty in obtaining an explicit expression for the supremum in (2.2) depends on the nature of $\mathbb{S}$ and the matrix norm $\|\cdot\|$. For example, when $\|\cdot\|$ is the Frobenius norm or the matrix 2-norm and for unstructured perturbations (i.e., $\mathbb{S}=\mathbb{C}^{n\times n}$), the supremum in (2.2) is attained by $E=\epsilon yx^*$, which implies the well-known formula [20]

$$\kappa_\nu(A,\lambda) = 1/|y^*x|, \quad \nu = 2, F.$$

Note that $\kappa_\nu(A,\lambda)\ge 1$ always, but $\kappa_\nu(A,\lambda;\mathbb{S})$ can be less than 1 for $\nu=2,F$.

When $\mathbb{S}$ is a smooth manifold (see [12] for an introduction to smooth manifolds), the task of computing the supremum (2.2) simplifies to a linearly constrained optimization problem.

THEOREM 2.1. *Let λ be a simple eigenvalue of $A\in\mathbb{S}$, where $\mathbb{S}$ is a smooth real or complex submanifold of $\mathbb{K}^{n\times n}$ ($\mathbb{K}=\mathbb{R}$ or $\mathbb{C}$). Then for any norm $\|\cdot\|$ on $\mathbb{K}^{n\times n}$ the structured condition number of λ with respect to $\mathbb{S}$ is given by*

$$\kappa(A,\lambda;\mathbb{S}) = \frac{1}{|y^*x|}\max\{|y^*Hx| :\ H\in T_A\mathbb{S},\ \|H\|=1\}, \tag{2.3}$$

where $T_A\mathbb{S}$ is the tangent space of $\mathbb{S}$ at A.

Proof. We show that $\lim_{\epsilon\to 0}\beta_\epsilon = \phi$, where

$$\beta_\epsilon := \sup\left\{\frac{|y^*Ex|}{\epsilon} :\ A+E\in\mathbb{S},\ \|E\|\le\epsilon\right\}, \qquad \epsilon>0,$$
$$\phi := \max\{|y^*Hx| :\ H\in T_A\mathbb{S},\ \|H\|=1\}.$$

Let d denote the real dimension of $\mathbb{S}$. By definition of a smooth submanifold of a finite dimensional vector space there exist open neighborhoods $\mathcal{U}\subset\mathbb{R}^d$ of $0\in\mathbb{R}^d$ and $\mathcal{V}\subset\mathbb{K}^{n\times n}$ of A and a continuously differentiable map $F:\mathcal{U}\to\mathcal{V}$ with the following properties.

(i) $F(\mathcal{U}) = \mathbb{S}\cap\mathcal{V}$.
(ii) F is a homeomorphism between $\mathcal{U}$ and $\mathbb{S}\cap\mathcal{V}$.
(iii) If $D_0F:\mathbb{R}^d\to\mathbb{K}^{n\times n}$ denotes the differential of F at $0\in\mathbb{R}^d$, then
  (a) for all $\xi\in\mathcal{U}$, $F(\xi) = A + D_0F(\xi) + R(\xi)$ and the map $R:\mathcal{U}\to\mathbb{K}^{n\times n}$ satisfies

$$\lim_{\xi\to 0}\|R(\xi)\|/|||\xi||| = 0, \tag{2.4}$$

  where $|||\cdot|||$ is an arbitrary norm on $\mathbb{R}^d$,
  (b) D_0F is an injective linear map, i.e., $0<s:=\min_{|||\xi|||=1}\|D_0F(\xi)\|$,
  (c) $T_A\mathbb{S} = \operatorname{range}(D_0F)$.

A map F with all the properties (i)–(iii) is called a local parametrization of $\mathbb{S}$ at the point A. The neighborhoods $\mathcal{U}$ and $\mathcal{V}$ can be chosen such that

  (d) $\|R(\xi)\|\le\frac{1}{2}s|||\xi|||$ for all $\xi\in\mathcal{U}$.

Suppose now that $A + E \in \mathbb{S}$ and $0 < \|E\| \le \epsilon$. If ϵ is small enough, then, by (i) and (ii), there is a unique nonzero $\xi \in \mathcal{U}$ such that $A+E = F(\xi) = A+D_0F(\xi)+R(\xi)$. Hence

$$E = D_0F(\xi) + R(\xi). \tag{2.5}$$

By (b) and (d),

$$\epsilon \ge \|E\| \ge \|D_0F(\xi)\| - \|R(\xi)\| \ge \frac{s}{2}\,|||\xi|||. \tag{2.6}$$

This implies

$$\frac{\|D_0F(\xi)\|}{\epsilon} \le \frac{\|E\|}{\epsilon} + \frac{\|R(\xi)\|}{\epsilon} \le 1 + \frac{2}{s}\,\frac{\|R(\xi)\|}{|||\xi|||} \tag{2.7}$$

and

$$\frac{|y^*R(\xi)x|}{\epsilon} \le \frac{2}{s}\,\frac{|y^*R(\xi)x|}{|||\xi|||} \le \frac{2\,c}{s}\,\frac{\|R(\xi)\|}{|||\xi|||}, \tag{2.8}$$

where $c := \max\{|y^*Mx| : M \in \mathbb{K}^{n\times n},\ \|M\| = 1\}$. Using (2.5), (2.7), (2.8), and (c) we obtain the estimate

$$\begin{aligned}
\frac{|y^*Ex|}{\epsilon} &\le \frac{|y^*D_0F(\xi)x|}{\epsilon} + \frac{|y^*R(\xi)x|}{\epsilon} \\
&\le \left(1 + \frac{2}{s}\,\frac{\|R(\xi)\|}{|||\xi|||}\right)\frac{|y^*D_0F(\xi)x|}{\|D_0F(\xi)\|} + \frac{2\,c}{s}\,\frac{\|R(\xi)\|}{|||\xi|||} \\
&\le \left(1 + \frac{2}{s}\,\frac{\|R(\xi)\|}{|||\xi|||}\right)\phi + \frac{2\,c}{s}\,\frac{\|R(\xi)\|}{|||\xi|||}. \qquad (2.9)
\end{aligned}$$

The relations (2.4) and (2.9) yield $\lim_{\epsilon\to 0}\beta_\epsilon \le \phi$. In order to show equality let $\widehat{H} \in T_A\mathbb{S}$ be such that $\|\widehat{H}\| = 1$ and $|y^*\widehat{H}x| = \phi$. By (c) there exists a $\hat{\xi} \in \mathbb{R}^d$ with $D_0F(\hat{\xi}) = \widehat{H}$. For $t \ge 0$ let $E_t = D_0F(t\,\hat{\xi}) + R(t\,\hat{\xi})$ and $\epsilon_t = \|E_t\|$. Then $A + E_t = F(t\,\hat{\xi}) \in \mathbb{S}$, $\lim_{t\to 0}\epsilon_t = 0$, and $\lim_{t\to 0}|y^*E_tx|/\epsilon_t = |y^*\widehat{H}x| = \phi$. Thus, $\lim_{\epsilon\to 0}\beta_\epsilon \ge \phi$, and the proof is complete. ∎

It is convenient to introduce the notation

$$\phi(x, y; \mathbb{S}) = \max\Big\{|y^*Ex| :\ E \in \mathbb{S},\ \|E\| = 1\Big\} \tag{2.10}$$

so that (2.3) can be rewritten as

$$\kappa(A, \lambda; \mathbb{S}) = \phi(x, y; T_A\mathbb{S})/|y^*x|. \tag{2.11}$$

In a similar way to [19], an explicit expression for $\kappa(A, \lambda; \mathbb{S})$ can be obtained if one further assumes that the matrix norm $\|\cdot\|$ under consideration is the Frobenius norm $\|\cdot\|_F$. Let us rewrite

$$y^*Ex = \operatorname{vec}(y^*Ex) = (x^T \otimes y^*)\operatorname{vec}(E) = (\overline{x} \otimes y)^*\operatorname{vec}(E),$$

where $\otimes$ denotes the Kronecker product and vec denotes the operator that stacks the columns of a matrix into one long vector [8, p. 180]. Note that $T_A\mathbb{S}$ is a linear vector

space of dimension $m \leq n^2$. Hence, there is an $n^2 \times m$ matrix B such that for every $E \in T_A\mathbb{S}$ there exists a uniquely defined parameter vector p with

$$\operatorname{vec}(E) = Bp, \quad \|E\|_F = \|p\|_2. \tag{2.12}$$

Any matrix B satisfying these properties is called a *pattern matrix* for $T_A\mathbb{S}$; see also [10], [19], and [6]. The relationships in (2.12) together with (2.10) yield

$$\phi_F(x, y; T_A\mathbb{S}) = \max\left\{|(\overline{x} \otimes y)^* Bp| : \ \|p\|_2 = 1, \ p \in \mathbb{K}^m\right\}, \tag{2.13}$$

where $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$. We will use the subscripts F and 2 to refer to the use of the Frobenius and matrix 2-norm in (2.10).

When $\mathbb{K} = \mathbb{C}$ the supremum is taken over all $p \in \mathbb{C}^m$ and consequently, from (2.11),

$$\kappa_F(A, \lambda; \mathbb{S}) = \frac{1}{|y^* x|} \|(\overline{x} \otimes y)^* B\|_2. \tag{2.14}$$

Complications arise if $\mathbb{K} = \mathbb{R}$ but λ is a complex eigenvalue or if B is a complex matrix. In this case, the supremum is also taken over all $p \in \mathbb{R}^m$ but $(\overline{x} \otimes y)^* B$ may be a complex vector. In a similar way as in [5] for the standard eigenvalue condition number we can show that the real structured eigenvalue condition number is within a small factor of the complex one in (2.14). To be more specific,

$$\frac{1}{\sqrt{2}|y^* x|} \|(\overline{x} \otimes y)^* B\|_2 \leq \kappa_F(A, \lambda; \mathbb{S}) \leq \frac{1}{|y^* x|} \|(\overline{x} \otimes y)^* B\|_2; \tag{2.15}$$

see also [9], [17]. To obtain an exact expression for the real structured eigenvalue condition number, let us consider the relation

$$|(\overline{x} \otimes y)^* Bp|^2 = \left| \operatorname{Re}\left((\overline{x} \otimes y)^* B\right) p \right|^2 + \left| \operatorname{Im}\left((\overline{x} \otimes y)^* B\right) p \right|^2,$$

which together with (2.13) implies

$$\kappa_F(A, \lambda; \mathbb{S}) = \frac{1}{|y^* x|} \left\| \begin{bmatrix} \operatorname{Re}\left((\overline{x} \otimes y)^* B\right) \\ \operatorname{Im}\left((\overline{x} \otimes y)^* B\right) \end{bmatrix} \right\|_2. \tag{2.16}$$

For a real pattern matrix B, this formula can be rewritten as

$$\kappa_F(A, \lambda; \mathbb{S}) = \frac{1}{|y^* x|} \|[x_R \otimes y_R + x_I \otimes y_I, \ x_I \otimes y_R - x_R \otimes y_I]^T B\|_2, \tag{2.17}$$

where $x = x_R + \imath x_I$ and $y = y_R + \imath y_I$ with $x_R, x_I, y_R, y_I \in \mathbb{R}^n$. If additionally λ is real, we can choose x and y real and (2.17) reduces to (2.14).

The difficulty in computing (2.14), (2.16), or (2.17) lies in characterizing the tangent space $T_A\mathbb{S}$ and building the pattern matrix B. We show in section 5 how these tasks can be achieved when $\mathbb{S}$ is an automorphism group.

It is difficult to compare the explicit formula for $\kappa_F(A, \lambda; \mathbb{S})$ in (2.14) or (2.16) to that of the standard condition number $\kappa_F(A, \lambda) = 1/|y^* x|$ unless $\mathbb{S}$ has some special structure. Noschese and Pasquini [16] show that for perturbations having an assigned zero structure (or sparsity pattern), (2.14) reduces to

$$\kappa_F(A, \lambda; \mathbb{S}) = \|(yx^*)|_{\mathbb{S}}\|_F / |y^* x|,$$

where $(yx^*)|_{\mathbb{S}}$ means the restriction of the rank-one matrix yx^* to the sparsity structure of $\mathbb{S}$. For example if the perturbation is upper triangular, then $(yx^*)|_{\mathbb{S}}$ is the upper triangular part of yx^*.

Starting from (2.11) we compare in sections 4 and 5 the structured condition number to the unstructured one for structured matrices belonging to the Jordan algebra, Lie algebra, or automorphism group of a scalar product.

3. Structured matrices in scalar product spaces. In this paper a *scalar product* refers to any nondegenerate bilinear or sesquilinear form $\langle \cdot, \cdot \rangle$ on $\mathbb{K}^n$, where $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$. A real or complex bilinear form $\langle \cdot, \cdot \rangle$ has a unique matrix representation given by $\langle \cdot, \cdot \rangle = x^T M y$, while a sesquilinear form can be represented by $\langle \cdot, \cdot \rangle = x^* M y$, where the matrix M is nonsingular. We will denote $\langle \cdot, \cdot \rangle$ by $\langle \cdot, \cdot \rangle_{\mathrm{M}}$ as needed. A bilinear form is symmetric if $\langle x, y \rangle = \langle y, x \rangle$, and skew-symmetric if $\langle x, y \rangle = -\langle y, x \rangle$. Hence for a symmetric form $M = M^T$ and for a skew-symmetric form $M = -M^T$. A sesquilinear form is Hermitian if $\langle x, y \rangle = \overline{\langle y, x \rangle}$ and skew-Hermitian if $\langle x, y \rangle = -\overline{\langle y, x \rangle}$. The matrices associated with such forms are Hermitian and skew-Hermitian, respectively.

The *adjoint* $A^\star$ of $A \in \mathbb{K}^{n \times n}$ with respect to $\langle \cdot, \cdot \rangle_{\mathrm{M}}$ is the unique matrix satisfying

$$\langle Ax, y \rangle_{\mathrm{M}} = \langle x, A^\star y \rangle_{\mathrm{M}} \quad \text{for all } x, y \in \mathbb{K}^n.$$

It can be shown that the adjoint is given explicitly by

$$A^\star = \begin{cases} M^{-1} A^T M & \text{for bilinear forms,} \\ M^{-1} A^* M & \text{for sesquilinear forms.} \end{cases}$$

It is well known [1] that the set of self-adjoint matrices

$$\mathbb{J} = \left\{ S \in \mathbb{K}^{n \times n} : \langle Sx, y \rangle_{\mathrm{M}} = \langle x, Sy \rangle_{\mathrm{M}} \right\} = \left\{ S \in \mathbb{K}^{n \times n} : S^\star = S \right\}$$

forms a Jordan algebra, while the set of skew-adjoint matrices

$$\mathbb{L} = \left\{ L \in \mathbb{K}^{n \times n} : \langle Lx, y \rangle_{\mathrm{M}} = -\langle x, Ly \rangle_{\mathrm{M}} \right\} = \left\{ L \in \mathbb{K}^{n \times n} : L^\star = -L \right\}$$

forms a Lie algebra. The sets $\mathbb{L}$ and $\mathbb{J}$ are linear subspaces, but they are not closed under multiplication. A third class of matrices associated with $\langle \cdot, \cdot \rangle_{\mathrm{M}}$ are those preserving the form, i.e.,

$$\mathbb{G} = \left\{ G \in \mathbb{K}^{n \times n} : \langle Gx, Gy \rangle_{\mathrm{M}} = \langle x, y \rangle_{\mathrm{M}} \right\} = \{ G \in \mathbb{K}^{n \times n} : \ G^\star = G^{-1} \}.$$

They form a Lie group under multiplication. We refer to $\mathbb{G}$ as an automorphism group. Table 3.1 shows a sample of well-known structured matrices in $\mathbb{J}$, $\mathbb{L}$, or $\mathbb{G}$ associated with some scalar products. In the rest of this paper we concentrate on structures belonging to at least one of these three classes.

The eigenvalues of matrices in $\mathbb{J}$, $\mathbb{L}$, and $\mathbb{G}$ have interesting pairing properties as shown by the following theorem.

THEOREM 3.1 ([14, Thms. 7.2 and 7.6]). *Let $A \in \mathbb{L}$ or $A \in \mathbb{J}$. Then the eigenvalues of A occur in pairs as shown below, with the same Jordan structure for each eigenvalue in a pair.*

	Bilinear	Sesquilinear
$A \in \mathbb{J}$	"no pairing"	$\lambda, \overline{\lambda}$
$A \in \mathbb{L}$	$\lambda, -\lambda$	$\lambda, -\overline{\lambda}$
$A \in \mathbb{G}$	$\lambda, 1/\lambda$	$\lambda, 1/\overline{\lambda}$

TABLE 3.1
A sampling of structured matrices associated with scalar products $\langle \cdot, \cdot \rangle_{\mathbf{M}}$, where M is the matrix defining the scalar product.

Space	M	**Automorphism group** $\mathbb{G} = \{G : G^\star = G^{-1}\}$	**Jordan algebra** $\mathbb{J} = \{S : S^\star = S\}$	**Lie algebra** $\mathbb{L} = \{K : K^\star = -K\}$
Bilinear forms				
$\mathbb{R}^n$	I	Real orthogonals	Symmetrics	Skew-symmetrics
$\mathbb{C}^n$	I	Complex orthogonals	Complex symmetrics	Cplx skew-symmetrics
$\mathbb{R}^n$	$\Sigma_{p,q}$	Pseudo-orthogonals	Pseudo-symmetrics	Pseudo–skew-symmetrics
$\mathbb{C}^n$	$\Sigma_{p,q}$	Cplx pseudo-orthogonals	Cplx pseudo-symm.	Cplx pseudo skew-symm.
$\mathbb{R}^n$	R	Real perplectics	Persymmetrics	Perskew-symmetrics
$\mathbb{R}^{2n}$	J	Real symplectics	Skew-Hamiltonians	Hamiltonians
$\mathbb{C}^{2n}$	J	Complex symplectics	Cplx J-skew-symm.	Complex J-symmetrics
Sesquilinear forms				
$\mathbb{C}^n$	I	Unitaries	Hermitian	Skew-Hermitian
$\mathbb{C}^n$	$\Sigma_{p,q}$	Pseudo-unitaries	Pseudo-Hermitian	Pseudo–skew-Hermitian
$\mathbb{C}^{2n}$	J	Conjugate symplectics	J-skew-Hermitian	J-Hermitian

Here, $R = \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix}$ and $\Sigma_{p,q} = \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix} \in \mathbb{R}^{n\times n}$ are symmetric and $J = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}$ is skew-symmetric.

There is no eigenvalue structure property that holds for Jordan algebras of all bilinear forms. However, for certain special classes of $\mathbb{J}$ there may be additional structure in the eigenvalues. For example, it is known that the eigenvalues of any real or complex skew-Hamiltonian matrix all have even multiplicity [7]. More generally we have the following result.

PROPOSITION 3.2 ([14, Prop. 7.7]). *Let $\mathbb{J}$ be the Jordan algebra of any skew-symmetric bilinear form on $\mathbb{K}^n$. Then for any $A \in \mathbb{J}$, the eigenvalues of A all have even multiplicity. Furthermore, all Jordan blocks of a fixed size appear an even number of times.*

Hence we will not consider matrices in these algebras since they cannot have simple eigenvalues.

Many of the results presented in the next two sections require the scalar product defining the structure to be unitary and orthosymmetric: a scalar product $\langle \cdot, \cdot \rangle_{\mathrm{M}}$ is *unitary* if αM is unitary for some $\alpha > 0$; a scalar product is said to be *orthosymmetric* if

$$M = \begin{cases} \beta M^T, \ \beta = \pm 1 & \text{for bilinear forms,} \\ \beta M^*, \ |\beta| = 1 & \text{for sesquilinear forms.} \end{cases}$$

We refer to [14, Definitions A.4 and A.6] for a list of equivalent properties. Note that the classes of structured matrices listed in Table 3.1 are all associated with a scalar product which is both unitary and orthosymmetric with $\alpha = 1$ and $\beta \pm 1$.

Remark 3.3. For a sesquilinear form $\langle x, y \rangle_M = x^* M y$, orthosymmetry means that $M = \beta M^*$ for some $\beta \in \mathbb{C}$ with $|\beta| = 1$. Then the matrix $H := \bar{\beta}^{1/2} M$ is Hermitian

and $\langle x, y\rangle_H = \bar{\beta}^{1/2}\langle x, y\rangle_M$, for all $x, y \in \mathbb{C}^n$. Hence

$$\langle Ax, y\rangle_H = \langle x, Ay\rangle_H \;\Leftrightarrow\; \bar{\beta}^{1/2}\langle Ax, y\rangle_M = \bar{\beta}^{1/2}\langle x, Ay\rangle_M \;\Leftrightarrow\; \langle Ax, y\rangle_M = \langle x, Ay\rangle_M$$

showing that the Jordan algebra of $\langle\cdot,\cdot\rangle_H$ is identical to the Jordan algebra of $\langle\cdot,\cdot\rangle_M$. Similarly the Lie algebras of $\langle\cdot,\cdot\rangle_H$ and $\langle\cdot,\cdot\rangle_M$ are identical. Consequently, results for orthosymmetric sesquilinear forms just need to be established for Hermitian sesquilinear forms.

4. Jordan and Lie algebras. Let $\mathbb{S}$ be the Jordan algebra or Lie algebra of a scalar product on $\mathbb{K}^n$. Since $\mathbb{S}$ is a linear subspace of $\mathbb{K}^{n\times n}$, the tangent space at $A \in \mathbb{S}$ is $\mathbb{S}$ itself. Hence (2.11) becomes

$$\kappa(A, \lambda; \mathbb{S}) = \frac{1}{|y^*x|}\phi(x, y; \mathbb{S}) = \frac{1}{|y^*x|}\max\{|y^*Ex| : \; E \in \mathbb{S}, \; \|E\| = 1\}. \tag{4.1}$$

Clearly, if there exists $E \in \mathbb{S}$ such that $Ex = y$ and $\|E\| = 1$, then $\kappa(A, \lambda; \mathbb{S}) = \kappa(A, \lambda)$. When $\mathbb{S}$ is the Lie or Jordan algebra of an orthosymmetric scalar product, the next theorem gives necessary and sufficient conditions on two given vectors x and b for there to exist $E \in \mathbb{S}$ mapping x to b .

THEOREM 4.1 ([15, Thm. 3.2]). *Let $\mathbb{S}$ be the Lie algebra $\mathbb{L}$ or Jordan algebra $\mathbb{J}$ of an orthosymmetric scalar product $\langle\cdot,\cdot\rangle_M$ on $\mathbb{K}^n$. Then for any given pair of vectors x, $b \in \mathbb{K}^n$ with $x \neq 0$, there exists $E \in \mathbb{S}$ such that $Ex = b$ if and only if the conditions given in the following table hold:*

$\mathbb{S}$	**Bilinear forms**		**Sesquilinear forms**
	Symmetric	Skew-symmetric	Hermitian
$\mathbb{J}$	always	$b^TMx = 0$	$b^*Mx \in \mathbb{R}$
$\mathbb{L}$	$b^TMx = 0$	always	$b^*Mx \in \imath\mathbb{R}$

Mackey, Mackey, and Tisseur show that when the scalar product is both orthosymmetric and unitary and $\mathcal{S} = \{E \in \mathbb{S} : \; Ex = b\} \neq \emptyset$ then $\min_{E\in\mathcal{S}}\|E\|_2 = \|b\|_2/\|x\|_2$ [15, Thm. 5.10]. The minimal 2-norm structured mapping in $\mathcal{S}$ is in general not unique. An explicit characterization of the set $\mathcal{M} = \{E \in \mathcal{S} : \; \|E\|_2 = \min_{A\in\mathcal{S}}\|A\|_2\}$ is given in [15, Thm. 5.10] and it is shown that $\min_{E\in\mathcal{M}}\|E\|_F \leq \sqrt{2}\|b\|_2/\|x\|_2$. The next result follows.

LEMMA 4.2. *Let $\mathbb{S}$ be the Lie or Jordan algebra of a scalar product $\langle\cdot,\cdot\rangle_M$ which is both orthosymmetric and unitary and let $x, b \in \mathbb{K}^n$ of unit 2-norm be such that the relevant condition in Theorem 4.1 is satisfied. Then there exists $E \in \mathbb{S}$ such that $Ex = b$ with $\|E\|_2 = 1$ and $\|E\|_F \leq \sqrt{2}$.*

The next lemma will also be useful when $\mathbb{S} \subset \mathbb{R}^{n\times n}$ is a real algebra but the right and left eigenvectors are complex.

LEMMA 4.3 ([17, Lem. 2.5]). *Let $x \in \mathbb{C}^n$ with $\|x\|_2 = 1$ be given. Then there exists a real symmetric matrix S such that $Sx = \mu\overline{x}$ with $\mu \in \mathbb{C}$, $|\mu| = 1$ and $\|S\|_2 = 1$, $\|E\|_F = \sqrt{2}$.*

4.1. Jordan algebras. Graillat [9] and Rump [17] show that for the structures symmetric, complex symmetric, persymmetric, complex persymmetric, and Hermitian, the structured and unstructured eigenvalue condition numbers are equal for the 2-norm. These are examples of Jordan algebras (see Table 3.1). The next theorem extends these results to all Jordan algebras of a unitary and orthosymmetric scalar

product. Unlike the proofs in [9] and [17], our unifying proof does not need to consider each Jordan algebra individually.

THEOREM 4.4. *Let λ be a simple eigenvalue of $A \in \mathbb{J}$, where $\mathbb{J}$ is the Jordan algebra of an orthosymmetric and unitary scalar product $\langle \cdot, \cdot \rangle_{\mathrm{M}}$ on $\mathbb{K}^n$. Then, for the 2-norm,*

$$\kappa_2(A, \lambda; \mathbb{J}) = \kappa_2(A, \lambda).$$

Proof. Since the scalar product $\langle \cdot, \cdot \rangle_{\mathrm{M}}$ is unitary, αM is unitary for some $\alpha > 0$. Let x and y be right and left eigenvectors of A associated with λ normalized so that $\|x\|_2 = \|y\|_2 = 1$. From (4.1) and since $\phi_2(x, y; \lambda) \leq 1$, we just need to find $E \in \mathbb{J}$ of unit 2-norm such that $|y^* E x| = 1$.

For bilinear forms, orthosymmetry of $\langle \cdot, \cdot \rangle_{\mathrm{M}}$ means that $M = \pm M^T$. Suppose first that $M = M^T$, that is, the bilinear form is symmetric. When $\mathbb{K} = \mathbb{C}$, Lemma 4.2 says that there exists $E \in \mathbb{J}$ such that $Ex = y$ and $\|E\|_2 = 1$. Hence $|y^* E x| = |y^* y| = 1$.

When $\mathbb{K} = \mathbb{R}$, A is real but if λ is complex, then $x, y \in \mathbb{C}^n$ and we cannot use Lemma 4.2 to say that there exists a real $E \in \mathbb{J}$ of unit 2-norm sending x to y. However, $A \in \mathbb{J}$ implies $A = A^\star = M^{-1} A^T M$ so that

$$Ax = \lambda x \iff \overline{x}^* A^T = \lambda \overline{x}^* \iff \overline{x}^* M A = \lambda \overline{x}^* M$$

so that we can take $y = (\alpha M)^* \overline{x}$ as a normalized left eigenvector for A associated with λ. From Lemma 4.3 we know there exists a real symmetric S such that $Sx = \mu \overline{x}$, $|\mu| = 1$, and $\|S\|_2 = 1$. Let $E = \alpha M S \in \mathbb{R}^{n \times n}$. Since αM is real orthogonal and $M = M^T$ we have

$$E^\star = M^{-1} E^T M = (\alpha M)^{-1} S (\alpha M)^T (\alpha M) = \alpha M S = E$$

showing that $E \in \mathbb{J}$. Moreover $\|E\|_2 = \|\alpha M S\|_2 = \|S\|_2 = 1$ and $Ex = \alpha M S x = \mu \alpha M \overline{x}$ so that $|y^* E x| = |\mu x^T (\alpha M)^T (\alpha M) \overline{x}| = |\mu x^T \overline{x}| = 1$.

We do not need to consider the skew-symmetric bilinear case ($M = -M^T$) since from Proposition 3.2 the eigenvalues of matrices in Jordan algebras of skew-symmetric bilinear forms all have even multiplicity.

When $\langle \cdot, \cdot \rangle$ is an orthosymmetric sesquilinear form, Remark 3.3 says that we just need to establish the result for $M = M^*$, that is, for Hermitian sesquilinear forms. Let $\mu \in \mathbb{C}$, $|\mu| = 1$ be such that $(\mu y)^* M x \in \mathbb{R}$. Then from Lemma 4.2 there exists $E \in \mathbb{J}$ such that $Ex = \mu y$ and $\|E\|_2 = 1$. ∎

The proof above also shows that for the Frobenius norm,

$$\frac{1}{\sqrt{2}} \kappa_F(A, \lambda) \leq \kappa_F(A, \lambda; \mathbb{J}) \leq \kappa_F(A, \lambda).$$

For Jordan algebras $\mathbb{J}$ of sesquilinear forms, eigenvalues come in pairs λ and $\bar{\lambda}$ and if λ is simple so is $\bar{\lambda}$ (see Theorem 3.1). For unitary scalar products, αM is unitary for some $\alpha > 0$, and, if x and y are normalized right and left eigenvectors associated with λ, then $\alpha M y$ and $\alpha M x$ are normalized right and left eigenvectors associated with $\bar{\lambda}$. Hence, $|(\alpha M x)^* (\alpha M y)| = |x^* y|$ so that

$$\kappa(A, \lambda; \mathbb{J}) = \kappa(A, \bar{\lambda}; \mathbb{J}).$$

4.2. Lie algebras. We show that, with the exception of symmetric bilinear forms, incorporating structure does not affect the eigenvalue condition number for matrices in Lie algebras of scalar products that are both orthosymmetric and unitary. These include as special cases the skew-symmetric, complex skew-symmetric, and skew-Hermitian matrices considered by Rump [17].

THEOREM 4.5. *Let λ be a simple eigenvalue of $A \in \mathbb{L}$, where $\mathbb{L}$ is the Lie algebra of an orthosymmetric and unitary scalar product $\langle \cdot, \cdot \rangle_{\mathrm{M}}$ on $\mathbb{C}^n$.*

- *For symmetric bilinear forms,*

$$\kappa_2(A, \lambda; \mathbb{L}) = \left(\max_{\substack{b \in (\overline{Mx})^\perp \\ \|b\|_2 = 1}} |y^* b| \right) \kappa_2(A, \lambda),$$

- *For skew-symmetric bilinear forms or sesquilinear forms,*

$$\kappa_2(A, \lambda; \mathbb{L}) = \kappa_2(A, \lambda).$$

Proof. Since the scalar product $\langle \cdot, \cdot \rangle_{\mathrm{M}}$ is unitary, αM is unitary for some $\alpha > 0$. Let x and y be right and left eigenvectors of A associated with λ normalized so that $\|x\|_2 = \|y\|_2 = 1$.

For bilinear forms, orthosymmetry implies $M = \pm M^T$. Suppose first that $M = M^T$, that is, $\langle \cdot, \cdot \rangle_{\mathrm{M}}$ is a symmetric bilinear form. From (4.1) we just need to show that

$$\eta := \max \left\{ |y^* b| : \ b \in (\overline{Mx})^\perp, \ \|b\|_2 = 1 \right\}$$

is equal to $\phi_2(x, y; \mathbb{L})$. Let $E \in \mathbb{L}$ be of unit 2-norm and such that $|y^* E x| = \phi_2(x, y; \mathbb{L})$. Let $b = Ex$. Theorem 4.1 implies that $b^T M x = 0$, i.e., $b \in (\overline{Mx})^\perp$. Also, $\|b\|_2 = \|Ex\|_2 \le 1$. Hence $\phi_2(x, y; \mathbb{L}) \le \eta$. Let $b \in (\overline{Mx})^\perp$ be of unit 2-norm and such that $|y^* b| = \eta$. Lemma 4.2 then implies that there exists $E \in \mathbb{L}$ such that $Ex = b$ and $\|E\|_2 = 1$. Hence $\phi_2(x, y; \mathbb{L}) \ge |y^* E x| = |y^* b| = \eta$.

Now for skew-symmetric bilinear forms, Lemma 4.2 implies that there exists $E \in \mathbb{L}$ such that $Ex = y$ and $\|E\|_2 = 1$ so that $|y^* E x| = |y^* y| = 1$ and equality between the structured and unstructured eigenvalue condition numbers follows.

Finally when $\langle \cdot, \cdot \rangle_{\mathrm{M}}$ is an orthosymmetric sesquilinear form, Remark 3.3 says that we just need to prove the result for an Hermitian sesquilinear form ($M = M^*$). Let $\mu \in \mathbb{C}$, $|\mu| = 1$ be such that $\langle \mu y, x \rangle_{\mathrm{M}} = \bar{\mu} y^* M x \in \imath \mathbb{R}$. Then from Lemma 4.2 there exists $E \in \mathbb{L}$ such that $Ex = \mu y$ and $\|E\|_2 = 1$. Hence $|y^* E x| = |\mu y^* y| = 1$. The result follows then from (4.1). ∎

With a very similar proof we can show that for Lie algebras of orthosymmetric and unitary scalar products and for perturbations measured in the Frobenius norm,

$$\frac{1}{\sqrt{2}} \, \gamma_{\mathbb{L}} \, \kappa_F(A, \lambda) \le \kappa_F(A, \lambda; \mathbb{L}) \le \gamma_{\mathbb{L}} \, \kappa_F(A, \lambda),$$

where $\gamma_{\mathbb{L}} = \max_{\substack{b \in (\overline{Mx})^\perp \\ \|b\|_2 = 1}} |y^* b|$ for symmetric bilinear forms and $\gamma_{\mathbb{L}} = 1$ otherwise.

Note that Theorem 4.5 deals with complex perturbations only. However, for real bilinear forms the results still hold when λ is real. For complex λ, in view of (2.15) we know that the real structured eigenvalue condition number is within a small factor of the complex one.

Now suppose that $\langle \cdot, \cdot \rangle_{\mathrm{M}}$ is symmetric bilinear. For $A \in \mathbb{L}$ we have $A^\star = -A$ and

$$\lambda \langle x, x \rangle_{\mathrm{M}} = \langle \lambda x, x \rangle_{\mathrm{M}} = \langle Ax, x \rangle_{\mathrm{M}} = \langle x, A^\star x \rangle_{\mathrm{M}} = -\langle x, Ax \rangle_{\mathrm{M}} = -\lambda \langle x, x \rangle_{\mathrm{M}}$$

so that if $\lambda \neq 0$, $\langle x, x\rangle_{\mathrm{M}} = (Mx)^T x = 0$, that is, $x \in (\overline{Mx})^{\perp}$. Hence for $\lambda \neq 0$,

$$|y^* x| \leq \max_{\substack{b \in (\overline{Mx})^{\perp} \\ \|b\|_2 = 1}} |y^* b| \leq 1.$$

When $\lambda = 0$ is an eigenvalue of $A \in \mathbb{L}$,

$$Ax = 0 \iff -A^{\star}x = 0 \iff M^{-1}A^T Mx = 0 \iff (Mx)^T A = 0$$

so that we can take $y = \overline{Mx}$ as a left eigenvector of $\lambda = 0$. Hence if $\lambda = 0$ is simple,

$$\kappa_2(A, 0; \mathbb{L}) = 0 < \kappa_2(A, 0).$$

This result may be surprising but from Theorem 3.1 we know that eigenvalues of Lie algebras of bilinear forms come in pairs $\lambda, -\lambda$ so that for odd dimensions n, $\lambda = 0$ has to be an eigenvalue. Any perturbation of A leaves a simple 0 eigenvalue unchanged. For the special case where $M = I$, i.e., when $\mathbb{L}$ is the set of complex skew-symmetric matrices, Rump [17] exhibits a 3×3 example showing that the ratio $\kappa_2(A, \lambda; \mathbb{L})/\kappa_2(A, \lambda)$ for $\lambda \neq 0$ can be arbitrarily small. Our result shows that this ratio can be arbitrarily small for all Lie algebras of symmetric bilinear forms on $\mathbb{K}^n$.

Since the eigenvalues of matrices in $\mathbb{L}$ come in pairs $\lambda, -\lambda$ for bilinear forms and $\lambda, -\bar{\lambda}$ for sesquilinear forms (see Theorem 3.1) then if $0 \neq \lambda$ is simple so is $-\lambda$ (or $-\bar{\lambda}$). We can show that for unitary scalar products,

$$\kappa(A, \lambda; \mathbb{L}) = \begin{cases} \kappa(A, -\lambda; \mathbb{L}) & \text{for bilinear forms,} \\ \kappa(A, -\bar{\lambda}; \mathbb{L}) & \text{for sesquilinear forms.} \end{cases}$$

5. Automorphism groups. We now consider structured condition numbers for automorphism groups $\mathbb{G}$ associated with the scalar product $\langle \cdot, \cdot \rangle_{\mathrm{M}}$,

$$\mathbb{G} = \{A \in \mathbb{K}^{n \times n} : A^{\star} = A^{-1}\}.$$

This includes the groups of symplectic matrices ($M = J$), real and complex orthogonal matrices ($M = I$), as well as Lorentz transformations ($M = \mathrm{diag}(1, 1, 1, -1)$). We first show how to compute $\kappa_F(A, \lambda; \mathbb{G})$ in (2.14) and (2.16), then consider properties of the structured condition number, and finally provide lower bounds for $\kappa_2(A, \lambda; \mathbb{G})$.

5.1. Computation of $\kappa_F(A, \lambda; \mathbb{G})$. An automorphism group $\mathbb{G}$ forms a smooth manifold. The Jacobian of the function

$$\Phi(A) = \begin{cases} A^T MA - M & \text{for bilinear forms,} \\ A^* MA - M & \text{for sesquilinear forms} \end{cases}$$

at $A \in \mathbb{K}^{n \times n}$ can be represented as the linear function

$$J_A(X) = \begin{cases} A^T MX + X^T MA & \text{for bilinear forms,} \\ A^* MX + X^* MA & \text{for sesquilinear forms.} \end{cases}$$

The tangent space $T_A\mathbb{G}$ at $A \in \mathbb{G}$ coincides with the kernel of this Jacobian,

$$T_A\mathbb{G} = \{X \in \mathbb{K}^{n \times n} : \ J_A(X) = 0\} = \{AH \in \mathbb{K}^{n \times n} : \ H^{\star} = -H\} = A \cdot \mathbb{L}, \tag{5.1}$$

where $\mathbb{L}$ is the Lie algebra of $\langle \cdot, \cdot \rangle_{\mathrm{M}}$.

TABLE 5.1

Pattern matrices L_{M} for $M\cdot\mathbb{L} = \mathrm{Sym}(\mathbb{K})$, $\mathrm{Skew}(\mathbb{K})$, or $\mathrm{Herm}(\mathbb{C})$. L_{M} is such that for any $H \in M\cdot\mathbb{L}$ there exists a uniquely defined parameter vector q with $\mathrm{vec}(H) = L_{\mathrm{M}}q$, $\|H\|_F = \|q\|_2$. Here $n = 2$.

$M\cdot\mathbb{L}$	$\mathrm{Sym}(\mathbb{K})$	$\mathrm{Skew}(\mathbb{K})$	$\mathrm{Herm}(\mathbb{C})$
L_{M}	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & 0 \\ 0 & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1/\sqrt{2} \\ -1/\sqrt{2} \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/\sqrt{2} & -\imath/\sqrt{2} & 0 \\ 0 & 1/\sqrt{2} & \imath/\sqrt{2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

As the Lie algebra $\mathbb{L}$ in (5.1) is independent of A, it is often simple to explicitly construct a pattern matrix L such that for every $H \in \mathbb{L}$ there exists a uniquely defined parameter vector q with $\mathrm{vec}(H) = Lq$. To obtain a pattern matrix B for $A\cdot\mathbb{L}$ in the sense of (2.12), we can compute a QR decomposition $(I \otimes A)L = BR$, where the columns of B form an orthonormal basis for the space spanned by the columns of L, and R is an upper triangular matrix. Hence,

$$\mathrm{vec}(AH) = (I \otimes A)\,\mathrm{vec}(H) = (I \otimes A)Lq = Bp,$$

where $p = Rq$, and $\|AH\|_F = \|\,\mathrm{vec}(AH)\|_2 = \|p\|_2$.

According to (2.14) we have

$$\kappa_F(A,\lambda;\mathbb{G}) = \frac{1}{|y^*x|}\|(\overline{x}\otimes y)^*B\|_2 = \frac{|\lambda|}{|y^*x|}\|(\overline{x}\otimes y)^*LR^{-1}\|_2 \tag{5.2}$$

if $\mathbb{K} = \mathbb{C}$ or if $\mathbb{K} = \mathbb{R}$ with λ real. Otherwise, when $\mathbb{K} = \mathbb{R}$ and λ is complex or, when B is complex, (2.16) implies that

$$\kappa_F(A,\lambda;\mathbb{G}) = \frac{1}{|y^*x|}\left\|\begin{bmatrix}\mathrm{Re}\left(\lambda(\overline{x}\otimes y)^*LR^{-1}\right)\\ \mathrm{Im}\left(\lambda(\overline{x}\otimes y)^*LR^{-1}\right)\end{bmatrix}\right\|_2. \tag{5.3}$$

It is shown in [15, Lem. 5.9] that when the scalar product $\langle\cdot,\cdot\rangle_{\mathrm{M}}$ defining the structure is orthosymmetric, left multiplication by M is a bijection from $\mathbb{K}^{n\times n}$ to $\mathbb{K}^{n\times n}$ that maps $\mathbb{L}$ and $\mathbb{J}$ to $\mathrm{Skew}(\mathbb{K})$ and $\mathrm{Sym}(\mathbb{K})$ for bilinear forms and a scalar multiple of $\mathrm{Herm}(\mathbb{C})$ for sesquilinear forms, where

$$\mathrm{Skew}(\mathbb{K}) = \{A \in \mathbb{K}^{n\times n} : A^T = -A\}, \quad \mathrm{Sym}(\mathbb{K}) = \{A \in \mathbb{K}^{n\times n} : A^T = A\}$$

are the sets of symmetric and skew-symmetric matrices on $\mathbb{K}^{n\times n}$ and $\mathrm{Herm}(\mathbb{C})$ is the set of Hermitian matrices. More precisely, for bilinear forms on $\mathbb{K}^n$, $(\mathbb{K} = \mathbb{R}, \mathbb{C})$ write,

$$M\cdot\mathbb{L} = \begin{cases}\mathrm{Skew}(\mathbb{K}) & \text{if } M = M^T,\\ \mathrm{Sym}(\mathbb{K}) & \text{if } M = -M^T,\end{cases} \tag{5.4}$$

and for sesquilinear forms on $\mathbb{C}^n$,

$$M\cdot\mathbb{L} = \beta^{1/2}\,\imath\,\mathrm{Herm}(\mathbb{C}), \tag{5.5}$$

where, by orthosymmetry, β is such that $M = \beta M^*$, $|\beta| = 1$. For any $H \in \mathbb{L}$, $MH \in M\cdot\mathbb{L}$ and if L_{M} is pattern matrix for $M\cdot\mathbb{L}$, that is, $\mathrm{vec}(MH) = L_{\mathrm{M}}q$ where q is a uniquely defined vector of parameters, then

$$\mathrm{vec}(H) = \mathrm{vec}(M^{-1}MH) = (I \otimes M^{-1})\,\mathrm{vec}(MH) = (I \otimes M^{-1})L_{\mathrm{M}}q$$

so that $L := (I \otimes M^{-1})L_{\mathrm{M}}$ is a pattern matrix for $\mathbb{L}$. An advantage of using left multiplication by M is that pattern matrices for $\mathrm{Sym}(\mathbb{K})$, $\mathrm{Skew}(\mathbb{K})$, and $\mathrm{Herm}(\mathbb{C})$ are easy to construct (see Table 5.1 for examples of such matrices).

5.2. Properties of $\kappa(A,\lambda;\mathbb{G})$. The eigenvalues of $A \in \mathbb{G}$ come in pairs λ and $1/\lambda$ for bilinear forms, and in pairs λ and $1/\bar{\lambda}$ for sesquilinear forms. In both cases these pairs have the same Jordan structure, and hence the same algebraic and geometric multiplicities (see Theorem 3.1). Hence if λ is simple so is $1/\lambda$ or $1/\bar{\lambda}$. For unitary scalar products, there are interesting relations between the structured condition numbers of these eigenvalue pairings.

THEOREM 5.1. *Let λ be a simple eigenvalue of $A \in \mathbb{G}$, where $\mathbb{G}$ is the automorphism group of a unitary scalar product on $\mathbb{K}^n$. For any unitarily invariant norm, the (absolute) unstructured eigenvalue condition number satisfies*

$$\kappa(A,\lambda) = \begin{cases} \kappa(A,1/\lambda) & \text{for bilinear forms,} \\ \kappa(A,1/\bar{\lambda}) & \text{for sesquilinear forms,} \end{cases}$$

whereas the (absolute) structured eigenvalue condition number satisfies

$$\kappa(A,\lambda;\mathbb{G}) = \begin{cases} |\lambda|^2\,\kappa(A,1/\lambda;\mathbb{G}) & \text{for bilinear forms,} \\ |\lambda|^2\,\kappa(A,1/\bar{\lambda};\mathbb{G}) & \text{for sesquilinear forms.} \end{cases}$$

Proof. We just prove the bilinear case, the proof for the sesquilinear case being similar. The scalar product $\langle\cdot,\cdot\rangle_{\mathrm{M}}$ being unitary implies that αM is unitary for some $\alpha > 0$. If x and y are normalized right and left eigenvectors associated with λ, then $\widetilde{x} = \alpha\overline{My}$ and $\widetilde{y} = \alpha\overline{Mx}$ are right and left normalized eigenvectors belonging to the eigenvalue $1/\lambda$. It is easily checked that $|\widetilde{y}^*\widetilde{x}| = |y^*x|$, and since $\|\cdot\|$ is unitarily invariant, $\phi(\widetilde{x},\widetilde{y};\mathbb{K}^{n\times n}) = \phi(x,y;\mathbb{K}^{n\times n})$ so that $\kappa(A,\lambda) = \kappa(A,1/\lambda)$.

Let $E \in T_A\mathbb{G} = A\cdot\mathbb{L}$. Then $E = AH$ for some H in the Lie algebra $\mathbb{L}$ of $\langle\cdot,\cdot\rangle_{\mathrm{M}}$ and

$$|y^*Ex| = |\lambda|\,|y^*Hx|. \tag{5.6}$$

Also, $A \in \mathbb{G} \Rightarrow M^TA = A^{-T}M^T$, αM unitary $\Rightarrow M^{-T} = \alpha^2\overline{M}$, and $H \in \mathbb{L} \Rightarrow \alpha^2 M^T H\overline{M} = -H^T$. Hence,

$$\begin{aligned} |(\alpha\overline{Mx})^*E(\alpha\overline{My})| &= |\alpha^2 x^T M^T AH\overline{My}| \\ &= |\alpha^2 (x^TA^{-T})(M^TH\overline{M})\overline{y}| \\ &= \frac{1}{|\lambda|}|x^TH^T\overline{y}| \\ &= \frac{1}{|\lambda|}|y^*Hx| = \frac{1}{|\lambda|^2}|y^*Ex| \end{aligned}$$

so that from (2.10) and (2.11), $\kappa(A,\lambda;\mathbb{G}) = \kappa(A,1/\lambda;\mathbb{G})/|\lambda|^2$. ∎

Theorem 5.1 shows that the *relative structured* eigenvalue condition numbers for λ and $1/\lambda$ if the form is bilinear or λ and $1/\bar{\lambda}$ if the form is sesquilinear, are equal. On the other hand, the ratio between the *relative unstructured* eigenvalue condition numbers for λ and $1/\lambda$ (or λ and $1/\bar{\lambda}$) is $1/|\lambda|^2$. Hence, if we use a non–structure-preserving algorithm, we should compute the larger of λ and $1/\lambda$ (or $1/\bar{\lambda}$). In other words, we should compute whichever member of the pair $(\lambda, 1/\lambda)$ (or the pair $(\lambda, 1/\bar{\lambda})$) lies outside the unit circle and then obtain the other one by reciprocation.

5.3. Bounds for $\kappa(A,\lambda;\mathbb{G})$. Lower bounds for the eigenvalue structured condition number can be derived when $\langle\cdot,\cdot\rangle_{\mathrm{M}}$ is orthosymmetric and unitary.

THEOREM 5.2. *Let λ be a simple eigenvalue of $A \in \mathbb{G}$, where $\mathbb{G}$ is the automorphism group of an orthosymmetric and unitary scalar product $\langle \cdot, \cdot \rangle_{\mathrm{M}}$ on $\mathbb{K}^n$. If $\mathbb{K} = \mathbb{C}$ or, if $\mathbb{K} = \mathbb{R}$ with λ real we have for both the Frobenius norm and the 2-norm ($\nu = 2, F$),*

- *for symmetric bilinear forms,*

$$\frac{|\lambda|}{\|A\|_2} \max_{\substack{b \in (\overline{Mx})^\perp \\ \|b\|_2 = 1}} |y^* b| \, \kappa_\nu(A, \lambda) \leq \kappa_\nu(A, \lambda; \mathbb{G}) \leq \max_{\substack{b \in (\overline{Mx})^\perp \\ \|b\|_2 = 1}} |y^* b| \, \kappa_\nu(A, \lambda),$$

- *for skew-symmetric bilinear or sesquilinear forms,*

$$\frac{|\lambda|}{\|A\|_2} \, \kappa_\nu(A, \lambda) \leq \kappa_\nu(A, \lambda; \mathbb{G}) \leq \, \kappa_\nu(A, \lambda).$$

For $\mathbb{K} = \mathbb{R}$ and λ complex, the lower bounds for the Frobenius norm need to be multiplied by $1/\sqrt{2}$.

Proof. Let x and y be right and left eigenvectors of A associated with λ normalized so that $\|x\|_2 = \|y\|_2 = 1$. Let $\mathbb{L}$ be the Lie algebra of $\langle \cdot, \cdot \rangle_{\mathrm{M}}$. From (2.11) and (5.1) we have

$$\kappa(A, \lambda; \mathbb{G}) = \frac{1}{|y^* x|} \phi(x, y; A \cdot \mathbb{L}) = \frac{1}{|y^* x|} \max \{ |y^* A H x| : \; H \in \mathbb{L}, \; \|AH\| = 1 \}.$$

By definition of orthosymmetry and from Remark 3.3 we just need to prove the result for symmetric and skew-symmetric bilinear forms and for Hermitian sesquilinear forms.

Suppose first that $\langle \cdot, \cdot \rangle_{\mathrm{M}}$ is a symmetric bilinear form on $\mathbb{K}^n$. Let $H_\nu \in \mathbb{L}$ be such that $\|AH_\nu\|_\nu = 1$ and $|y^* A H_\nu x| = \phi_\nu(x, y; A \cdot \mathbb{L})$, $\nu = 2, F$. Let $b_\nu = AH_\nu x$. Theorem 4.1 implies that $(A^{-1} b_\nu)^T M x = 0$. Since $M = M^T$ and $A \in \mathbb{G}$, that is, $A^{-1} = A^\star = M^{-1} A^T M$, we have

$$(A^{-1} b_\nu)^T M x = 0 \Longleftrightarrow b_\nu^T M A M^{-1} M x = \lambda \, b_\nu^T M x = 0$$

so that $b_\nu \in (\overline{Mx})^\perp$. Also, $\|b_\nu\|_2 = \|AH_\nu x\|_2 \leq 1$. Hence

$$\phi_\nu(x, y; A \cdot \mathbb{L}) = |y^* A H_\nu x| = |y^* b_\nu| \leq \max \big\{ |y^* b| : \; b \in (\overline{Mx})^\perp, \|b\|_2 = 1 \big\},$$

which proves the upper bound. For the lower bound we take $v \in (\overline{Mx})^\perp$ of unit 2-norm and such that $|y^* v| = \max \big\{ |y^* b| : \; b \in (\overline{Mx})^\perp, \|b\|_2 = 1 \big\}$. From Lemma 4.2 there exists $S \in \mathbb{L}$ such that $Sx = v$ and $\|S\|_2 = 1$. Let $\widetilde{H}_\nu = \xi_\nu S$ with $\xi_\nu > 0$ such that $\|A \widetilde{H}_\nu\|_\nu = 1$, $\nu = 2, F$. From $\|A \widetilde{H}_\nu\|_\nu \leq \|A\|_\nu \|\widetilde{H}_\nu\|_2$ we have that $\xi_\nu \geq 1/\|A\|_\nu$. Hence

$$\begin{aligned} \phi_\nu(x, y; A \cdot \mathbb{L}) &= |\lambda| \max \{ |y^* H x| : \; H \in \mathbb{L}, \; \|AH\|_\nu = 1 \} \\ &\geq |\lambda| |y^* \widetilde{H}_\nu x| \\ &\geq \frac{|\lambda|}{\|A\|_\nu} |y^* v| \\ &= \frac{|\lambda|}{\|A\|_\nu} \max \big\{ |y^* b| : \; b \in (\overline{Mx})^\perp, \|b\|_2 = 1 \big\} \end{aligned}$$

proving the lower bound.

The lower bound for the skew-symmetric bilinear or Hermitian sesquilinear cases is derived in a similar way to that for the symmetric bilinear case. The only difference being that, from Lemma 4.2, there exists $S \in \mathbb{L}$ of unit 2-norm such that $Sx = y$ if the form is skew-symmetric bilinear and $Sx = \mu y$ for some $\mu \in \mathbb{C}$ such that $(\mu y)^* Mx \in \imath\mathbb{R}$, $|\mu| = 1$ when the form is Hermitian sesquilinear. ∎

Note that $A \in \mathbb{G}$ implies

$$\lambda \langle x, x \rangle_{\scriptscriptstyle M} = \langle Ax, x \rangle_{\scriptscriptstyle M} = \langle x, A^{-1}x \rangle_{\scriptscriptstyle M} = \frac{1}{\lambda} \langle x, x \rangle_{\scriptscriptstyle M}. \tag{5.7}$$

Hence if $\lambda \neq \pm 1$ we have, for bilinear forms, $\langle x, x \rangle_{\scriptscriptstyle M} = x^T Mx = 0$, that is, $x \in (\overline{Mx})^{\perp}$ so that

$$|y^* x| \leq \max_{\substack{b \in (\overline{Mx})^{\perp} \\ \|b\|_2 = 1}} |y^* b| \leq 1, \qquad \lambda \neq \pm 1. \tag{5.8}$$

If $\lambda = \pm 1$, then

$$Ax = \pm x \Leftrightarrow x = \pm A^{-1}x \Leftrightarrow x = \pm A^{\star}x \Leftrightarrow Mx = A^T\, Mx \Leftrightarrow (\overline{Mx})^* = \pm (\overline{Mx})^* A$$

so that $y = \overline{Mx}$ is a left eigenvector of A associated with λ. If $M = M^T$, then Theorem 5.2 implies that for both the 2-norm and Frobenius norm,

$$\kappa_{\nu}(A, \lambda; \mathbb{G}) = 0 \quad \text{ for } \lambda = \pm 1. \tag{5.9}$$

When $M = I$ and $\langle \cdot, \cdot \rangle$ is a sesquilinear form, $\mathbb{G}$ is the set of unitary matrices (see Table 3.1). But unitary matrices are normal and therefore $\kappa_{\nu}(A, \lambda) = 1$, $\nu = 2, F$. Thus we can expect $\kappa_{\nu}(A, \lambda; \mathbb{G}) \leq 1$. Theorem 5.2 implies that the structured condition number is exactly 1. If $\langle \cdot, \cdot \rangle_M$ with $M = I$ is a real (symmetric) bilinear form, $\mathbb{G}$ is the set of orthogonal matrices. Theorem 5.2 combined with (5.8) and (5.9) says that $\kappa_{\nu}(A, \lambda; \mathbb{G}) = 0$ if $\lambda = \pm 1$ and $\kappa_{\nu}(A, \lambda; \mathbb{G}) = 1$ otherwise. We refer to [3] for a more general perturbation analysis of orthogonal and unitary eigenvalue problems, based on the Cayley transform.

Suppose $\mathbb{G}$ is the automorphism group of a skew-symmetric bilinear form $\langle \cdot, \cdot \rangle_{\scriptscriptstyle M}$ ($M = -M^T$). For an eigenvalue λ of A with $|\lambda| \approx \|A\|_2$, the bounds in Theorem 5.2 imply

$$\kappa_{\nu}(A, \lambda; \mathbb{G}) \approx \kappa_{\nu}(A, \lambda), \quad \nu = 2, F.$$

From Theorem 5.1 we then have

$$|\lambda|^2 \kappa_{\nu}(A, 1/\lambda; \mathbb{G}) \approx \kappa_{\nu}(A, 1/\lambda), \quad \nu = 2, F$$

showing that if $|\lambda|$ is large, the unstructured eigenvalue condition number for $1/\lambda$ is much larger than the structured one. The lower bounds in Theorem 5.2 may not be tight when $\max(|\lambda|, 1/|\lambda|) \ll \|A\|_{\nu}$ as shown by the following example. Suppose that $M = J$ and that $\langle \cdot, \cdot \rangle_{\scriptscriptstyle J}$ is a real bilinear form ($\mathbb{K} = \mathbb{R}$). Then $\mathbb{G}$ is the set of real symplectic matrices (see Table 3.1). Let us consider the symplectic matrix

$$A = \begin{bmatrix} D & D \\ 0 & D^{-1} \end{bmatrix}, \quad D = \operatorname{diag}(10^4, 10^2, 2). \tag{5.10}$$

Define the ratio

$$\rho = \kappa_F(A, \lambda; \mathbb{G}) / \kappa_F(A, \lambda) \leq 1$$

TABLE 5.2
Condition numbers for the eigenvalues of the symplectic matrix A in (5.10), ratio ρ between the structured and unstructured condition number, and lower bound γ for this ratio.

λ	10^4	10^2	2	1/2	10^{-2}	10^{-4}
$\kappa_F(A,\lambda;\mathbb{G})$	1.2	1.2	1.5	0.4	1.2×10^{-4}	1.2×10^{-8}
ρ	0.87	0.87	0.89	0.22	8.7×10^{-5}	8.7×10^{-9}
γ	0.5	5×10^{-3}	1×10^{-4}	2.5×10^{-5}	5×10^{-7}	5×10^{-9}

between the structured and unstructured eigenvalue condition numbers. $\kappa_F(A,\lambda;\mathbb{G})$ is computed using (5.2) and its values and these of ρ are displayed in Table 5.2 together with the lower bound $\gamma = |\lambda|/(\sqrt{2}\|A\|_2)$ of Theorem 5.2. This example demonstrates the looseness of the bounds of Theorem 5.2 for eigenvalues in the interior of the spectrum. Hence for these eigenvalues the computable expressions in section 5.1 are of interest.

6. Conclusions. We have derived directly computable expressions for structured eigenvalue condition numbers on a smooth manifold of structured matrices. Furthermore, we have obtained meaningful bounds on the ratios between the structured and unstructured eigenvalue condition numbers for a number of structures related to Jordan algebras, Lie algebras, and automorphism groups. We have identified classes of structured matrices for which this ratio is 1 or close to 1. Hence for these structures, the usual unstructured perturbation analysis is sufficient.

The important task of finding computable expressions for structured backward errors of nonlinearly structured eigenvalue problems is still largely open and remains to be addressed.

Acknowledgments. The authors sincerely thank Nick Higham, Siegfried Rump, and Ji-guang Sun for helpful remarks and discussions on earlier versions of this paper. The hospitality of the Department of Computing Science, Umeå University, where the second author was staying during the work on this paper, is gratefully acknowledged.

REFERENCES

[1] E. ARTIN, *Geometric Algebra*, Interscience Tracts in Pure and Applied Mathematics, Interscience Publishers Inc., New York, 1957.
[2] 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.
[3] B. BOHNHORST, A. BUNSE-GERSTNER, AND H. FASSBENDER, *On the perturbation theory for unitary eigenvalue problems*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 809–824.
[4] 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.
[5] R. BYERS AND D. KRESSNER, *On the condition of a complex eigenvalue under real perturbations*, BIT, 44 (2004), pp. 209–214.
[6] P. I. DAVIES, *Structured conditioning of matrix functions*, Electron. J. Linear Algebra, 11 (2004), pp. 132–161.
[7] H. FASSBENDER, D. S. MACKEY, N. MACKEY, AND H. XU, *Hamiltonian square roots of skew-Hamiltonian matrices*, Linear Algebra Appl., 287 (1999), pp. 125–159.
[8] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[9] S. GRAILLAT, *Structured Condition Number and Backward Error for Eigenvalue Problems*, Research Report RR2005-01, LP2A, University of Perpignan, Perpignan, France, 2005.

[10] D. J. Higham and N. J. Higham, *Backward error and condition of structured linear systems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 162–175.
[11] 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.
[12] J. M. Lee, *Introduction to Smooth Manifolds*, Grad. Texts in Math. 218, Springer-Verlag, New York, 2003.
[13] D. S. Mackey, N. Mackey, and D. M. Dunlavy, *Structure preserving algorithms for perplectic eigenproblems*, Electron. J. Linear Algebra, 13 (2005), pp. 10–39.
[14] D. S. Mackey, N. Mackey, and F. Tisseur, *Structured factorizations in scalar product spaces*, SIAM J. Matrix Anal. Appl., 27 (2006), pp. 821–850.
[15] D. S. Mackey, N. Mackey, and F. Tisseur, *Structured Mapping Problems for Matrices Associated with Scalar Products, Part* I: *Lie and Jordan Algebras*, MIMS EPrint 2006.44, Manchester Institute for Mathematical Sciences, The University of Manchester, Manchester, UK, 2006.
[16] S. Noschese and L. Pasquini, *Eigenvalue condition numbers: Zero-structured versus traditional*, J. Comput. Appl. Math., 185 (2006), pp. 174–189.
[17] S. M. Rump, *Eigenvalues, pseudospectrum, and structured perturbations*, Linear Algebra Appl., 413 (2006), pp. 567–593.
[18] G. W. Stewart and J. Sun, *Matrix Perturbation Theory*, Academic Press, London, 1990.
[19] 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.
[20] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 4, pp. 1069–1082

INEXACT INVERSE ITERATION WITH VARIABLE SHIFT FOR NONSYMMETRIC GENERALIZED EIGENVALUE PROBLEMS*

JÖRG BERNS-MÜLLER† AND ALASTAIR SPENCE‡

Abstract. In this paper we analyze inexact inverse iteration for the nonsymmetric generalized eigenvalue problem $\mathbf{Ax} = \lambda\mathbf{Mx}$, where $\mathbf{M}$ is symmetric positive definite and the problem is diagonalizable. Our analysis is designed to apply to the case when $\mathbf{A}$ and $\mathbf{M}$ are large and sparse and preconditioned iterative methods are used to solve shifted linear systems with coefficient matrix $\mathbf{A} - \sigma\mathbf{M}$. We prove a convergence result for the variable shift case (for example, where the shift is the Rayleigh quotient) which extends current results for the case of a fixed shift. Additionally, we consider the approach from [V. Simoncini and L. Eldén, *BIT*, 42 (2002), pp. 159–182] to modify the right-hand side when using preconditioned solves. Several numerical experiments are presented that illustrate the theory and provide a basis for the discussion of practical issues.

Key words. eigenvalue approximation, inverse iteration, iterative methods

AMS subject classifications. 65F10, 65F15

DOI. 10.1137/050623255

1. Introduction.

Consider the generalized eigenvalue problem

$$\mathbf{Ax} = \lambda\mathbf{Mx}, \tag{1.1}$$

where $\mathbf{A}$ is an $n \times n$ nonsymmetric matrix, and $\mathbf{M}$ is an $n \times n$ symmetric positive definite matrix with $\mathbf{x} \in \mathbb{C}^n$, $\lambda \in \mathbb{C}$. In our analysis we restrict ourselves to the case where $\mathbf{M}^{-1}\mathbf{A}$ is diagonalizable; that is, (1.1) has a full set of eigenvectors. Here n is large and $\mathbf{A}$ and $\mathbf{M}$ are assumed to be sparse.

Large-scale eigenvalue problems arise in many applications, such as the determination of linearized stability of a three-dimensional fluid flow. Typically only a few eigenvalues are of interest to the user, and therefore iterative projection methods such as Arnoldi's method [1] and its modern variants [11, 7], or Davidson-type methods [13, 22], and subspace iteration [8, 24, 12] are applied. However, to speed up the convergence (see [2, section 3.3]), often these methods are applied to a "shift-invert" form of (1.1) with the resulting large, sparse linear systems solved iteratively. To obtain a reliable and efficient eigenvalue solver one requires a good understanding of the interaction between the iterative linear solver and the iterative eigenvalue solver. In this paper we study inexact inverse iteration, the simplest inexact iterative method, as a first step in helping to understand more sophisticated inexact eigenvalue techniques.

The classical inverse iteration algorithm to find a single eigenvalue of (1.1) is given as follows.

ALGORITHM 1. inverse iteration.
Given $\mathbf{x}^{(0)}$*, then iterate:*
(1) *Choose* $\sigma^{(i)}$*.*

*Received by the editors January 25, 2005; accepted for publication (in revised form) by I. C. F. Ipsen December 22, 2005; published electronically December 18, 2006. This work was supported by the Engineering and Physical Sciences Research Council, UK, grant GR/M59075.
http://www.siam.org/journals/simax/28-4/62325.html
†Fachbereich Mathematik, JWG-Universität Frankfurt, Postfach 11 19 32, D-60054 Frankfurt, Germany (berns@math.uni-frankfurt.de).
‡Department of Mathematical Sciences, University of Bath, Bath, United Kingdom (as@maths.bath.ac.uk).

(2) *Solve* $(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{M}\mathbf{x}^{(i)}$.
(3) *Set* $\mathbf{x}^{(i+1)} = \mathbf{y}^{(i)}/\varphi(\mathbf{y}^{(i)})$.

Here $\varphi(\mathbf{y}^{(i)})$ denotes a scalar normalizing function. Common choices for φ are $\varphi(\mathbf{y}^{(i)}) = \| \mathbf{y}^{(i)} \|_{\mathbf{M}}$ and $\varphi(\mathbf{y}^{(i)}) = \mathbf{z}^H\mathbf{y}^{(i)}$ for some fixed vector $\mathbf{z}$. Often the choice $\mathbf{z} = \mathbf{e}_k$ is made, where $\mathbf{e}_k$ denotes the kth canonical unit vector and k corresponds to a component of large modulus in the desired eigenvector. One can keep $\sigma^{(i)}$ fixed, so that $\sigma^{(i)} = \sigma^{(0)}$, to obtain a fixed shift method. Alternatively, one can obtain a variable shift method by updating $\sigma^{(i)}$, typically by the Rayleigh quotient or by $\sigma^{(i+1)} = \sigma^{(i)} + 1/(\mathbf{z}^H\mathbf{M}\mathbf{y}^{(i)})$ if $\varphi(\mathbf{y}^{(i)}) = \mathbf{z}^H\mathbf{M}\mathbf{y}^{(i)}$; see [25, p. 637], [6]. An early fundamental paper on Rayleigh quotient iteration for nonsymmetric problems with exact solves is [16].

We consider the following inexact version of inverse iteration.

ALGORITHM 2. inexact inverse iteration.
Given $\mathbf{x}^{(0)}$, *then iterate:*
(1) *Choose* $\sigma^{(i)}$ *and* $\tau^{(i)}$.
(2) *Find* $\mathbf{y}^{(i)}$ *such that* $\|(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} - \mathbf{M}\mathbf{x}^{(i)}\| \leq \tau^{(i)}$.
(3) *Set* $\mathbf{x}^{(i+1)} = \mathbf{y}^{(i)}/\varphi(\mathbf{y}^{(i)})$.

Algorithm 2 is an example of an "inner-outer" iterative algorithm; see, for example, [5]. Here the outer iteration being indexed by i is the standard step in inverse iteration, and the inner iteration refers to the iterative solution of the linear system $(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{M}\mathbf{x}^{(i)}$ to a prescribed accuracy. Since most iterative linear solvers have stopping conditions based on the residual we use the residual condition $\|(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} - \mathbf{M}\mathbf{x}^{(i)}\| \leq \tau^{(i)}$. In practice there are various ways to formulate the inner iteration stopping condition (usually as a relative condition). Here we use an absolute stopping condition to simplify the analysis.

An early paper on inexact inverse iteration for the standard symmetric eigenvalue problem is [19]. More recently [23, 21, 14, 9, 3] various aspects of inexact inverse iteration for the symmetric eigenvalue problem have been considered, usually with the shift chosen as the Rayleigh quotient. It is known (see [10, 6]) that with a fixed and not too accurate shift one needs to solve the shifted linear equations more and more accurately. Additionally, for nonsymmetric generalized eigenvalue problems, the analysis in [6] shows how the accuracy of the inner solves affects the convergence of the outer iteration. Here we extend the convergence theory to the case of variable shifts, for example, when the Rayleigh quotient is used. In this case we show that the tolerance for the inexact solve need not decrease, provided the shift tends towards the desired eigenvalue. The analysis in this paper will be independent of a specific linear solver; we assume only that the residual of the inexact linear solve can be controlled.

The plan of the paper is as follows. Section 2 gives some basic results and notation. Section 3 contains a convergence analysis for inexact inverse iteration. In particular, if Rayleigh quotient shifts are chosen, we see how to regain the quadratic convergence that is achieved using exact linear solves. Alternatively, we show that if the linear systems are solved to a fixed tolerance, we can still achieve a convergent method but with the rate of convergence being only linear. In section 4 we extend the approach of [21] based on modifying the right-hand side of the standard inverse iteration formulation with the aim of reducing the number of inner iterations needed per outer iteration but maintaining the variable shift. This idea is motivated by the work in [20] and has proven to be effective for the symmetric eigenvalue problem. We give a convergence theory and compare it with more standard approaches. In the paper several numerical examples are given to both illustrate the theory and aid the

discussion.

Throughout this paper we use $\|\cdot\|$ for $\|\cdot\|_2$; however, most results are norm independent.

2. Some basic results. We restrict our attention to the case where the generalized eigenvalue problem $\mathbf{Ax} = \lambda\mathbf{Mx}$ is diagonalizable; that is, there exist an invertible matrix $\mathbf{V}$ and a diagonal matrix $\mathbf{\Lambda}$ (both possibly complex) such that

$$\mathbf{AV} = \mathbf{MV\Lambda}, \tag{2.1}$$

and so the eigenvalues of $\mathbf{A}$ lie on the diagonal of $\mathbf{\Lambda}$ and the columns of $\mathbf{V}$ are the right eigenvectors, that is, $\mathbf{Av}_j = \lambda_j\mathbf{Mv}_j$, $j = 1, \ldots, n$. The corresponding decomposition in terms of the left eigenvectors is

$$\mathbf{UA} = \mathbf{\Lambda UM}, \tag{2.2}$$

where $\mathbf{U}$ can be chosen as $\mathbf{U} = \mathbf{V}^{-1}\mathbf{M}^{-1}$ and so $\mathbf{UMV} = \mathbf{I}$. Hence the rows of $\mathbf{U}$ are the left eigenvectors, that is, $\mathbf{u}_j = \mathbf{U}^T\mathbf{e}_j$ with $\mathbf{u}_j^T\mathbf{A} = \lambda_j\mathbf{u}_j^T\mathbf{M}$, $j = 1, \ldots, n$. Note that for the theory we leave the scaling of the eigenvectors open, but we could ask that $\|\mathbf{v}_j\| = 1$ or $\|\mathbf{v}_j\|_{\mathbf{M}} = 1$. In either case $\mathbf{UMV} = \mathbf{I}$ provides the corresponding scaling for $\mathbf{u}_j$.

Using the decomposition (2.1) and assuming that σ is not an eigenvalue of (1.1) we can write

$$\begin{aligned} &(\mathbf{A} - \sigma\mathbf{M})\mathbf{V} = \mathbf{MV}(\mathbf{\Lambda} - \sigma\mathbf{I}) \\ \Leftrightarrow \quad &\mathbf{V}(\mathbf{\Lambda} - \sigma\mathbf{I})^{-1} = (\mathbf{A} - \sigma\mathbf{M})^{-1}\mathbf{MV}. \end{aligned} \tag{2.3}$$

Similarly we can use (2.2) to obtain

$$\begin{aligned} &\mathbf{U}(\mathbf{A} - \sigma\mathbf{M}) = (\mathbf{\Lambda} - \sigma\mathbf{I})\mathbf{UM} \\ \Leftrightarrow \quad &(\mathbf{\Lambda} - \sigma\mathbf{I})^{-1}\mathbf{U} = \mathbf{UM}(\mathbf{A} - \sigma\mathbf{M})^{-1}. \end{aligned} \tag{2.4}$$

2.1. The generalized tangent. In order to analyze the convergence of inexact inverse iteration described in Algorithm 2 we use the following splitting:

$$\mathbf{x}^{(i)} = \alpha^{(i)}(c^{(i)}\mathbf{v}_1 + s^{(i)}\mathbf{w}^{(i)}), \tag{2.5}$$

where $\mathbf{w}^{(i)} \in \operatorname{span}(\mathbf{v}_2, \ldots, \mathbf{v}_n)$ and $\|\mathbf{UMw}^{(i)}\| = 1$. The splitting implies that $\mathbf{V}^{-1}\mathbf{w}^{(i)} \in \operatorname{span}(\mathbf{e}_2, \ldots, \mathbf{e}_n)$ and scaling implies that $\|\mathbf{V}^{-1}\mathbf{w}^{(i)}\| = \|\mathbf{UMw}^{(i)}\| = 1$. Defining

$$\alpha^{(i)} := \|\mathbf{UMx}^{(i)}\|$$

gives $|s^{(i)}|^2 + |c^{(i)}|^2 = 1$, since from (2.5) we have

$$\mathbf{UMx}^{(i)} = \alpha^{(i)}c^{(i)}\mathbf{UMv}_1 + \alpha^{(i)}s^{(i)}\mathbf{UMw}^{(i)}, \tag{2.6}$$

and so

$$\begin{aligned} 1 = \frac{\|\mathbf{UMx}^{(i)}\|}{\alpha^{(i)}} \quad &= \quad \|c^{(i)}\mathbf{e}_1 + s^{(i)}\mathbf{UMw}^{(i)}\| \\ &= \left(|c^{(i)}|^2 + |s^{(i)}|^2\right)^{\frac{1}{2}} \end{aligned}$$

since $\mathbf{e}_1 \perp \mathbf{UMw}^{(i)}$. Thus we interpret $s^{(i)}$ as a generalized sine and $c^{(i)}$ as a generalized cosine, which is in the spirit of the orthogonal decomposition in [17] used for the symmetric eigenvalue problem analysis. For convenience we introduce the matrix $\mathbf{F}$, defined by

$$\mathbf{F} := (\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^T)\mathbf{UM} \;=\; \mathbf{UM}(\mathbf{I} - \mathbf{v}_1\mathbf{u}_1^T\mathbf{M}), \tag{2.7}$$

and note that $\mathbf{Fv}_1 = \mathbf{0}$ and $\mathbf{Fv}_j = \mathbf{e}_j$, so that

$$(\mathbf{UM} - \mathbf{F})\mathbf{x}^{(i)} = \alpha^{(i)}c^{(i)}\mathbf{e}_1, \tag{2.8}$$

and

$$\mathbf{Fx}^{(i)} = \alpha^{(i)}s^{(i)}\mathbf{UMw}^{(i)}. \tag{2.9}$$

Hence $\|(\mathbf{UM} - \mathbf{F})\mathbf{x}^{(i)}\|$ measures the length of the component of $\mathbf{x}^{(i)}$ in the direction of $\mathbf{v}_1$ and $\mathbf{Fx}^{(i)}$ picks out the second term in (2.6). So it is natural to introduce as a measure for convergence of $\mathbf{x}^{(i)}$ to $\mathrm{span}(\mathbf{v}_1)$ the generalized tangent (cf. [6, section 2.1])

$$t^{(i)} \;:=\; \frac{|s^{(i)}|}{|c^{(i)}|} \;=\; \frac{\|\mathbf{Fx}^{(i)}\|}{\|(\mathbf{UM} - \mathbf{F})\mathbf{x}^{(i)}\|}. \tag{2.10}$$

Clearly $\| \frac{1}{c^{(i)}\alpha^{(i)}}\mathbf{x}^{(i)} - \mathbf{v}_1 \| = \; t^{(i)} \; \| \mathbf{w}^{(i)} \|$, and so $t^{(i)}$ measures the quality of the approximation of $\mathbf{x}^{(i)}$ to $\mathbf{v}_1$. Note that $t^{(i)}$ is independent of the factor $\alpha^{(i)}$ and that in the inverse iteration algorithm $\mathbf{x}^{(i)}$ is scaled so that $\varphi(\mathbf{x}^{(i)}) = 1$.

For future reference we recall that for $\mathbf{x} \in \mathbb{C}^n$ the Rayleigh quotient for (1.1) is defined by

$$\varrho(\mathbf{x}) := \frac{\mathbf{x}^H\mathbf{Ax}}{\mathbf{x}^H\mathbf{Mx}} \tag{2.11}$$

and that

$$\varrho(\mathbf{x}^{(i)}) - \lambda_1 = \frac{(\mathbf{x}^{(i)})^H(\mathbf{A} - \lambda_1\mathbf{M})\mathbf{x}^{(i)}}{(\mathbf{x}^{(i)})^H\mathbf{Mx}^{(i)}} \;=\; O(|s^{(i)}|) \tag{2.12}$$

since $(A - \lambda_1\mathbf{M})\mathbf{x}^{(i)} = \alpha^{(i)}s^{(i)}(\mathbf{A} - \lambda_1\mathbf{M})\mathbf{w}^{(i)}$, using (2.5). Thus, the Rayleigh quotient converges linearly in $|s^{(i)}|$ to λ_1. Also, since

$$(\mathbf{A} - \varrho(\mathbf{x}^{(i)})\mathbf{M})\mathbf{x}^{(i)} = (\mathbf{A} - \lambda_1\mathbf{M})\mathbf{x}^{(i)} + (\lambda_1 - \varrho(\mathbf{x}^{(i)}))\mathbf{Mx}^{(i)} \tag{2.13}$$

we have that the eigenvalue residual $\mathbf{r}^{(i)}$ defined by

$$\mathbf{r}^{(i)} := \big(\mathbf{A} - \varrho(\mathbf{x}^{(i)})\mathbf{M}\big)\mathbf{x}^{(i)} \tag{2.14}$$

satisfies

$$\|\mathbf{r}^{(i)}\| = O(|s^{(i)}|). \tag{2.15}$$

Note that while both (2.12) and (2.15) indicate that convergence is linear in $|s^{(i)}|$, it is often the case that convergence to an eigenvalue is faster than convergence to the corresponding eigenvalue residual.

3. Convergence of inexact inverse iteration. In this section we provide the convergence analysis for inexact inverse iteration using a variable shift strategy. In section 3.1 we provide a lemma which gives a bound on the generalized tangent $t^{(i+1)}$. This bound is then used in the convergence theorem in section 3.2. Numerical experiments are presented to illustrate the theory.

Practical choices for $\sigma^{(i)}$ are the update technique

$$\sigma^{(i+1)} = \sigma^{(i)} + 1/\varphi(\mathbf{y}^{(i)}), \tag{3.1}$$

the Rayleigh quotient given by (2.11), or the related

$$\sigma^{(i)} = \frac{\mathbf{z}^H \mathbf{A}\mathbf{x}^{(i)}}{\mathbf{z}^H \mathbf{M}\mathbf{x}^{(i)}}, \tag{3.2}$$

where $\mathbf{z}$ is some fixed vector chosen to maximize $|\mathbf{z}^H\mathbf{M}\mathbf{x}^{(i)}|$. For $\mathbf{M} = \mathbf{I}$ it is common to take $\mathbf{z} = \mathbf{e}_k$, where k corresponds to the component of maximum modulus of $\mathbf{x}^{(i)}$ (for example, see [18]). If the choice $\varphi(\mathbf{y}^{(i)}) = \mathbf{z}^H\mathbf{M}\mathbf{y}^{(i)}$ is made, then for exact solves it is easily shown that

$$\sigma^{(i+1)} = \sigma^{(i)} + \frac{1}{\mathbf{z}^H \mathbf{M}\mathbf{y}^{(i)}} \;=\; \frac{\mathbf{z}^H \mathbf{A}\mathbf{x}^{(i+1)}}{\mathbf{z}^H \mathbf{M}\mathbf{x}^{(i+1)}}, \tag{3.3}$$

so that (3.1) and (3.2) are equivalent. For inexact solves we use (3.2), and it is easily shown that $\lambda_1 - \sigma^{(i)} = O(t^{(i)})$ (cf. (2.12)).

3.1. One step bound. Let us assume that the sought eigenvalue, say λ_1, is simple and well separated. Next, we assume the starting vector $\mathbf{x}^{(0)}$ is neither the solution itself nor is it deficient in the sought eigendirection, that is, $0 <| s^{(i)} |< 1$. Further, we assume that the shift $\sigma^{(i)}$ satisfies

$$|\lambda_1 - \sigma^{(i)}| \le \frac{1}{2}\,|\lambda_2 - \lambda_1| \quad \forall i, \tag{3.4}$$

where $|\lambda_2 - \lambda_1| = \min_{j\neq 1} |\lambda_j - \lambda_1|$. Hence $|\lambda_1 - \sigma^{(i)}| < |\lambda_2 - \sigma^{(i)}|$.

Now consider step (2) of inexact inverse iteration, given by Algorithm 2, and define

$$\mathbf{d}^{(i)} := \mathbf{M}\mathbf{x}^{(i)} - (\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)}. \tag{3.5}$$

Rearranging this equation and using the scaling of $\mathbf{x}^{(i+1)}$ from step (3) in Algorithm 2 together with the fact that $\mathbf{A} - \sigma^{(i)}\mathbf{M}$ is invertible we obtain the update equation

$$\varphi(\mathbf{y}^{(i)})\mathbf{x}^{(i+1)} = (\mathbf{A} - \sigma^{(i)}\mathbf{M})^{-1}(\mathbf{M}\mathbf{x}^{(i)} - \mathbf{d}^{(i)}). \tag{3.6}$$

This is the equation on which the following analysis is based.

LEMMA 3.1. *Assume the shifts satisfy* (3.4) *and that the bound on the residual* $\tau^{(i)}$ *in Algorithm* 2 *satisfies*

$$\|\mathbf{d}^{(i)}\| \;\le\; \tau^{(i)} \;<\; \beta\,|\mathbf{u}_1^T\mathbf{M}\mathbf{x}^{(i)}| \,/\, \|\mathbf{u}_1\| \tag{3.7}$$

for some $\beta \in (0,1)$. *Then*

$$t^{(i+1)} \le \frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \sigma^{(i)}|}\,\frac{|\alpha^{(i)}s^{(i)}| + \|\mathbf{U}\mathbf{d}^{(i)}\|}{(1-\beta)\,|\mathbf{u}_1^T\mathbf{M}\mathbf{x}^{(i)}|}. \tag{3.8}$$

Proof. Recall that $\mathbf{u}_1^T\mathbf{M}\mathbf{x}^{(i+1)} = \alpha^{(i+1)}c^{(i+1)}$, and $\mathbf{u}_1^T = \mathbf{e}_1^T\mathbf{U}$. Hence premultiplying the update equation (3.6) by $\mathbf{u}_1^T\mathbf{M}$ and using $\mathbf{U}\mathbf{M}(\mathbf{A}-\sigma^{(i)}\mathbf{M})^{-1} = (\mathbf{\Lambda}-\sigma^{(i)}\mathbf{I})\mathbf{U}$ (see (2.4)), we obtain

$$\begin{aligned}\varphi(\mathbf{y}^{(i)})\alpha^{(i+1)}c^{(i+1)} &= \mathbf{e}_1^T(\mathbf{\Lambda}-\sigma^{(i)}\mathbf{I})^{-1}\mathbf{U}(\mathbf{M}\mathbf{x}^{(i)}-\mathbf{d}^{(i)})\\ &= (\lambda_1-\sigma^{(i)})^{-1}\mathbf{u}_1^T(\mathbf{M}\mathbf{x}^{(i)}-\mathbf{d}^{(i)}).\end{aligned} \tag{3.9}$$

Further, using (3.7)

$$|\mathbf{u}_1^T\mathbf{M}\mathbf{x}^{(i)}| - |\mathbf{u}_1^T\mathbf{d}^{(i)}| \geq (1-\beta)\,|\mathbf{u}_1^T\mathbf{M}\mathbf{x}^{(i)}|\,. \tag{3.10}$$

Hence

$$\begin{aligned}|\varphi(\mathbf{y}^{(i)})||\alpha^{(i+1)}c^{(i+1)}| &\geq \frac{|\mathbf{u}_1^T\mathbf{M}\mathbf{x}^{(i)}| - |\mathbf{u}_1^T\mathbf{d}^{(i)}|}{|\lambda_1-\sigma^{(i)}|}\\ &\geq (1-\beta)\frac{|\mathbf{u}_1^T\mathbf{M}\mathbf{x}^{(i)}|}{|\lambda_1-\sigma^{(i)}|}.\end{aligned} \tag{3.11}$$

To obtain an upper bound on $|s^{(i+1)}|$ we apply $\mathbf{F}$, defined by (2.7), to (3.6) to obtain

$$\varphi(\mathbf{y}^{(i)})\mathbf{F}\mathbf{x}^{(i+1)} = (\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^T)\mathbf{U}\mathbf{M}(\mathbf{A}-\sigma^{(i)}\mathbf{M})^{-1}(\mathbf{M}\mathbf{x}^{(i)}-\mathbf{d}^{(i)}) \tag{3.12}$$

and using (2.4),

$$\begin{aligned}\varphi(\mathbf{y}^{(i)})\mathbf{F}\mathbf{x}^{(i+1)} &= (\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^T)(\mathbf{\Lambda}-\sigma^{(i)}\mathbf{I})^{-1}\mathbf{U}(\mathbf{M}\mathbf{x}^{(i)}-\mathbf{d}^{(i)})\\ &= (\mathbf{\Lambda}-\sigma^{(i)}\mathbf{I})^{-1}(\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^T)\mathbf{U}(\mathbf{M}\mathbf{x}^{(i)}-\mathbf{d}^{(i)}).\end{aligned} \tag{3.13}$$

Taking norms we obtain

$$\begin{aligned}\|\varphi(\mathbf{y}^{(i)})\mathbf{F}\mathbf{x}^{(i+1)}\| &= \|(\mathbf{\Lambda}-\sigma^{(i)}\mathbf{I})^{-1}(\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^T)\mathbf{U}(\mathbf{M}\mathbf{x}^{(i)}-\mathbf{d}^{(i)})\|\\ &\leq \|(\mathbf{\Lambda}-\sigma^{(i)}\mathbf{I})^{-1}(\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^T)\|\;\|(\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^T)\mathbf{U}(\mathbf{M}\mathbf{x}^{(i)}-\mathbf{d}^{(i)})\|\\ &\leq \frac{1}{|\lambda_2-\sigma^{(i)}|}\Big(\,|\alpha^{(i)}s^{(i)}| + \|(\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^T)\mathbf{U}\mathbf{d}^{(i)}\|\,\Big).\end{aligned} \tag{3.14}$$

With $t^{(i+1)}$ defined by (2.9), and using (2.8), we have

$$\begin{aligned}t^{(i+1)} &= \frac{\|\varphi(\mathbf{y}^{(i)})\mathbf{F}\mathbf{x}^{(i+1)}\|}{\|\varphi(\mathbf{y}^{(i)})(\mathbf{U}\mathbf{M}-\mathbf{F})\mathbf{x}^{(i+1)}\|}\\ &\leq \frac{\|\varphi(\mathbf{y}^{(i)})\mathbf{F}\mathbf{x}^{(i+1)}\|}{|\varphi(\mathbf{y}^{(i)})\alpha^{(i+1)}c^{(i+1)}|}.\end{aligned}$$

Hence, using (3.10), (3.11), and (3.14),

$$t^{(i+1)} \leq \frac{|\lambda_1-\sigma^{(i)}|}{|\lambda_2-\sigma^{(i)}|}\,\frac{|\alpha^{(i)}s^{(i)}| + \|(\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^T)\mathbf{U}\mathbf{d}^{(i)}\|}{|\mathbf{u}_1^T\mathbf{M}\mathbf{x}^{(i)}| - |\mathbf{u}_1^T\mathbf{d}^{(i)}|}. \qquad \square$$

This result is similar to results in [23, 21, 3] in the symmetric case and [6, 15] in the unsymmetric case. One advantage of our approach over that in [6, 15] is that it

can be applied to both fixed and variable shift strategies, though here we concentrate on the variable shift analysis.

Condition (3.7) asks that $\tau^{(i)}$ be bounded in terms of $|\,\mathbf{u}_1^T\mathbf{M}\mathbf{x}^{(i)}\,| = \alpha^{(i)}\,|\,c^{(i)}\,|$ which is related to the cosine of the angle between $\mathbf{v}_1$ and $\mathbf{x}^{(i)}$, the exact and the approximate eigenvectors. In Algorithm 2 we used an absolute tolerance criteria for the inexact solves involving $\tau^{(i)}$. Now Lemma 3.1 shows that this constraint naturally should be relative to the scaling of $\mathbf{x}^{(i)}$.

In the case where $\mathbf{d}^{(i)} = \mathbf{0}$, we can take $\beta = 0$ in (3.7), and (3.8) reduces to $t^{(i+1)} \leq |\frac{\lambda_1-\sigma^{(i)}}{\lambda_2-\sigma^{(i)}}|t^{(i)}$, the familiar expression when exact solves are employed. If (3.4) holds and $\|\mathbf{d}^{(i)}\| \leq \tau^{(i)} \leq C\,|s^{(i)}|$, as is the case if the solve tolerance is bounded by $\|\mathbf{r}^{(i)}\|$ defined in (2.14), then (3.8) indicates that we can expect Algorithm 2 to achieve quadratic convergence, the same asymptotic rate of convergence as the exact solves case. However, if (3.4) holds and $\|\mathbf{d}^{(i)}\| \leq \tau^{(i)} \leq$ *constant*, then we would expect a reduced rate of convergence in Algorithm 2. These expectations about the (outer) convergence rate of Algorithm 2 are made precise in the following section.

3.2. Convergence theorem for variable shifts. The following theorem provides sufficient conditions under which an inexact inverse iteration algorithm with linearly converging shifts achieves linear convergence, even if the residual tolerance is fixed.

THEOREM 3.2. *Given* $\mathbf{A}$, $\mathbf{M} \in \mathbb{R}^{n\times n}$ *with* $\mathbf{M}$ *symmetric positive definite. Let the generalized eigenvalue problem* $\mathbf{A}\mathbf{x} = \lambda\mathbf{M}\mathbf{x}$ *be diagonalizable and have simple eigenpair* $(\lambda_1, \mathbf{v}_1)$. *Further let* $\mathbf{x}^{(i)} = \alpha^{(i)}(c^{(i)}\mathbf{v}_1 + s^{(i)}\mathbf{w}^{(i)})$ *with* $|s^{(0)}| < 1$ *and let the shift updates satisfy*

$$|\lambda_1 - \sigma^{(i)}| \leq \frac{|\lambda_1 - \lambda_2|}{2}\,|s^{(i)}| \quad \forall i. \tag{3.15}$$

Assume that, for $\mathbf{d}^{(i)}$ *defined by* (3.5), $\|\mathbf{d}^{(i)}\| \leq \tau^{(i)}$ *with*

$$\tau^{(i)} < \alpha^{(i)}\beta c^{(i)} / \,\|\mathbf{U}\|, \tag{3.16}$$

where

$$0 \quad \leq \quad \beta \quad < \quad \frac{1 - |s^{(0)}|}{2}. \tag{3.17}$$

Then inexact inverse iteration as given in Algorithm 2 *using a variable shift converges (at least) linearly,* $t^{(i+1)} \leq qt^{(i)} \leq q^{i+1}t^{(0)}$, *where*

$$q := \frac{|s^{(0)}| + \beta}{1 - \beta} \quad < \quad 1. \tag{3.18}$$

Proof. With $|\lambda_1 - \sigma^{(i)}| \leq \frac{1}{2}\,|\lambda_1 - \lambda_2|\,|s^{(i)}|$ and hence $|\lambda_2 - \sigma^{(i)}| > \frac{1}{2}\,|\lambda_2 - \lambda_1|$, we have

$$\frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \sigma^{(i)}|} < |s^{(i)}|\,. \tag{3.19}$$

Thus, from (3.8),

$$
\begin{aligned}
t^{(i+1)} &\leq |s^{(i)}| \ \frac{|\alpha^{(i)} s^{(i)}| + \|\mathbf{U}\| \, \tau^{(i)}}{(1-\beta) \, |\alpha^{(i)} c^{(i)}|} \\
&\leq t^{(i)} \ \frac{|s^{(i)}| + \beta \, |c^{(i)}|}{1-\beta} \\
&\leq t^{(i)} \ \frac{|s^{(0)}| + \beta}{1-\beta}. \qquad (3.20)
\end{aligned}
$$

Set $q = (|s^{(0)}| + \beta)/(1-\beta)$. If β satisfies (3.17), then $q < 1$, and linear convergence is proved by induction. □

This theorem shows that for a close enough starting guess, namely $|s^{(0)}| < 1 - 2\beta$, and for a shift converging linearly, say using (3.2) or (2.11), then we obtain a linearly converging method, provided the inner iteration is solved to a strict enough tolerance (which itself does not tend to zero).

Not surprisingly, if we ask that the bound on the tolerances $\tau^{(i)}$ is linear in $|\, s^{(i)} \,|$ instead of being held fixed as allowed by (3.16), then one achieves quadratic convergence. This is stated in the following corollary.

COROLLARY 3.3. *Assume the conditions of Theorem* 3.2 *are satisfied but that* (3.16) *is replaced by*

$$
\tau^{(i)} \leq \alpha^{(i)} \min(\beta c^{(0)} / \|\mathbf{U}\|, \gamma \, |s^{(i)}|) \qquad (3.21)
$$

for some constant $\gamma \geq 0$*; then the convergence is (at least quadratic), that is,* $t^{(i+1)} \to 0$ *(monotonically) with* $t^{(i+1)} \leq q (t^{(i)})^2$ *for some* $q > 0$.

Conditions (3.16), (3.17), and (3.18) make precise statements such as "$\tau^{(i)}$ is small enough" and "$\mathbf{x}^{(0)}$ is close enough to $\mathbf{v}_1$." Those are unlikely to be of any quantitative use since they are probably too restrictive and contain quantities that are unknown (for example $\|\mathbf{U}\|$ and $|\lambda_2 - \lambda_1|$). Of course, the conditions (3.16), (3.18), and (3.21) are not necessary, and in our experiments considerably larger values for $\tau^{(i)}$ have been used successfully. Condition (3.15) is easily satisfied if $\sigma^{(i)}$ is given by (3.2) and if $\mathbf{z}$ is sufficiently close to the left eigenvector $\mathbf{u}_1$. However, this is a theoretically sufficient condition, and as is the case in many practical situations convergence occurs without this condition being fulfilled.

We now present some numerical results to illustrate the theory given in Theorem 3.2 and Corollary 3.3. In our experiments different choices of shift produced no significant changes in the results, so we present numerical results for the Rayleigh quotient shift only.

Example 1. Consider $\mathbf{A}$ and $\mathbf{M}$ derived by discretizing

$$
\begin{aligned}
-\Delta u + 5u_x + 5u_y &= \lambda u \quad \text{in } D := [0,1] \times [0,1], \\
u &= 0 \quad \text{on } \Gamma := \partial D,
\end{aligned}
$$

using the Galerkin FEM on regular triangular elements with piecewise linear functions. This eigenvalue problem is also discussed in [6]. Here we use a 32 by 32 grid which leads to 961 degrees of freedom. For the discrete eigenvalue problem it is known that $\lambda_1 \approx 32.2$ and $\lambda_2 \approx 61.7$ with all other eigenvalues satisfying $Re(\lambda_j) > 61.8$. Note that the eigenvalue residual $\mathbf{r}^{(i)}$ defined by (2.14) is proportional to $|\, s^{(i)} \,|$ (using (2.15)), and so this provides a practical way to implement a decreasing tolerance. As inexact linear solver we use preconditioned full GMRES (that is, without restarts), where the preconditioner $\mathbf{P} \approx \mathbf{A}$ is obtained by an incomplete modified LU decomposition

TABLE 3.1
Generalized tangent $t^{(i)}$ and number of inner iterations $k^{(i)}$ for **RQIf** (a) *and* (b) *and* **RQId** (c). *In* (a) $\tau_0 = 0.1$, *in* (b) $\tau_0 = 0.001$, *and in* (c) $\tau_0 = 0.2$ *and* $\tau_1 = 0.5$.

	(a)		(b)		(c)	
	$t^{(i)}$	$k^{(i-1)}$	$t^{(i)}$	$k^{(i-1)}$	$t^{(i)}$	$k^{(i-1)}$
0	5.0e-02		5.0e-02		5.0e-02	
1	9.0e-03	11	4.4e-04	23	1.6e-02	13
2	2.4e-04	19	8.0e-08	36	2.8e-05	35
3	4.6e-06	29	7.7e-12	51	2.9e-10	54
4	2.6e-08	37			6.8e-12	51
5	4.7e-11	47				
6	1.0e-11	52				
$\sum k^{(i-1)}$		195		110		153

with drop tolerance = 0.1. In Table 3.1 we present numerical results obtained when calculating λ_1. Each row in Table 3.1 provides the generalized tangent, $t^{(i)}$ (calculated knowing the exact solution $\mathbf{v}_1$), and $k^{(i-1)}$ the number of inner iterations used by preconditioned GMRES to satisfy the residual condition. We use the following two versions of Algorithm 2.

RQIf, Rayleigh quotient iteration with fixed tolerance, that is, $\sigma^{(i)} = \varrho(\mathbf{x}^{(i)})$ and $\tau^{(i)} = \tau_0 \|\mathbf{M}\mathbf{x}^{(i)}\|$.

RQId, Rayleigh quotient iteration with decreasing tolerance, that is, $\sigma^{(i)} = \varrho(\mathbf{x}^{(i)})$ and $\tau^{(i)} = \min\{\tau_0, \tau_1 \|\mathbf{r}^{(i)}\|/\sigma^{(i)}\} \|\mathbf{M}\mathbf{x}^{(i)}\|$.

As $\|\mathbf{r}^{(i)}\| \,/\, |\varrho^{(i)}|$ is proportional to $|s^{(i)}|$ and $\|\mathbf{M}\mathbf{x}^{(i)}\|$ is proportional to $\alpha^{(i)}$ we expect according to Theorem 3.2 linear convergence for **RQIf** and according to Corollary 3.3 quadratic convergence for **RQId**.

In Table 3.1, cases (a) and (b) illustrate the behavior of **RQIf** with $\tau_0 = 0.1$ and 0.001, respectively. Case (c) gives results for **RQId**, that is, Rayleigh quotient shifts and a decreasing tolerance based on the eigenvalue residual (2.14). We present results for the approximation of $(\lambda_1, \mathbf{v}_1)$ and stop the entire calculation once the relative eigenvalue residual $\|\mathbf{r}^{(i)}\| \,/\varrho^{(i)}$ is smaller than $\tau_{outer} = 10^{-14}$.

Discussion of results. Case (a) shows that the Rayleigh quotient iteration with fixed tolerance $\tau_0 = 0.1$ achieves linear convergence (indeed, in this experiment, superlinear convergence). Case (c) shows that the Rayleigh quotient iteration with linearly decreasing tolerance based on the eigenvalue residual achieves quadratic convergence as predicted by Corollary 3.3. Thus we recover the convergence rate attained for nonsymmetric problems if the Rayleigh quotient iteration is used with exact solves. We point out that the last iteration in (c) is stopped due to the fact that the relative outer tolerance condition is satisfied within GMRES, and so quadratic convergence is lost in the final step. Case (b) shows results obtained using the Rayleigh quotient iteration with a small fixed tolerance. First, we note that since τ_0 is small the method behaves very similarly to the exact solves case. Further, case (b) exhibits initially quadratic convergence as the $s^{(i)}$ dominates $\tau^{(i)}$ in the numerator of (3.8). However, this quadratic convergence is lost when the tangent, $t^{(i)}$, has reduced to the order of the stopping tolerance, and then $\tau^{(i)}$ dominates $s^{(i)}$.

4. Modified right-hand side. In this section we analyze a variation of inexact inverse iteration where the right-hand side is altered with the aim of improving the performance of the preconditioned iterative solver at the risk of slowing down the

outer convergence rate. This idea has been used in [20] and [21]. Instead of solving

$$(\mathbf{A} - \sigma\mathbf{M})\mathbf{y}^{(i)} = \mathbf{M}\mathbf{x}^{(i)} \tag{4.1}$$

[20] used the system

$$(\mathbf{A} - \sigma\mathbf{M})\mathbf{y}^{(i)} = \mathbf{x}^{(i)} \tag{4.2}$$

with no theoretical justification but with the remark that computational time is saved with the modified right-hand side. Also, for the solution of the standard symmetric eigenvalue problem $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ using a preconditioner $\mathbf{P} \approx (\mathbf{A} - \sigma\mathbf{I})$, Simoncini and Eldén [21] solve

$$\mathbf{P}^{-1}(\mathbf{A} - \sigma\mathbf{I})\mathbf{y}^{(i)} = \mathbf{x}^{(i)} \tag{4.3}$$

rather than the obvious system

$$\mathbf{P}^{-1}(\mathbf{A} - \sigma\mathbf{I})\mathbf{y}^{(i)} = \mathbf{P}^{-1}\mathbf{x}^{(i)}. \tag{4.4}$$

The motivation for this alteration is that in (4.3) the right-hand side $\mathbf{x}^{(i)}$ is both close to a null vector of $\mathbf{P}^{-1}(\mathbf{A} - \sigma\mathbf{I})$ and close to a scaled version of the solution. The vector $\mathbf{P}^{-1}\mathbf{x}^{(i)}$ has neither of these properties. Here we combine the two ideas. Let $\mathbf{P} \approx (\mathbf{A} - \sigma\mathbf{M})$ be a preconditioner for use within GMRES. Given an approximate eigenvector $\mathbf{x}^{(i)}$ to obtain an improved eigendirection using preconditioned GMRES we solve

$$\mathbf{P}^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{x}^{(i)} \tag{4.5}$$

rather than the obvious $\mathbf{P}^{-1}(\mathbf{A}-\sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{P}^{-1}\mathbf{M}\mathbf{x}^{(i)}$. As we shall show below, by changing the right-hand side from $\mathbf{P}^{-1}\mathbf{M}\mathbf{x}^{(i)}$ to $\mathbf{x}^{(i)}$ the convergence theory changes. The expected gain is that (4.5) will prove to be significantly cheaper to solve in terms of inner iterations. For the standard symmetric eigenvalue problem where the shift was chosen as the Rayleigh quotient this was indeed the case. We shall see that for nonsymmetric problems the situation is not so clear-cut. In this paper we shall concentrate on the outer convergence theory. The algorithm derived from solving (4.5) which uses the Rayleigh quotient shift is defined as follows.

ALGORITHM 3. inexact inverse iteration with modified right-hand side.
Given $\mathbf{x}^{(0)}$*, then iterate:*
(1) *Choose* $\tau^{(i)}$*, and set* $\sigma^{(i)} = \varrho(\mathbf{x}^{(i)})$.
(2) *Find* $\mathbf{y}^{(i)}$ *such that* $\|\mathbf{x}^{(i)} - \mathbf{P}^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)}\| \le \tau^{(i)}$.
(3) *Set* $\mathbf{x}^{(i+1)} = \mathbf{y}^{(i)}/\varphi(\mathbf{y}^{(i)})$.

Note that we use a standard residual condition rather than the stopping condition used in [21, section 7]. We define the residual obtained by solving (4.5) approximately as

$$\mathbf{d}^{(i)} := \mathbf{x}^{(i)} - \mathbf{P}^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} \tag{4.6}$$

so that the inexact solve step can be written as

$$(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{P}\mathbf{x}^{(i)} - \mathbf{P}\mathbf{d}^{(i)}, \tag{4.7}$$

which should be compared with the inexact solve step

$$(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{M}\mathbf{x}^{(i)} - \mathbf{d}^{(i)} \tag{4.8}$$

in section 3. From (4.8) we obtain

$$\varphi(\mathbf{y}^{(i)})\mathbf{x}^{(i+1)} = (\mathbf{A} - \sigma^{(i)}\mathbf{M})^{-1}\mathbf{P}(\mathbf{x}^{(i)} - \mathbf{d}^{(i)}) \tag{4.9}$$

(cf. (3.6)), which is used in the following analysis. First, assume the residual $\mathbf{d}^{(i)}$ satisfies the bound

$$\|\mathbf{d}^{(i)}\| \quad \leq \quad \tau^{(i)} \quad \leq \quad \beta' \, |\mathbf{u}_1^T\mathbf{P}\mathbf{x}^{(i)}| \, / \, \|\mathbf{U}\mathbf{P}\| \tag{4.10}$$

for some $\beta' \in ([0, 1)$ (cf. (3.7)), and hence it is easily shown that

$$|\mathbf{u}_1^T\mathbf{P}\mathbf{x}^{(i)}| - |\mathbf{u}_1^T\mathbf{P}\mathbf{d}^{(i)}| \geq (1 - \beta') \, |\mathbf{u}_1^T\mathbf{P}\mathbf{x}^{(i)}| \, . \tag{4.11}$$

Next, we introduce the expression

$$T_P(\mathbf{z}) := \frac{\|(\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^T)\mathbf{U}\mathbf{P}\mathbf{z}\|}{|\mathbf{u}_1^T\mathbf{P}\mathbf{z}|}, \tag{4.12}$$

where $\mathbf{z} \in \mathbb{C}^n$. By analogy with (2.7) and (2.10), $T_P(\mathbf{z})$ looks like a generalized tangent with respect to $\mathbf{P}$ rather than $\mathbf{M}$. However, for a general preconditioner $T_P(\mathbf{v}_1) \neq 0$. In fact, $T_P(\mathbf{v}_1)$ measures the effect of $\mathbf{P}$ on the eigenvector $\mathbf{v}_1$, and we shall see in Theorem 4.2 that large values of $T_P(\mathbf{v}_1)$ will slow down or possibly destroy the convergence of Algorithm 3. Note that, under (4.11),

$$T_P(\mathbf{x}^{(i)} - \mathbf{d}^{(i)}) \leq \frac{1}{1 - \beta'} \left(T_P(\mathbf{x}^{(i)}) + \frac{\|\mathbf{U}\mathbf{P}\mathbf{d}^{(i)}\|}{|\mathbf{u}_1^T\mathbf{P}\mathbf{x}^{(i)}|} \right). \tag{4.13}$$

Now we give a one step bound for Algorithm 3 using a variable shift $\sigma^{(i)}$.

LEMMA 4.1. *Assume $\sigma^{(i)}$ satisfies* (3.4) *and* (3.15). *Further assume that* (4.11) *holds. Then*

$$\begin{aligned} t^{(i+1)} &\leq \frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \sigma^{(i)}|} \, T_P(\mathbf{x}^{(i)} - \mathbf{d}^{(i)}) \\ &\leq |s^{(i)}| \; T_P(\mathbf{x}^{(i)} - \mathbf{d}^{(i)}), \end{aligned} \tag{4.14}$$

where $T_P(\cdot)$ is given by (4.12).

Proof. With the notation in sections 2 and 3 we have

$$\begin{aligned} t^{(i+1)} &= \frac{\|\mathbf{F}\varphi(\mathbf{y}^{(i)})\mathbf{x}^{(i+1)}\|}{\|(\mathbf{U}\mathbf{M} - \mathbf{F})\varphi(\mathbf{y}^{(i)})\mathbf{x}^{(i+1)}\|} \\ &= \frac{\|\mathbf{F}(\mathbf{A} - \sigma^{(i)}\mathbf{M})^{-1}(\mathbf{P}\mathbf{x}^{(i)} - \mathbf{P}\mathbf{d}^{(i)})\|}{\|(\mathbf{U}\mathbf{M} - \mathbf{F})(\mathbf{A} - \sigma^{(i)}\mathbf{M})^{-1}(\mathbf{P}\mathbf{x}^{(i)} - \mathbf{P}\mathbf{d}^{(i)})\|} \\ &= \frac{\|(\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^T)\mathbf{U}\mathbf{M}(\mathbf{A} - \sigma^{(i)}\mathbf{M})^{-1}(\mathbf{P}\mathbf{x}^{(i)} - \mathbf{P}\mathbf{d}^{(i)})\|}{|\mathbf{e}_1^T\mathbf{U}\mathbf{M}(\mathbf{A} - \sigma^{(i)}\mathbf{M})^{-1}(\mathbf{P}\mathbf{x}^{(i)} - \mathbf{P}\mathbf{d}^{(i)})|} \\ &= \frac{\|(\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^T)(\mathbf{\Lambda} - \sigma^{(i)}\mathbf{I})^{-1}\mathbf{U}(\mathbf{P}\mathbf{x}^{(i)} - \mathbf{P}\mathbf{d}^{(i)})\|}{|\mathbf{e}_1^T(\mathbf{\Lambda} - \sigma^{(i)}\mathbf{I})^{-1}\mathbf{U}(\mathbf{P}\mathbf{x}^{(i)} - \mathbf{P}\mathbf{d}^{(i)})|} \\ &\leq \frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \sigma^{(i)}|} \, \frac{\|(\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^T)\mathbf{U}\mathbf{P}(\mathbf{x}^{(i)} - \mathbf{d}^{(i)})\|}{|\mathbf{e}_1^T\mathbf{U}\mathbf{P}(\mathbf{x}^{(i)} - \mathbf{d}^{(i)})|}, \end{aligned}$$

TABLE 4.1
Generalized tangent $t^{(i)}$ and number of inner iterations $k^{(i)}$ for **RQIf** (a) *and* **RQImodrhs** (b) *with $\tau_0 = 0.05$ for both methods.*

	(a)		(b)	
	$t^{(i)}$	$k^{(i-1)}$	$t^{(i)}$	$k^{(i-1)}$
0	2.0e-02		2.0e-02	
1	1.6e-02	30	1.6e-02	30
2	2.9e-05	41	1.2e-04	37
3	4.7e-08	47	9.8e-07	37
4	2.0e-08	47	4.4e-08	36
5			1.7e-08	24
$\sum k^{(i-1)}$		165		164

from which the required result follows. ∎

Clearly a formal statement of the convergence of Algorithm 3 merely requires conditions that ensure the second term on the right-hand side of (4.14) remains bounded below 1 for all i. For completeness we present such a theorem.

THEOREM 4.2. *Assume that the conditions of Lemma* 4.1 *hold, and let $\tau^{(i)}$ satisfy* (4.10) *with $\beta' \in [0, 1)$. Assume that $T_P(\mathbf{v}_1) \neq 0$ and*

$$q := \frac{1}{1-\beta'}\left(2T_P(\mathbf{v}_1) + \beta'\right) < 1. \tag{4.15}$$

Then, for $\mathbf{x}^{(0)}$ close enough to $\mathbf{v}_1$, Algorithm 3 *converges linearly with $t^{(i+1)} \leq q t^{(i)}$.*

Proof. Due to the condition on $\tau^{(i)}$, (4.10), we can use (4.13) and (4.10) (again) to give $T_P(\mathbf{x}^{(i)} + \mathbf{d}^{(i)}) \leq (1-\beta')^{-1}(T_P(\mathbf{x}^{(i)}) + \beta')$. Hence it remains to show that $T_P(\mathbf{x}^{(i)}) \leq 2T_P(\mathbf{v}_1)$, which is valid for $\mathbf{x}^{(0)}$ close enough to $\mathbf{v}_1$ as $T_P(\mathbf{v}_1) \neq 0$. ∎

Lemma 4.1 and Theorem 4.2 show that the quantity $T_P(\mathbf{v}_1)$ plays an important role in the convergence of Algorithm 3, and ideally $T_P(\mathbf{v}_1)$ should be small. In practical situations we will have little knowledge of the effect of $\mathbf{P}$ on $\mathbf{v}_1$, but it is clear that if $\mathbf{u}_1^T \mathbf{P} \mathbf{v}_1$ is small, and hence $T_P(\mathbf{v}_1)$ is large; then Algorithm 3 may converge slowly or may possibly fail to converge. Note that we ignore the unlikely case $T_P(\mathbf{v}_1) = 0$ in Theorem 4.2, though in this case one could recover quadratic convergence using a decreasing tolerance. We present numerical values for $T_P(\mathbf{v}_1)$ in Table 4.2. First, we compare the performance of Algorithm 3 with the variable shift method **RQIf** discussed in Example 1.

Example 2. Again we consider the convection diffusion problem of Example 1; however, now we seek the interior eigenvalue $\lambda_{20} = 337.7$. Here we use preconditioned full GMRES with multigrid as preconditioner to solve the linear systems that arise. The preconditioner consists of one V-cycle and uses 3 Jacobi iterations for both pre- and postsmoothing on each grid. In case (a) of Table 4.1 we use **RQIf** with $\tau_0 = 0.05$ and in (b) we use **RQImodrhs** with $\tau_0 = 0.05$.

RQImodrhs, Algorithm 3 with $\sigma^{(i)} = \varrho(\mathbf{x}^{(i)})$ and tolerance $\tau^{(i)} = \tau_0 \, \|\mathbf{P}\mathbf{x}^{(i)}\|$.

We present numerical results for calculating λ_{20} up to a relative outer tolerance of $\tau_{outer} = 10^{-10}$ in Table 4.1.

Discussion of results. From case (a) we observe that the number of inner iterations $k^{(i)}$ increases as the outer process proceeds. This effect was already observed when calculating the eigenvalue λ_1 of the same example; see Table 3.1. However, the rate of increase here is not as substantial due to the fact that the multigrid preconditioner is a much better preconditioner than the one constructed by the incomplete LU decomposition. Case (b) shows that even though the right-hand side

TABLE 4.2
Generalized tangent $t^{(i)}$ for **RQImodrhs** *with $\tau_0 = 0.01$ using two different preconditioners. In* (a) *milu*$(\mathbf{A}, 0.1)$*, where $T_P(\mathbf{v}_1) = 0.34$, and in* (b) *milu*$(\mathbf{A} - 320\mathbf{M}, 10^{-4})$*, where $T_P(\mathbf{v}_1) = 0.045$.*

	(a)	(b)
	$t^{(i)}$	$t^{(i)}$
0	2.0e-02	2.0e-02
1	3.1e-04	1.9e-04
2	4.7e-05	5.8e-07
3	2.6e-06	1.5e-09
4	1.3e-07	
5	1.1e-08	

has been modified **RQImodrhs** still provides a linearly converging method as stated in Theorem 4.2. Further, the number of inner iterations used at each outer iteration by **RQImodrhs** does not increase with i, which leads to an efficient iteration process. (The link between the outer convergence and the cost of the inner solves using GMRES is discussed further in [4].) We also observe, however, that **RQImodrhs** requires more outer iterations. This is to be expected from the convergence theory because of the nonzero term $T_P(\mathbf{v}_1)$ in (4.15) and is observed in other experiments; see Table 8.6 in [4]. Note that the choices for τ_0 in Example 2 are not optimal for either method. For **RQImodrhs** the optimal value (that is, the value producing the smallest total number of inner iterations) is $\tau_0 = 0.1$, and for **RQIf** the optimal value is $\tau_0 = 0.001$. However, there was little difference in the performance of the methods. In both cases the total number of inner iterations was around 130.

We remark that in our experience with several different examples for the generalized nonsymmetric eigenvalue problem the choice of the constant τ_0 as used in the bound on the tolerance is important for both the convergence and efficiency of Algorithm 3. This is in contrast to the standard symmetric eigenvalue problem where the corresponding algorithms are less sensitive to the choice of τ_0, as reported in [3].

Next, we provide an example to demonstrate the effect of $T_P(\mathbf{v}_1)$ on the rate of convergence.

Example 3. Again we consider the convection diffusion problem discussed in Example 2, and we seek the interior eigenvalue $\lambda_{20} = 337.7$. To demonstrate the effect of $T_P(\mathbf{v}_1)$ on the convergence of **RQImodrhs** we consider two different preconditioner. In case (a) of Table 4.2 we use a modified incomplete LU decomposition constructed from the unshifted system $\mathbf{A}$ using a drop tolerance of 0.1; we denote this by milu$(\mathbf{A}, 0.1)$. The other preconditioner, which we use in case (b), is also a modified incomplete LU decomposition constructed now from the shifted system $\mathbf{A} - 320\mathbf{M}$ using a drop tolerance of 10^{-4} (milu$(\mathbf{A} - 320\mathbf{M}, 10^{-4})$). In Table 4.2 we present numerical results obtained using **RQImodrhs** with $\tau_0 = 0.01$ using in (a) the "unshifted" preconditioner which has for this example $T_P(\mathbf{v}_1) = 0.34$ and in (b) the "shifted" preconditioner which has $T_P(\mathbf{v}_1) = 0.045$.

Note that in our experience parameter values for τ_0 smaller than 0.01 did not alter the outer convergence. This is not surprising since $\tau_0 \ll T_P(\mathbf{v}_1)$, and hence according to Theorem 4.2 the effect of the inexact solves on the rate of convergence should not be significant.

Discussion of results. From Table 4.2 we observe that the outer convergence in case (a) is linear with a rate $t^{(i+1)}/t^{(i)} \approx 0.05$. Comparing this with the results for case (b) we observe a significant improvement in the outer rate of convergence, which results in a reduced number of outer iterations. In Algorithms 1 and 2 the

preconditioner merely makes the solution of the linear system more efficient, whereas in Algorithm 3 the preconditioner also affects the outer convergence rate, as seen by the presence of $T_P(\mathbf{v}_1)$ term on the right-hand side in (4.15).

5. Conclusion. In this paper we provided a convergence theory for inexact inverse iteration with varying shifts applied to the nonsymmetric generalized eigenvalue problem. Additionally we extended the approach from [21] of modifying the right-hand side to the nonsymmetric generalized eigenvalue problem, presented a convergence theory, and showed that the preconditioner affects the outer convergence rate.

REFERENCES

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

[2] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, SIAM, Philadelphia, 2000.

[3] J. Berns-Müller, I. G. Graham, and A. Spence, *Inverse iteration and inexact solves*, Linear Algebra Appl., to appear.

[4] J. Berns-Müller and A. Spence, *Inexact Inverse Iteration and GMRES*, Tech. report maths0507, University of Bath, Bath, UK, 2005.

[5] G. H. Golub and Q. Ye, *Inexact preconditioned conjugate gradient method with inner-outer iteration*, SIAM J. Sci. Comput., 21 (1999), pp. 1305–1320.

[6] G. H. Golub and Q. Ye, *Inexact inverse iteration for generalized eigenvalue problems*, BIT, 40 (2000), pp. 671–684.

[7] Z. Jia, *A refined iterative algorithm based on the block Arnoldi process for large unsymmetric eigenproblems*, Linear Algebra Appl., 270 (1998), pp. 171–189.

[8] H.-J. Jung, M.-C. Kim, and I.-W. Lee, *An improved subspace iteration method with shifting*, Comput. & Structures, 70 (1999), pp. 625–633.

[9] A. V. Knyazev, *Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method*, SIAM J. Sci. Comput., 23 (2001), pp. 517–541.

[10] 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.

[11] R. B. Lehoucq, *Analysis and Implementation of an Implicitly Restarted Arnoldi Iteration*, Ph.D. thesis, Rice University, Houston, TX, 1995.

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

[13] R. B. Morgan and D. S. Scott, *Generalizations of Davidson's method for computing eigenvalues of sparse symmetric matrices*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 817–825.

[14] K. Neymeyr, *A geometric theory for preconditioned inverse iteration* I: *Extrema of the Rayleigh quotient*, Linear Algebra Appl., 322 (2001), pp. 61–85.

[15] K. Neymeyr, *A geometric theory for preconditioned inverse iteration* II: *Convergence estimates*, Linear Algebra Appl., 322 (2001), pp. 87–104.

[16] B. N. Parlett, *The Rayleigh quotient iteration and some generalizations for nonnormal matrices*, Math. Comp., 28 (1974), pp. 679–693.

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

[18] G. Peters and J. H. Wilkinson, *Inverse iteration, ill-conditioned equations and Newton's method*, SIAM Rev., 21 (1979), pp. 339–360.

[19] A. Ruhe and T. Wiberg, *The method of conjugate gradients used in inverse iteration*, BIT, 12 (1972), pp. 543–554.

[20] D. S. Scott, *Solving sparse symmetric generalized eigenvalue problems without factorization*, SIAM J. Numer. Anal., 18 (1981), pp. 102–110.

[21] V. Simoncini and L. Eldén, *Inexact Rayleigh quotient-type methods for eigenvalue computations*, BIT, 42 (2002), pp. 159–182.

[22] 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.

[23] 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.

[24] X. Wang and J. Zhou, *An accelerated subspace iteration method for generalized eigenproblems*, Comput. & Structures, 71 (1999), pp. 293–301.

[25] J. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, UK, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 4, pp. 1083–1096

SPECTRAL ANALYSIS OF SYMPLECTIC MATRICES WITH APPLICATION TO THE THEORY OF PARAMETRIC RESONANCE*

S. K. GODUNOV† AND M. SADKANE‡

Abstract. The numerical analysis of the spectral structure of symplectic matrices is proposed. The strong stability of symplectic matrices and linear Hamiltonian systems with periodic coefficients is discussed. Applications to the theory of parametric resonance are illustrated by spectral portraits calculation.

Key words. symplectic matrix, linear Hamiltonian system, strong stability, spectral dichotomy, spectral portrait, parametric resonance

AMS subject classifications. 65F15, 15A57, 37J25, 70J40

DOI. 10.1137/040618503

1. Introduction.

In [4], an algorithm for computing invariant subspaces of matrices and matrix pencils using spectral dichotomy techniques was proposed. The aim of the present paper is to show that this algorithm can be used to analyze the stability of symplectic matrices and linear Hamiltonian systems with periodic coefficients. It can also be used to illustrate the theory of parametric resonance.

The spectral dichotomy methods originate from Lyapunov's theory. It allows one to determine whether a regular matrix pencil $\lambda B - A$ has eigenvalues on or in a neighborhood of a curve γ in the complex plane. In the regions where no such eigenvalues exist, the projector $P(\rho)$ onto the deflating subspace of $\lambda B - A$ is computed along with the matrix integral

$$F(\rho) = \oint_{\gamma=\gamma(\rho)} \left((\lambda B - A)^{-1}\right)^* (\lambda B - A)^{-1} \, |d\lambda| = F(\rho)^*. \tag{1.1}$$

Up to a multiplicative constant, the matrix $F \equiv F(\rho)$ and the projector $P \equiv P(\rho)$ are related by Lyapunov-type equations in the case where γ is a circle or the imaginary axis. See [2, Chap. 10]. The 2-norm of F, denoted hereafter by $\|F\|$, is used to measure the numerical quality of the projector P. The larger this norm, the less accurate the computed projector P, or in other words, a large $\|F\|$ means that $\lambda B - A$ has eigenvalues on or in a neighborhood of $\gamma(\rho)$. The matrix F can be computed iteratively as shown in [4, Thm. 2.2]. We mention that the algorithm proposed in [6] computes P and $\|F\|$ efficiently.

The spectral portrait is defined as the graph of the function

$$\rho \longmapsto f(\rho) = \|F(\rho)\|. \tag{1.2}$$

When the pencil $\lambda B - A$ has an eigenvalue $\lambda \in \gamma(\rho_0)$, then the function $f(\rho_0)$ goes to ∞. In other words, the graph of f has an asymptote on the line $\rho = \rho_0$. It can be shown [2, sect. 13.6] that the function $\log f$ is convex on each interval where f is

*Received by the editors November 8, 2004; accepted for publication (in revised form) by K. Veselic December 3, 2005; published electronically December 18, 2006.
http://www.siam.org/journals/simax/28-4/61850.html

†Sobolev Institute of Mathematics, 630090 Novosibirsk, Russia (godunov@math.nsc.ru).
‡Université de Brest. Département de Mathématiques, 6 Av. Le Gorgeu, CS 93837, 29238 Brest Cedex 3, France (sadkane@univ-brest.fr).

finite. Numerically, the graph of f is composed of several strips separated by intervals where f takes a prescribed large number. The larger this interval, the more ill-conditioned the corresponding eigenvalues and deflating subspaces. The other values of ρ correspond to regions of absence of eigenvalues where the trace of the projector P remains constant.

Throughout this paper, the projector $P(\rho)$ and the norm $\|F(\rho)\|$ are computed with the algorithms proposed in [4]. See in particular Algorithm 1 in [4]. The following notation will be used: If $H = H^*$ is a Hermitian (symmetric) matrix, then the notation $H > 0$ ($H \geq 0$) means that H is positive definite (semidefinite). Unless otherwise stated, the matrix $J_{2k} = \left(\begin{smallmatrix} 0_k & -I_k \\ I_k & 0_k \end{smallmatrix}\right)$ will be denoted by J. The identity (null) matrix of order k will be denoted by I_k (0_k) or just I (0) when the order is clear from the context.

2. Stability of symplectic matrices. Let J be a real skew-symmetric nonsingular $2N \times 2N$ matrix.[1] A real $2N \times 2N$ matrix W is said to be J-symplectic if $W^*JW = J$. The spectrum of W is generally composed of three groups: (i) N_∞ eigenvalues outside the unit circle, (ii) $N_0 = N_\infty$ eigenvalues inside the unit circle and placed symmetrically with respect to the previous group, and (iii) $N_1 = 2N - 2N_0$ eigenvalues on the unit circle.

The symplectic matrices arise in several applications, among which are optimal control (see, e.g., [5, Chap. 12]) and the theory of parametric resonance (see, e.g., [7]), which we discuss in section 3. In the first application, the unit circle is free of eigenvalues, and it is important to construct, in a stable way, the projectors P_0 and P_∞ onto the invariant subspaces of W associated to the eigenvalues respectively inside and outside the unit circle. In the second application, all the eigenvalues should be on the unit circle, i.e., $P_0 = P_\infty = 0$ and $P_1 = I$, where P_1 is the projector onto the invariant subspace of W associated to the eigenvalues on the unit circle. These necessary conditions are not sufficient to ensure the stability of W. Before going through the details, we recall some definitions.

DEFINITION 2.1. *An eigenvalue λ of W on the unit circle is said to be of the first (second) kind if any corresponding eigenvector x satisfies $(iJx, x) > 0$ ($(iJx, x) < 0$), where (iJx, x) stands for the Euclidean inner product. When $(Jx, x) = 0$, then λ is said to be of the mixed kind.*

DEFINITION 2.2. *The J-symplectic matrix W is stable if $\|W^k\| < \infty$ for all $k > 0$. It is strongly stable if $W + \Delta$ remains stable under small perturbations Δ which conserve the symplecticity of $W + \Delta$.*

It is clear that the stability implies that all eigenvalues of W lie on the unit circle and are not defective. It was shown by Krein, Gelfand, and Lidskii (see [7, pp. 161, 192]) that the strong stability is equivalent to the following conditions, called hereafter the KGL criterion:

- all the eigenvalues of W are on the unit circle;
- the eigenvalues of W are either of the first or second kind.

To these two conditions, one should actually add a third: the eigenvalues of the first and second kinds must be well separated and separated from ± 1, which are eigenvalues of mixed kind (see below).

The symplectic matrix W often results from some Hamiltonian systems as explained below. Its spectrum is a priori unknown and its numerical computation may

[1] Although not relevant to what follows, we assume that J has the form indicated in the introduction.

be sensitive to perturbations. Thus, in practice, it is not easy to determine whether or not W is strongly stable using the KGL criterion. From a computational point of view, it is better to proceed as follows. Let

$$S_0(W) = \frac{1}{2} J \left(W - W^{-1} \right). \tag{2.1}$$

Some simple but important properties of the matrix $S_0 \equiv S_0(W)$ are summarized in the following proposition. These properties can be found in (or deduced from) Theorem 2.1 in [3].

PROPOSITION 2.3.

1. *The matrix S_0 is symmetric and satisfies $W^* S_0 W = S_0$.*
2. *If λ and μ are two eigenvalues of W such that $\lambda\bar{\mu} \neq 1$, then the corresponding eigenvectors x and y are J-orthogonal, i.e., $(Jx, y) = 0$.*
3. *Let the columns of the rectangular matrices X and Y be invariant by W:*

$$WX = XA \quad \text{and} \quad WY = YB.$$

If the eigenvalues $\lambda_i(A)$ and $\lambda_j(B)$ of A and B, respectively, satisfy $\lambda_i(A)\bar{\lambda}_j(B) \neq 1$ for all i and j, then $X^* J Y = X^* S_0 Y = 0$.
4. *If the symmetric matrix S_0 is positive (negative) definite, then all eigenvalues of W lie on the unit circle.*

It is clear that S_0 is singular if and only if W has eigenvalues ± 1. Moreover, if (λ, x) is an eigenpair of W with $\lambda = e^{i\theta},\ \theta \in (-\pi, \pi]$, then

$$(S_0 x, x) = \sin\theta (iJx, x). \tag{2.2}$$

It follows therefore that if $(S_0 x, x) > 0$, then $\lambda = e^{i|\theta|}$ is an eigenvalue of the first kind, whereas $\lambda = e^{-i|\theta|}$ is of the second kind, and, when $(S_0 x, x) < 0$, then $\lambda = e^{-i|\theta|}$ is of the first kind and $\lambda = e^{i|\theta|}$ is of the second kind. Note that if $\theta = 0$ or $\theta = \pi$, then $\lambda = \pm 1$. Such eigenvalues are necessarily of mixed kind. Indeed, the corresponding eigenvector x is real and satisfies $(Jx, x) = (x, J^* x) = -(x, Jx) = -(Jx, x)$, whence $(Jx, x) = 0 = (S_0 x, x)$. Therefore the matrix S_0 is nonsingular when the KGL criterion is satisfied. It was shown in [1] that the KGL criterion is equivalent to the existence of a symmetric positive definite matrix $\widehat{S}$ whose construction shows that the quadratic form $(S_0 x, x)$ is either positive or negative definite. This leads us to a different classification of the spectrum of W.

DEFINITION 2.4. *An eigenvalue λ of W on the unit circle is an r-eigenvalue (eigenvalue with a red color) if $(S_0 x, x) > 0$ for all corresponding eigenvectors x. It is a g-eigenvalue (eigenvalue with a green color) if $(S_0 x, x) < 0$ for all corresponding eigenvectors x.*

The classification given in Definition 2.4 appears more convenient in practice than that of Definition 2.1 since it deals with symmetric matrices and avoids complex vectors. The invariant subspace associated with an r-eigenvalue (g-eigenvalue) can be chosen real. The main difference between Definitions 2.1 and 2.4 is as follows: if $\lambda = e^{i|\theta|}$ and $\bar{\lambda} = e^{-i|\theta|}$ are eigenvalues of W of, respectively, the first and second kinds, then it easily follows from (2.2) and Proposition 2.3 that $(S_0 z, z)$ has the same sign for all z in the invariant subspace associated with λ and $\bar{\lambda}$. Therefore the eigenvalues λ and $\bar{\lambda}$ will have the same color (green or red). On the other hand, the eigenvector x is associated to a mixed eigenvalue if and only if $(Jx, x) = 0$, which, by (2.2), turns out to be equivalent to $(S_0 x, x) = 0$ (note that if $\theta = 0$ or $\theta = \pi$, then

$S_0 x = 0$). Therefore the classification in Definition 2.4 does not modify the mixed eigenvalues, to which no color will be assigned.

Assume that the spectrum of W lies on the unit circle and is formed only by the r- and g-eigenvalues. Let us denote by P_r (P_g) the projector associated to the r- (g-) eigenvalues. Then, $P_1 = I = P_g + P_r$ and Proposition 2.3 yields $P_r^* J P_g = 0$ and $P_r^* S_0 P_g = 0$. It follows that

- $P_r^* J = J P_r = P_r^* J P_r$ and $P_g^* J = J P_g = P_g^* J P_g$;
- $P_r^* S_0 = S_0 P_r = P_r^* S_0 P_r$ and $P_g^* S_0 = S_0 P_g = P_g^* S_0 P_g$.

The KGL criterion can be reformulated as follows.

THEOREM 2.5. *The KGL criterion is equivalent to* $P_1 = I = P_g + P_r$.

Proof. If the KGL criterion is satisfied, then it is clear that $P_0 = P_\infty = 0$ and $P_1 = I$. Moreover, the eigenvalues $\lambda = e^{i\theta}$ of W are either of the first or second kind. In particular, $e^{i\theta} \neq \pm 1$, i.e., $\sin\theta \neq 0$. It follows from (2.2) that $(S_0 x, x)$ takes nonzero values of the same sign for all $x \in \mathrm{Null}\,(W - \lambda I)$, which, according to Definition 2.4, means that λ is of either red or green color. Therefore $P_1 = I = P_g + P_r$.

Conversely, the condition $P_1 = I = P_g + P_r$ means that the spectrum of W lies on the unit circle and is formed only by r- and/or g-eigenvalues. Therefore if $\left(e^{i\theta}, x\right)$ is an eigenpair of W, then either $(S_0 x, x) > 0$ or $(S_0 x, x) < 0$, which, by (2.2), implies that either $\sin\theta\,(iJx, x) > 0$ or $\sin\theta\,(iJx, x) < 0$. This means that $\sin\theta \neq 0$, i.e., $e^{i\theta} \neq \pm 1$ (i.e., S_0 is nonsingular), and $(iJx, x) > 0$ or $(iJx, x) < 0$. Thus the second condition of the KGL criterion is fulfilled. ∎

Theorem 2.5 shows that W is strongly stable if and only if its spectrum lies on the unit circle and is formed only by r- and/or g-eigenvalues. In practice, the strong stability requires that the r- and g-eigenvalues should be well separated from each other and from ± 1. The properties of P_r and P_g and Theorem 2.5 imply that

- $S_0 = P_r^* S_0 P_r + P_g^* S_0 P_g$;
- $P_r^* S_0 P_r \geq 0$ and $P_g^* S_0 P_g \leq 0$;
- $\mathrm{rank}\,(P_r^* S_0 P_r) = \mathrm{rank}\,(P_r) \equiv \mathrm{tr} P_r$ and $\mathrm{rank}\left(P_g^* S_0 P_g\right) = \mathrm{tr} P_g$;
- $P_r - P_g$ is nonsingular and $\|P_r\| = \|P_g\|$;
- $P_r^* S_0 P_r - P_g^* S_0 P_g > 0$.

In the next subsection, we discuss the use of spectral dichotomy methods to analyze the spectral structure of symplectic matrices along the lines described above.

2.1. Spectral structure of symplectic matrices. First of all, note that if λ_0 is an eigenvalue of the J-symplectic matrix W, then so are λ_0^{-1}, $\bar{\lambda}_0$, and $\bar{\lambda}_0^{-1}$. Assuming $\lambda_0 \neq \pm 1$ and using the identity

$$\frac{\lambda_0 - 1}{\lambda_0 + 1} = -\frac{1/\lambda_0 - 1}{1/\lambda_0 + 1}, \quad \frac{\bar{\lambda}_0 - 1}{\bar{\lambda}_0 + 1} = \left(\frac{\overline{\lambda_0 - 1}}{\lambda_0 + 1}\right),$$

we see that

$$\left|\frac{\lambda_0 - 1}{\lambda_0 + 1}\right| = \left|\frac{1/\lambda_0 - 1}{1/\lambda_0 + 1}\right| = \left|\frac{\bar{\lambda}_0 - 1}{\bar{\lambda}_0 + 1}\right| = \left|\frac{1/\bar{\lambda}_0 - 1}{1/\bar{\lambda}_0 + 1}\right|.$$

Thus, the eigenvalues of W are on some circle of the equation

$$\left|\frac{\lambda - 1}{\lambda + 1}\right| = \xi \quad (0 < \xi < \infty).$$

Two eigenvalues λ and μ of W such that $\lambda\bar{\mu} = 1$ are on the same circle since $\lambda\bar{\mu} = 1$ implies that $\left|\frac{\mu-1}{\mu+1}\right| = \left|\frac{\lambda-1}{\lambda+1}\right|$. The idea is to gather all the eigenvalues belonging to

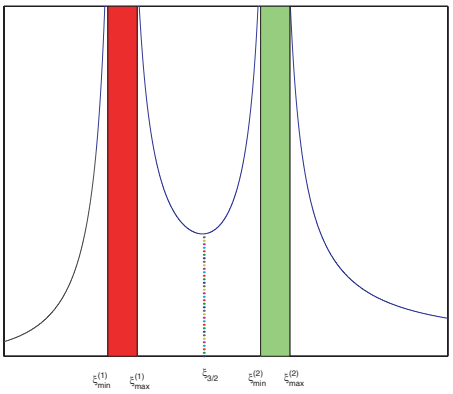


FIG. 2.1. *Shape of a spectral portrait. (In the print version of this figure, the dark shaded area represents red, and the light shaded area represents green.)*

such circles. The eigenvectors and invariant subspaces associated to eigenvalues on different circles are J-orthogonal (Proposition 2.3).

Using Algorithm 1 of [4], we plot the one-dimensional spectral portrait

$$\xi \mapsto \|E(\xi)\|,$$

where

$$E(\xi) = \frac{1}{2\pi}\int_0^{2\pi}\left(I - e^{i\theta}\frac{A^*}{\xi}\right)^{-1}\left(I - e^{-i\theta}\frac{A}{\xi}\right)^{-1} d\theta, \quad A = (P_1W + I)^{-1}(P_1W - I).$$

In general, the spectral portrait looks like Figure 2.1. On the real ξ-axis, we consider the interval of the form $\xi_{\min}^{(j)} < \xi < \xi_{\max}^{(j)}$ with $\|E(\xi)\| > E_{\max}$, where $E_{\max}$ is a large number, and the corresponding projector $P^{(j)}$ onto the invariant subspace associated to the nonzero eigenvalues of P_1W in the annulus

$$\xi_{\min}^{(j)} < \left|\frac{\lambda - 1}{\lambda + 1}\right| < \xi_{\max}^{(j)}.$$

Note that the number of nonzero eigenvalues of P_1W in this annulus equals $n^{(j)} = \mathrm{tr}P^{(j)}$.

We also construct the symmetric matrix

$$S^{(j)} = \left(P^{(j)}\right)^* S_0 P^{(j)},$$

whose number of nonzero eigenvalues is at most equal to $n^{(j)}$.

If $n^{(j)}$ eigenvalues of $S^{(j)}$ are positive (negative) we will say that the interval $\xi_{\min}^{(j)} < \xi < \xi_{\max}^{(j)}$ is of red (green) color. If $S^{(j)}$ has $n^{(j)}$ positive and negative eigenvalues, then the interval $\xi_{\min}^{(j)} < \xi < \xi_{\max}^{(j)}$ is indefinite.

If there are indefinite intervals, then the spectrum of W is not on the unit circle (see [7, Chap. III, sect. 3]).

When all the intervals have only definite colors (red or green) and the intervals with different colors are well separated, then W is strongly stable (see Theorem 2.5

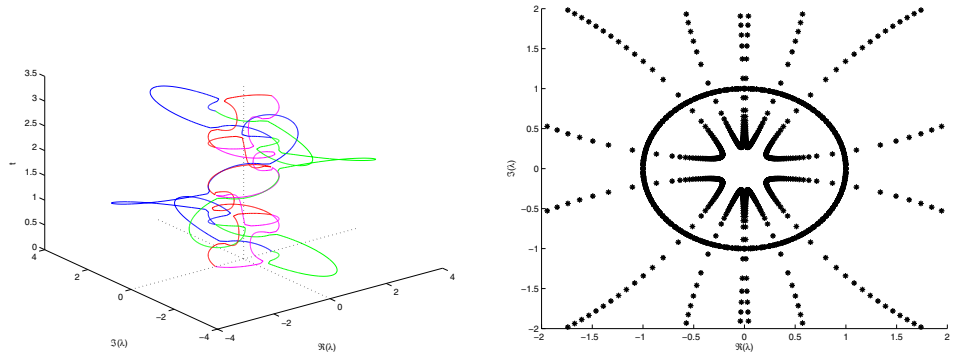


FIG. 2.2. *Movement of eigenvalues of* $W(t), t \in [0, \pi]$ *(left) and spectrum of* $W(t), t \in [0, 2\pi]$ *(right).*

and the discussion thereafter). The projectors P_r and P_g are then given by

$$P_r = \sum_{j \text{ with } S^{(j)} \geq 0} P^{(j)} \quad \text{and} \quad P_g = \sum_{j \text{ with } S^{(j)} \leq 0} P^{(j)}.$$

In each region $\xi_{\max}^{(j)} < \xi < \xi_{\min}^{(j+1)}$ which separates two intervals of different colors, we find the points $\xi = \xi_{j+\frac{1}{2}}$, where $\|E(\xi)\|$ has a local minimum (these points separate the r- and g-eigenvalues), and use $\|E_m(W)\| = \max_j \|E(\xi_{j+\frac{1}{2}})\|$ as a criterion for the distance between r- and g-eigenvalues. This criterion is, however, not sufficient since the eigenvalues should also be separated from ± 1, i.e., the matrix $S_0(W)$ should be well conditioned. Therefore, a good criterion is $\Phi(W) < \infty$, where

$$\Phi(W) = \max\left(\|E_m(W)\|, \|S_0(W)\|\|S_0^{-1}(W)\|\right). \tag{2.3}$$

In practice, the condition $\Phi(W) < \infty$ is replaced by $\Phi(W) < tol$, where tol is a "small" quantity.

2.2. Example. Let

$$b(t) = 2\sin 4t, \quad \alpha(t) = \frac{\pi}{2}\cos t \quad (0 \leq t \leq 2\pi),$$

$$9ptB(t) = \begin{pmatrix} b(t) & b(t)-1 \\ b(t)+1 & b(t) \end{pmatrix}, \quad W(t) = \begin{pmatrix} \cos\alpha(t)B(t) & -\sin\alpha(t)B(t) \\ \sin\alpha(t)B(t) & \cos\alpha(t)B(t) \end{pmatrix}.$$

The matrix $W(t)$ is J-symplectic for all t where $J = \operatorname{diag}(J_2, J_2)$. Its eigenvalues are shown in Figure 2.2.

Figures 2.3–2.5 illustrate the spectral structure of $W(t)$, where t is around 0.13. More precisely, Figure 2.3 (top left and right) shows the spectral portraits of $W(t)$, i.e., the graphs of function

$$\rho \mapsto \|F_{W(t)}(\rho)\| \text{ with } F_{W(t)}(\rho) = \frac{1}{2\pi}\int_0^{2\pi} \left(I - e^{i\theta}\frac{W(t)^*}{\rho}\right)^{-1}\left(I - e^{-i\theta}\frac{W(t)}{\rho}\right)^{-1} d\theta$$

for $t = 0.13069$ and $t = 0.13089$.

From these graphs we see that the eigenvalues lie on the unit circle, i.e., $P_1(t) = I$. Likewise, Figure 2.3 (bottom left and right) shows the spectral portraits of $W(t)$

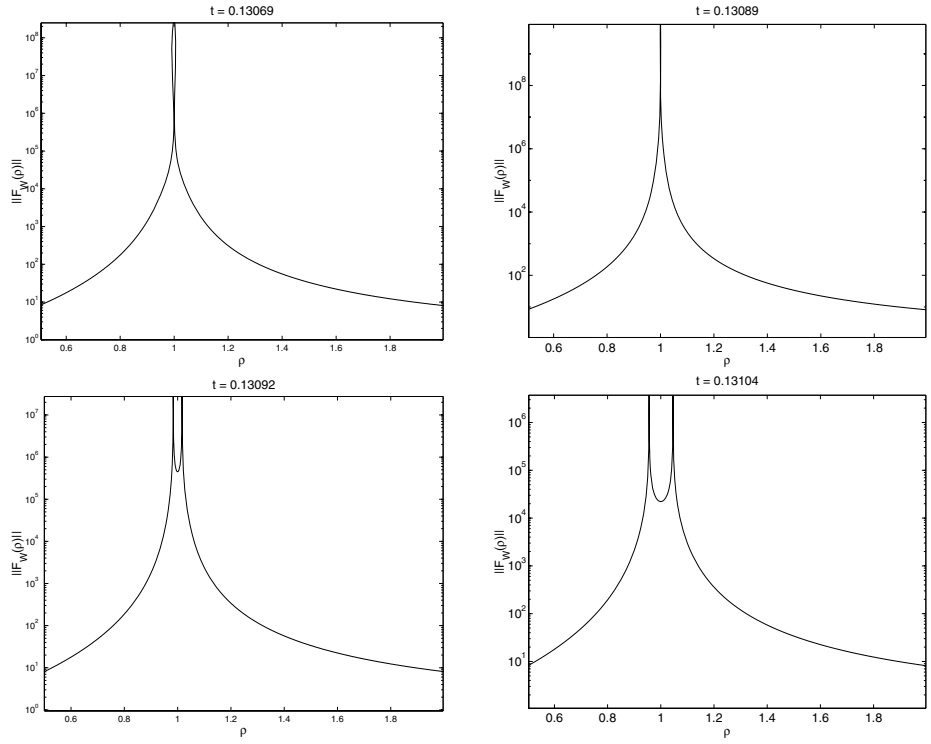


FIG. 2.3. *Spectral portraits of* $W(t)$ *for* $t \approx 0.13$.

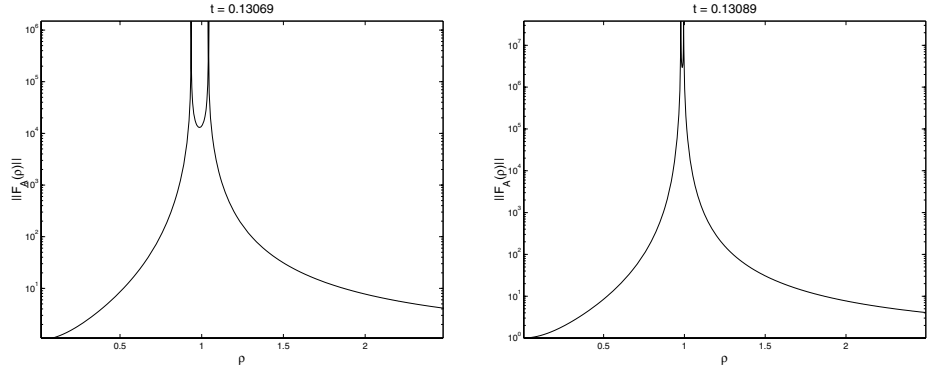


FIG. 2.4. *Spectral portraits of* $A(t)$ *for* $t \approx 0.13$.

for $t = 0.13092$ and $t = 0.13104$, and the graphs show that the spectrum is now outside the unit circle, i.e., $P_1(t) = 0$. Figure 2.4 shows the spectral portraits of $A(t) = (W(t) + I)^{-1}(W(t) - I)$ for $t = 0.13069$ and $t = 0.13089$. As explained in section 2.1, these graphs help to find annuli which contain eigenvalues λ of $W(t)$ such that $|\lambda - 1|/|\lambda + 1| = \text{const}$. The projectors P_r and P_g are obtained from the ones associated to eigenvalues in these annuli. Figure 2.5 shows the evolution of eigenvalues of $W(t)$, where $t \approx 0.13$. Figure 2.5 (top left and right) shows that $W(t)$ has two r- and two g-eigenvalues. Figure 2.5 (bottom left and right) shows that the r- and g-

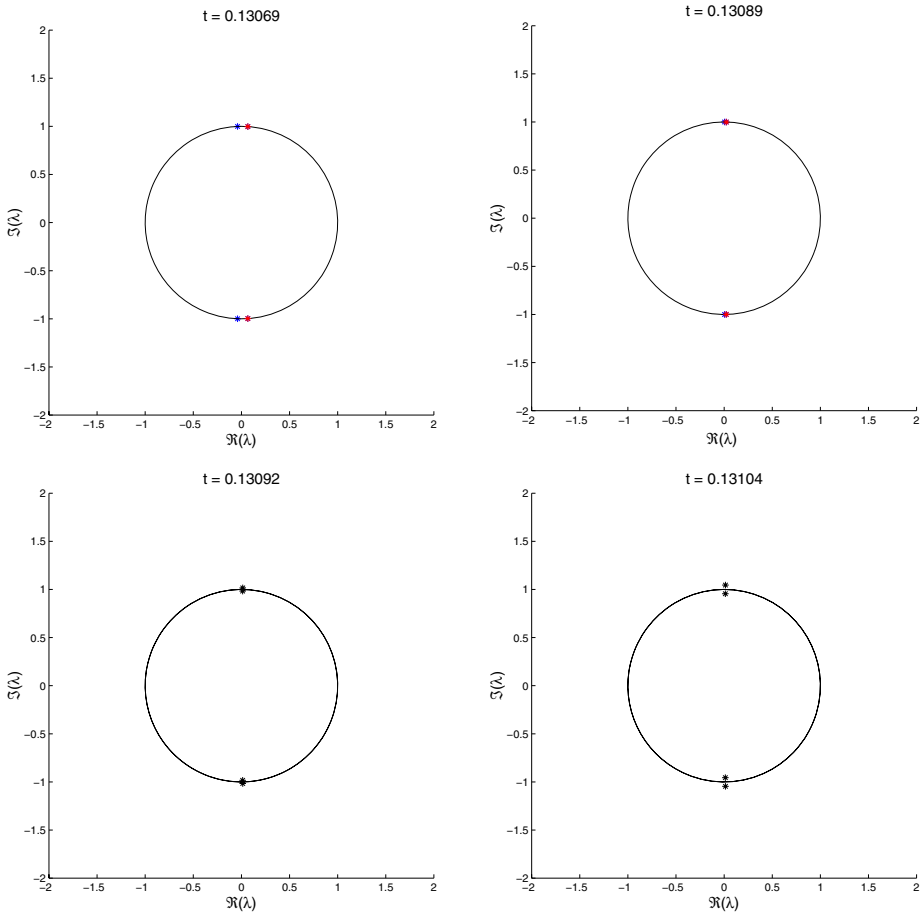


FIG. 2.5. *Eigenvalues of* $W(t)$ *for* $t \approx 0.13$.

eigenvalues move off the unit circle after a meeting (of r- and g-eigenvalues) produced for a certain t between 0.13089 and 0.13092.

Let us discuss the cases when $t = 0.13069$, $t = 0.1308996937$, and $t = 0.13104$ in more detail.

- At $t = 0.13069$:
  - The computed projectors are $P_1 = I$, $P_0 = P_\infty = 0$. See the spectral portrait of $W(t)$ in Figure 2.3 (top left).
  - The computed matrix S_0 is given by

$$S_0 = \begin{pmatrix} -2.68\ 10^{-2} & -2.60\ 10^{-18} & 2.22\ 10^{-16} & 9.98\ 10^{-1} \\ 3.47\ 10^{-18} & -1.92\ 10^{-5} & -9.98\ 10^{-1} & -3.92\ 10^{-17} \\ -3.33\ 10^{-16} & -9.98\ 10^{-1} & -2.68\ 10^{-2} & -1.73\ 10^{-18} \\ 9.98\ 10^{-1} & 1.15\ 10^{-17} & 2.60\ 10^{-18} & -1.92\ 10^{-5} \end{pmatrix},$$

    $\|S_0\|\|S_0^{-1}\| = 1.027$.

  - The computed projectors $P^{(j)}$, $j = 1, 2$, corresponding to the eigenvalues of $W(t)$ in the annuli $0.468 < \left|\frac{\lambda-1}{\lambda+1}\right| < 0.989$ and $0.989 < \left|\frac{\lambda-1}{\lambda+1}\right| < 1.54$

are given by

$$P^{(1)} = \begin{pmatrix} 5.00\ 10^{-1} & -1.03\ 10^{-9} & 1.29\ 10^{-15} & 1.34\ 10^{-2} \\ -1.44\ 10^{-6} & 5.00\ 10^{-1} & -1.87\ 10^{+1} & -1.05\ 10^{-15} \\ -1.05\ 10^{-15} & -1.34\ 10^{-2} & 5.00\ 10^{-1} & -1.03\ 10^{-9} \\ 1.87\ 10^{+1} & 1.29\ 10^{-15} & -1.44\ 10^{-6} & 5.00\ 10^{-1} \end{pmatrix},$$

$$P^{(2)} = I - P^{(1)},$$

$$\mathrm{tr}P^{(1)} = \mathrm{tr}P^{(2)} = 2, \quad P^{(1)}P^{(2)} = 0.$$

See the spectral portrait of $A(t)$ in Figure 2.4 (left). We have $E_m(W) = 3.35\ 10^4$ and therefore the quantity $\Phi(W)$ defined in (2.3), at $t = 0.13068$, is given by $\Phi(W) = \max(E_m(W), \|S_0\|\|S_0^{-1}\|) = 3.35\ 10^4$.

– The nonzero eigenvalues of $S^{(j)} \equiv \left(P^{(j)}\right)^* S_0 P^{(j)}$, $j = 1, 2$, are, respectively, $1.8649\ 10^{+1}$ (double) and $-1.8675\ 10^{+1}$ (double).

We conclude that $W(t)$ has two r-eigenvalues and two g-eigenvalues, $P_r = P^{(1)}$, $P_g = P^{(2)}$. See Figure 2.5 (top left). The matrix $W(t)$ is strongly stable.

- At $t = 0.1308996937$:
  – The Euclidean inner products in Definition 2.1 are all of order 10^{-5}, and one might wonder if the KGL criterion is satisfied or not.
  – The computed projectors and the matrix S_0 are such that $P_1 = I$, $P_\infty = P_0 = 0$, $\|S_0\|\|S_0^{-1}\| = 1.0272$:

$$P_r = \begin{pmatrix} 5.00\ 10^{-1} & -1.34\ 10^{-13} & -8.95\ 10^{-14} & 1.31\ 10^{-5} \\ -1.93\ 10^{-4} & 5.00\ 10^{-1} & -1.90\ 10^{4} & -4.45\ 10^{-13} \\ -4.45\ 10^{-13} & -1.31\ 10^{-5} & 5.00\ 10^{-1} & -1.33\ 10^{-13} \\ 1.90\ 10^{4} & -8.96\ 10^{-14} & -1.93\ 10^{-4} & 5.00\ 10^{-1} \end{pmatrix},$$

$$P_g = I - P_r, \quad \mathrm{tr}P_r = \mathrm{tr}P_g = 2.$$

  – The nonzero eigenvalues of $P_r^* S_0 P_r$ and $P_g^* S_0 P_g$ are, respectively, $1.9014\ 10^4$ (double) and $-1.9014\ 10^4$ (double). We conclude that $W(t)$ is strongly stable and has two r-eigenvalues and two g-eigenvalues.
- At $t = 0.13104$:
  – The computed projectors are

$$P_1 = 0,$$

$$P_\infty = \begin{pmatrix} 5.00\ 10^{-1} & 1.12\ 10^{-2} & -8.19\ 10^{-15} & -9.40\ 10^{-18} \\ 2.22\ 10^{+1} & 5.00\ 10^{-1} & 1.87\ 10^{-14} & 8.19\ 10^{-15} \\ -7.88\ 10^{-15} & -6.88\ 10^{-18} & 5.00\ 10^{-1} & 1.12\ 10^{-2} \\ 1.36\ 10^{-14} & 7.88\ 10^{-15} & 2.22\ 10^{+1} & 5.00\ 10^{-1} \end{pmatrix},$$

$$P_0 = I - P_\infty, \quad \mathrm{tr}P_0 = \mathrm{tr}P_\infty = 2.$$

  – See the spectral portrait of $W(t)$ in Figure 2.3 (bottom right). The spectrum is shown on Figure 2.5 (bottom right). The matrix $W(t)$ is not stable.

3. Stability of linear Hamiltonian systems. Consider a linear Hamiltonian system with T-periodic coefficients, i.e., a differential equation of the form

$$J\frac{dx(t)}{dt} = H(t)x(t), \tag{3.1}$$

where the matrix $H(t)$ is $2N \times 2N$ real symmetric and T-periodic: $H(t+T) = H(t) = (H(t))^*$. It is known that the matrizant $X(t)$ of (3.1), i.e., the fundamental matrix solution of (3.1) defined by the initial condition $X(0) = I$, is J-symplectic for all t, i.e., $X(t)^* J X(t) = J$. Moreover, for all t, $X(t+T) = X(t)X(T)$. Therefore for all real t and integer k,

$$X(t + kT) = X(t)X^k(T). \tag{3.2}$$

DEFINITION 3.1. *Equation* (3.1) *is stable if each of its solutions is bounded on* $(-\infty, +\infty)$. *It is strongly stable if there exists* $\delta > 0$ *such that each Hamiltonian system with* T*-periodic coefficient of the form*

$$J\frac{dx(t)}{dt} = \widetilde{H}(t)x(t)$$

satisfying

$$\int_0^T \|H(t) - \widetilde{H}(t)\|dt < \delta$$

is stable.

It follows from (3.2) that the stability is actually equivalent to the stability of the monodromy matrix $X(T)$, i.e., the matrizant evaluated at the period T [7, p. 162]. It can be shown that the strong stability is equivalent to the strong stability of $X(T)$ [7, p. 196] (see Definition 2.2).

Note that in general the stability or strong stability of (3.1) does not necessarily imply the stability or strong stability of $X(t)$ for $t < T$. In the next theorem, we propose (see properties 2 and 3) some sufficient and easily verifiable conditions that ensure the strong stability of $X(t)$ for $0 < t \leq T$.

To generalize (2.1), we consider the symmetric matrix

$$S(t) = \frac{1}{2}J\left(X(t) - X(t)^{-1}\right) \tag{3.3}$$

defined for $t \geq 0$. Some properties of $S(t)$ and $X(t)$ are summarized in the following theorem.

THEOREM 3.2.

1. *For all* $t \geq 0$

$$X(t)^* S(t) X(t) = S(t) = S(t)^*.$$

$S(t)$ satisfies the differential system

$$\frac{dS(t)}{dt} = \frac{1}{2}\left(H(t)X(t) + X(t)^* H(t)\right), \quad S(0) = 0.$$

2. *In a neighborhood of* 0, *we have*

$$S(t) = tH(0) + \frac{t^2}{2}H'(0) + \mathcal{O}(t^3).$$

Thus, if $H(0) > 0$ *(resp.,* $H(0) < 0$*) and* $S(t)$ *is nonsingular for* $0 < t \leq \tilde{t}$, *then the spectrum of* $X\left(\tilde{t}\right)$ *has only r-eigenvalues (resp., g-eigenvalues).*

3. *If* $\Phi(X(t)) = \max\left(E_m(X(t)), \|S(t)\|\|S^{-1}(t)\|\right) < \infty$ *for* $0 < t \leq T$, *then* $X(t)$ *is strongly stable for* $0 < t \leq T$.
4. *In a neighborhood of* 0, *if the spectrum of* $X(t)$ *lies on the unit circle, then the spectrum of* $J^{-1}H(0)$ *lies on the imaginary axis. From* 2 *we see that this neighborhood can be chosen such that the number of r- and g-eigenvalues of* $X(t)$, *respectively, coincides with the number of positive and negative eigenvalues of* $H(0)$.

Proof.

1. The first property has been mentioned in Proposition 2.3. The second one is easily obtained as

$$\begin{aligned}\frac{dS(t)}{dt} &= \frac{1}{2}\left(J\frac{dX(t)}{dt} + \left(J\frac{dX(t)}{dt}\right)^*\right)\\ &= \frac{1}{2}\left(H(t)X(t) + (H(t)X(t))^*\right).\end{aligned}$$

2.

$$\begin{aligned}\frac{dS}{dt}(0) &= \frac{1}{2}\left(H(0)X(0) + (H(0)X(0))^*\right) = H(0),\\ \frac{d^2S(t)}{dt^2} &= \frac{1}{2}\left[\frac{d}{dt}(H(t)X(t)) + \left(\frac{d}{dt}(H(t)X(t))\right)^*\right],\\ \frac{d}{dt}(H(t)X(t)) &= \frac{dH(t)}{dt}X(t) + H(t)J^{-1}H(t)X(t).\end{aligned}$$

   Hence $\frac{d^2S}{dt^2}(0) = \frac{dH}{dt}(0)$. It is clear that if $H(0) > 0$ ($H(0) < 0$) and $S(t)$ is nonsingular for $0 < t \leq \tilde{t}$, then $S(t)$ will remain positive (negative) definite for all $t \in]0, \tilde{t}]$, and from Proposition 2.3 we obtain that the spectrum of $X(t)$ lies on the unit circle.
3. As explained at the end of section 2.1, the condition "$\Phi(X(t)) < \infty$ for $0 < t \leq T$" means that the r- and g-eigenvalues of $X(t)$ are separated and well separated from ± 1 for $0 < t \leq T$. In particular, the symplectic matrix $X(T)$ and therefore the system (3.1) are strongly stable.
4. Since $J^{-1}S(t) = \frac{1}{2}\left(X(t) - X(t)^{-1}\right)$, it is clear that $J^{-1}S(t)$ and $X(t)$ have the same eigenvectors. Let $e^{i\alpha(t)}$ with $\alpha(t) \in \mathbf{R}$ be an eigenvalue of $X(t)$ corresponding to an eigenvector $u(t)$. Since $X(0) = I$ we obtain by continuity that $\alpha(0) = 2k\pi$ with $k \in \mathbf{Z}$. On the other hand,

$$J^{-1}S(t)u(t) = \frac{1}{2}\left(X(t)u(t) - X(t)^{-1}u(t)\right) = i\sin\alpha(t)u(t).$$

   The derivative of this expression is

$$J^{-1}\left(\frac{dS(t)}{dt}u(t) + S(t)\frac{du(t)}{dt}\right) = i\frac{d\alpha(t)}{dt}\cos\alpha(t)u(t) + i\sin\alpha(t)\frac{du(t)}{dt}.$$

   At $t = 0$, we obtain

$$J^{-1}\frac{dS(0)}{dt}u(0) = i\frac{d\alpha(0)}{dt}u(0)$$

   or

$$J^{-1}H(0)u(0) = i\alpha'(0)u(0). \qquad \square$$

Property 2 in Theorem 3.2 shows in particular that the r- and g-eigenvalues cannot meet in the interval $(0, \tilde{t}]$. If, for example, the conditions mentioned in property 2 are satisfied for $\tilde{t} = T$, then all the symplectic matrices $X(t),\ 0 < t \leq T$, are strongly stable. In particular the strong stability of $X(T)$ means that the system (3.1) is strongly stable (see the discussion after Definition 3.1). However, it is possible that for $t_0 > T$ and $t_0 \neq kT,\ k = 2, 3, \ldots$, the matrix $X(t_0)$ becomes unstable because of a "meeting" of an r-eigenvalue and a g-eigenvalue or because an eigenvalue becomes ± 1. These eigenvalues should move off the unit circle, as in Figure 2.5 (see [7, Chap. III, sect. 3]). In such a case, we say that parametric resonance sets in. The zone around t_0 is a "dangerous" zone where $X(t_0)$ is not strongly stable. In order to detect these points, one can monitor the function $\Phi\left(X(t)\right),\ t > 0$.

3.1. Example. Consider the following differential system:

$$\left\{\begin{array}{l} \frac{d^2\eta_1}{dt^2} + 4\eta_1 + \epsilon\eta_1 \cos 7t + \epsilon\eta_3 \cos 14t = 0, \\ \frac{d^2\eta_2}{dt^2} + 3\eta_2 + \epsilon\eta_3 \sin 35t = 0, \\ \frac{d^2\eta_3}{dt^2} + 2\eta_3 + \epsilon\eta_1 \cos 14t + \epsilon\eta_2 \sin 35t = 0. \end{array}\right. \tag{3.4}$$

Let

$$\eta = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix} \quad \text{and} \quad x = \begin{pmatrix} \eta \\ \frac{d\eta}{dt} \end{pmatrix},$$

where ϵ is a nonnegative parameter.

System (3.4) can be written as a Hamiltonian system with T-periodic coefficients of the form (3.1) with $T = \frac{2\pi}{7} \approx 0.8976$ and

$$H(t) = H_0 + \epsilon H_1(t) \equiv \begin{pmatrix} K_0 + \epsilon K_1(t) & 0 \\ 0 & I \end{pmatrix},$$

where

$$K_0 = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix},\ K_1(t) = \begin{pmatrix} \cos 7t & 0 & \cos 14t \\ 0 & 0 & \sin 35t \\ \cos 14t & \sin 35t & 0 \end{pmatrix}.$$

Since $\epsilon \geq 0$, a simple calculation shows that $H(0) > 0$ if and only if $0 \leq \epsilon < 4$. Figure 3.1 shows the movement of eigenvalues of the matrizant $X_\epsilon(t)$ with $0 \leq t \leq 4T$, $\epsilon = 1$ and $\epsilon = 2$. For these parameters, the function $t \in (0, 4T] \mapsto \Phi(X_\epsilon(t))$ is plotted in Figure 3.2. These figures show that $\Phi(X_\epsilon(t)) < \infty$ for $\epsilon = 1, 2$ and $0 < t \leq T \approx 0.8976$. The zones where $\Phi(X_\epsilon(t))$ is large are emphasized in Figure 3.3. The strong stability is a consequence of Theorem 3.2 (properties 2 or 3).

Remarks.

- The unperturbed system $J\frac{dx(t)}{dt} = H_0 x(t)$ has the monodromy matrix $X_0(T) = e^{TJ^{-1}H_0}$ whose eigenvalues are $\lambda_j\left(X_0(T)\right) = e^{\pm i\omega_j T}$ with $\omega_1 = \sqrt{2},\ \omega_2 = \sqrt{3},\ \omega_3 = 2$.
  According to the terminology used in [7], the numbers $\omega_1,\ \omega_2,\ \omega_3$ are called natural frequencies of the unperturbed system. The numbers

$$\omega_{k,l,m} = \frac{\omega_k + \omega_l}{m}, \quad k, l = 1, 2, 3, \quad m = 1, 2, \ldots,$$

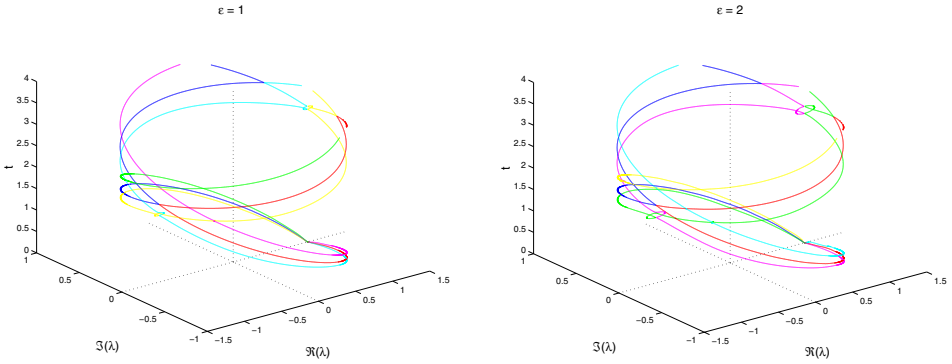


FIG. 3.1. *Movement of eigenvalues of* $X_\epsilon(t)$ *for* $t \in [0, 4T]$, $\epsilon = 1, 2$.

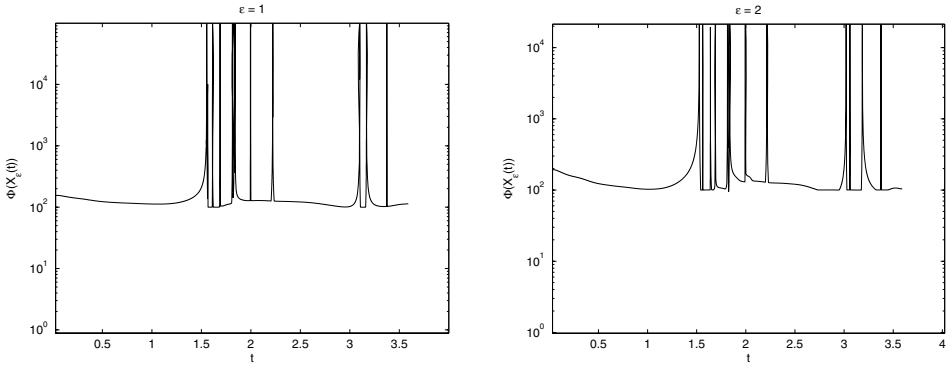


FIG. 3.2. *Function* $t \mapsto \Phi(X_\epsilon(t))$ *for* $t \in (0, 4T]$, $\epsilon = 0, 1, 2$.

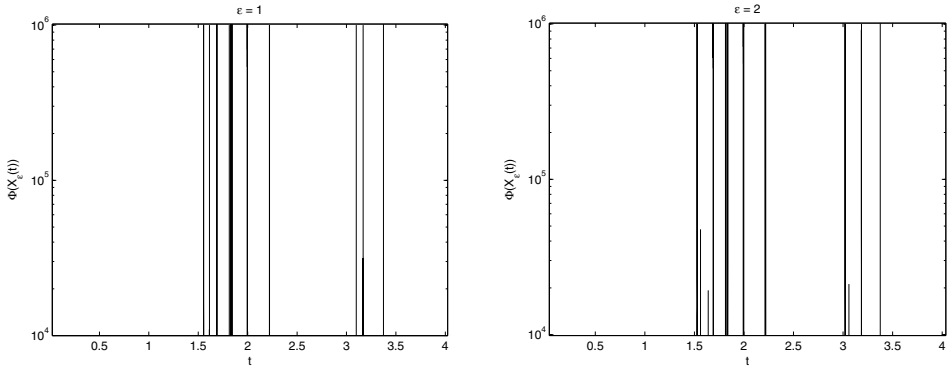


FIG. 3.3. *Values of* t *where* $\Phi(X_\epsilon(t)) \geq 10^4$, $\epsilon = 1, 2$.

are called critical frequencies of the perturbed system (i.e., $\epsilon \neq 0$). They have the following interpretation: for a given ϵ, there is no $\delta = \delta(\omega_{k,l,m})$ such that all the solutions of the unperturbed system are bounded for $0 < \epsilon < \delta$. In other words, there exists a perturbation ϵ such that at t with $t(\omega_k + \omega_l) = 0 \pmod{2\pi}$, $\Phi(X_\epsilon(t))$ is large.

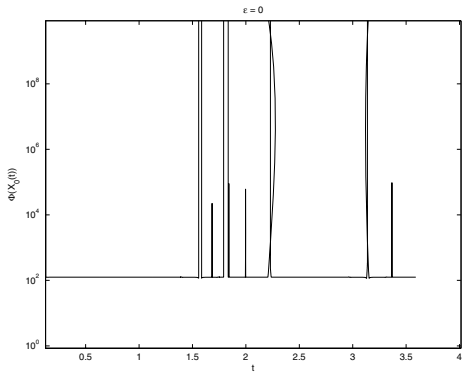


FIG. 3.4. *Function* $t \mapsto \Phi(X_0(t))$ *for* $t \in (0, 4T]$.

- The parameter t for which the resonance sets in corresponds to the situation where $\Phi(X_\epsilon(t)) = \infty$, that is, situations where either $\|E_m(X_\epsilon(t)\| = \infty$ or $S_0(X_\epsilon(t))$ is singular. The control of the first condition is the subject of subsection 2.1. The second condition can be controlled, for example, if $X_\epsilon(t)$ is known analytically. However, this corresponds to the particular case when $X_\epsilon(t)$ has the mixed eigenvalues ± 1. This case is important but is not the general one.
  When $\epsilon = 0$ in the system (3.4), the parameters t where $S_0(X_0(t))$ is singular are given by $t = k\pi/\omega$, where $k \geq 1$ and $\omega = \sqrt{2},\ \sqrt{3},\ 2$. This gives the parameters $t = 1.5708,\ 1.8138,\ 2.2214$, and 3.1416, which are in the interval $(0, 4T]$. Figure 3.4 shows that the function $t \longmapsto \Phi(X_0(t))$ has asymptotes corresponding to these parameters and takes "large but not quite large" values for other parameters. When ϵ is close to 0, the situation remains almost the same. We see from the figure that the zones where $X_\epsilon(t)$ ceases to be strongly stable can be predicted from the case $\epsilon = 0$.

Acknowledgment. The authors wish to thank the referees for their helpful comments and suggestions.

REFERENCES

[1] S. K. GODUNOV, *Stability of iterations of symplectic transformations*, Siberian Math. J., 30 (1989), pp. 54–63.
[2] S. K. GODUNOV, *Modern Aspects of Linear Algebra*, The Scientific Book, Novosibirsk, 1997 (in Russian); English translation, Transl. Math. Monogr. 175, AMS, Providence, RI, 1998.
[3] S. K. GODUNOV AND M. SADKANE, *Numerical determination of a canonical form of a symplectic matrix*, Siberian Math. J., 42 (2001), pp. 629–647.
[4] S. K. GODUNOV AND M. SADKANE, *Some new algorithms for the spectral dichotomy methods*, Linear Algebra Appl., 358 (2003), pp. 173–194.
[5] B. HASSIBI, A. H. SAYED, AND T. KAILATH, *Indefinite-Quadratic Estimation and Control: A Unified Approach to* H^2 *and* H^∞ *Theories*, SIAM Stud. Appl. Math. 16, SIAM, Philadelphia, 1999.
[6] A. N. MALYSHEV, *Parallel algorithm for solving some spectral problems of linear algebra*, Linear Algebra Appl., 188 (1993), pp. 489–520.
[7] V. A. YAKUBOVICH AND V. M. STARZHINSKII, *Linear Differential Equations with Periodic Coefficients*, Vols. 1 and 2, Wiley, New York, 1975.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 4, pp. 1097–1125

ON EIGENVALUE AND EIGENVECTOR ESTIMATES FOR NONNEGATIVE DEFINITE OPERATORS*

LUKA GRUBIŠIĆ†

Abstract. We present a perturbation approach to the Rayleigh–Ritz approximations in the sense of Davis, Kahan, and Weinberger. We restrict ourselves to nonnegative definite self-adjoint operators and obtain sharp bounds of relative type for both eigenvalues and eigenvectors. The operators are allowed to have nontrivial null-spaces, and the test spaces need not be contained in the domain of the considered operator.

Key words. eigenvalues, estimation of eigenvalues, upper and lower bounds, variational methods for eigenvalues of operators

AMS subject classifications. 65F15, 34L15, 35P15, 49R50

DOI. 10.1137/050626533

1. Introduction.

A perturbation approach to Rayleigh–Ritz approximation was introduced by Kahan in [12]. The main idea is to represent the eigenvalues (vectors), which we do not know (but want to approximate), as perturbations of the Ritz values (vectors), which we have computed. This concept lies behind the standard subspace approximation theory of Davis and Kahan [3] and of Davis, Kahan, and Weinberger [4]. In our previous paper [9] we have shown a way to apply this concept to less regular test spaces than those which were considered in [3, 4]. In the present note we continue this study and both improve and generalize the perturbation estimates from [9].

Let us introduce some preliminary notation. Let h be a positive definite symmetric form in a possibly infinite dimensional Hilbert space $\mathcal{H}$. The form h generates the positive definite operator H such that $h(u,v) = (H^{1/2}u, H^{1/2}v)$. The test space for the Rayleigh–Ritz method will be $\mathsf{ran}(X)$, where $X : \mathbb{C}^n \to \mathcal{H}$ is an isometry such that $\mathsf{ran}(X) \subset \mathcal{Q}(H) := \mathcal{D}(H^{1/2})$. Set $P = XX^*$, $P_\perp = \mathbf{I} - XX^*$ and define

- the *block diagonal part of* h as the positive definite form $h'(u,v) = h(Pu, Pv) + h(P_\perp u, P_\perp v)$,
- the *block diagonal part of* H as the operator H' such that $h'(u,v) = (H'^{1/2}u, H'^{1/2}v)$,
- the *Rayleigh quotient* as the matrix $\Xi = (H^{1/2}X)^* H^{1/2}X \in \mathbb{C}^{n\times n}$.

The standard theory of [3, 4] uses

$$\max_{\|x\|=1} |(x, Hx - H'x)| = \|R\| < \infty, \qquad R = HX - X\Xi = HX - H'X \tag{1.1}$$

to obtain spectral estimates. The operator R is called the *residual* of the test subspace $\mathsf{ran}(X)$.

*Received by the editors March 10, 2005; accepted for publication (in revised form) by K. Veselic March 22, 2006; published electronically December 18, 2006. This work was partially supported by grant 0037122 of the Ministry of Science, Education, and Sport, Croatia. This work is a part of the author's Ph.D. thesis which was written under the supervision of Prof. Dr. Krešimir Veselić, Hagen, in partial fulfillment of the requirements for the degree Dr. rer. nat. This research was performed while the author was a member of LG Mathematische Physik, Fernuniversitaet in Hagen.

http://www.siam.org/journals/simax/28-4/62653.html

†Institut für reine und angewandte Mathematik, RWTH Aachen, Templergraben 55, D-52056 Aachen, Germany (luka.grubisic@iram.rwth-aachen.de). On leave from PMF-Department of Mathematics, University of Zagreb, Zagreb, Croatia.

It has already been demonstrated—in [9]—that Kahan's concept can yield nontrivial estimates even when $H - H'$ is not a bona fide operator, that is to say when $\|R\| = \infty$. We now continue the study from [9] and both sharpen the estimates and extend the applicability of the theory to nonnegative H. Our results are generalizations of the known estimates for finite matrices [6, 16]. Familiarity with the paper [9] is not a prerequisite for this work.

As a start we review some finite dimensional results from [6]. For the forms h and h' we have (in our notation)

$$\max_x \frac{|(H^{1/2}x, H^{1/2}x) - (H'^{1/2}x, H'^{1/2}x)|}{(x, H'x)} = \sin\Theta(H^{1/2}X, H^{-1/2}X), \tag{1.2}$$

$$\max_x \frac{|(H^{1/2}x, H^{1/2}x) - (H'^{1/2}x, H'^{1/2}x)|}{(x, Hx)} = \frac{\sin\Theta(H^{1/2}X, H^{-1/2}X)}{1 - \sin\Theta(H^{1/2}X, H^{-1/2}X)}, \tag{1.3}$$

where $\mathsf{sin\Theta}(H^{1/2}X, H^{-1/2}X)$ is the sine of the maximal canonical angle between the subspaces $\mathsf{ran}(H^{1/2}X)$ and $\mathsf{ran}(H^{-1/2}X)$. We will slightly stretch the terminology and (colloquially) call (1.2) and (1.3) the *energy-scaled residual measures*.

Eigenvalue estimates obtained from (1.1) are of the "absolute" type, i.e.,

$$|\lambda - \mu| \leq \|R\|, \tag{1.4}$$

whereas the estimates obtained from (1.2)–(1.3) will be of the "relative" type,

$$|\lambda - \mu| \leq \mu \ \sin\Theta, \qquad |\lambda - \mu| \leq \lambda \ \frac{\sin\Theta}{1 - \sin\Theta}, \tag{1.5}$$

which tacitly supposes that the operator H is nonnegative definite. (It would certainly make sense to obtain similar estimates for indefinite operators as well—a typical application would be, e.g., the Dirac operator in the quantum mechanics. However, related finite dimensional considerations in [20] indicate that this case is technically rather difficult, and our knowledge is far from being exhaustive. This lies in contrast to our nonnegative definite case, where we believe ourselves to have reached a kind of "optimal" answers.)

We identify the following building blocks in (1.5):

- H and H' are considered as symmetric forms $h(u, v) = (H^{1/2}u, H^{1/2}v)$ and $h'(u, v) = (H'^{1/2}u, H'^{1/2}v)$,
- monotonicity of the spectrum implies the estimates.

In [9] the perturbation estimate (1.3) was shown to hold for a *positive definite* operator in an infinite dimensional Hilbert space. We now prove the sharper estimate (1.2) for a nonnegative definite operator in a Hilbert space. That is, we allow H to have a nonzero finite dimensional null-space. This generalization is technically not trivial. We also give an alternative proof of (1.3) as a spinoff and generalize some further results which were derived from (1.3) in reference [9].

The restriction $\|R\| < \infty$, necessary for (1.1) to give useful information in the unbounded operator setting, incurs $\mathsf{ran}(X) \subset \mathcal{D}(H)$. For (1.2) and (1.3) to be applicable we need to assume only

$$\sin\Theta(H^{1/2}X, H^{-1/2}X) < 1.$$

This new residual measure will give nontrivial information even when $\mathsf{ran}(X) \subset \mathcal{Q}(H)$ is such that $\mathsf{ran}(X) \not\subset \mathcal{D}(H)$; see [9] and section 7 of this paper.

Notably, both approaches to measuring the residual share the following property:

TABLE 1.1
Lower estimates for $\lambda_1(H_\eta)$ which can be obtained from the Ritz value $\mu_e = 10^{-2}$ by use of the Temple–Kato estimate and by use of the $\mathsf{sin}\Theta$ approach.

η	Temple–Kato	$\sin\Theta$
1	0.000000999801	0.009004962810
2	0.007500015625	0.009500623831
3	0.008888890261	0.009666851698
4	0.009375000244	0.009750078088
5	0.009600000064	0.009800039988

- $\mathsf{sin}\Theta(H^{1/2}X, H^{-1/2}X) = 0$ if and only if $\mathsf{ran}(X)$ is an invariant subspace of H,
- $R = 0$ if and only if $\mathsf{ran}(X)$ is an invariant subspace of H.

An important feature of our theory is that it gives an abstract framework for a consideration of both eigenvalue and eigenvector estimates. To get a better feeling for the estimate (1.5) consider a simple example. Let

$$H_\eta = \begin{bmatrix} \frac{1}{100} & -\frac{1}{100} \\ -\frac{1}{100} & 1+\eta^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{100} & 0 \\ 0 & \eta^2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \tag{1.6}$$

and $e = \begin{bmatrix} 1 & 0 \end{bmatrix}^*$. We will analyze an approximation of the first eigenvalue of the matrix H_η by the Ritz value $\mu_e = (e, H_\eta e) = 10^{-2}$ for η large.

As a starting point for developing a practical procedure to compute the estimates (1.5) we use the formula

$$\sin^2\Theta(H^{1/2}X, H^{-1/2}X) = \max_{x \in \mathsf{ran}(X)} \frac{(x, H^{-1}x) - (x, H'^{-1}x)}{(x, H^{-1}x)}, \tag{1.7}$$

which is implicit in [9, section 4]. Since

$$H_\eta'^{-1} = \begin{bmatrix} 100 & 0 \\ 0 & \frac{1}{1+\eta^2} \end{bmatrix}, \qquad H_\eta^{-1} = \begin{bmatrix} 100+\eta^{-2} & \eta^{-2} \\ \eta^{-2} & \eta^{-2} \end{bmatrix}$$

we compute, with the help of (1.7),

$$\begin{aligned} \lambda_1(H_\eta) &= \frac{1 + 50\,\eta^2 - \sqrt{1 + 2500\,\eta^4}}{100}, \\ \lambda_2(H_\eta) &= \frac{1 + 50\,\eta^2 + \sqrt{1 + 2500\,\eta^4}}{100}, \\ \sin\Theta(H_\eta^{1/2}e, H_\eta^{-1/2}e) &= \frac{1}{\sqrt{100\eta^2 + 1}}. \end{aligned}$$

As a comparison we will use an estimate which can be obtained from the Temple–Kato inequality from [19]; see (1.8) below. The obtained lower bounds for $\lambda_1(H_\eta)$ are displayed in Table 1.1.

We can observe in Table 1.1 the same behavior which was shown on an infinite dimensional model problem from [9]. Namely, the estimate

$$(1 - \sin\Theta)\mu_e \le \lambda_1,$$

which is linear in $\sin\Theta$, outperforms the estimate

$$\mu_e - \frac{\|H_\eta e - H'_\eta e\|^2}{\lambda_2 - \mu} \le \lambda_1, \tag{1.8}$$

which is quadratic in $\|H_\eta e - H'_\eta e\|$.

As an infinite dimensional analogue of (1.6) we consider the following operator. Let $\chi_{[1,2]}$ be the characteristic function of the interval $[1,2] \subset \mathbb{R}$. We consider $\mathbf{H}_\eta$, which is defined by

$$(\mathbf{H}_\eta^{1/2} u, \mathbf{H}_\eta^{1/2} v) = \int_0^2 (1 + \eta^2 \chi_{[1,2]}) u' v' \, dx, \tag{1.9}$$

and we choose

$$u_1(x) = \begin{cases} \sqrt{2}\sin(\pi x), & 0 \le x \le 1, \\ 0, & 1 \le x, \end{cases} \tag{1.10}$$

as a test function. Now $u_1 \in \mathcal{Q}(\mathbf{H}_\eta)$ but $u_1 \notin \mathcal{D}(\mathbf{H}_\eta)$, so neither of the Temple–Kato estimates (for eigenvectors or eigenvalues) applies, since $\|\mathbf{H}_\eta u_1 - \mu u_1\| = \infty$.

Improved eigenvalue and eigenvector approximation estimates can be summed up in the following procedure:[1]

- Let $\mathbf{H}$ be positive definite, and let P be an orthogonal projection such that $\mathsf{ran}(P) \subset \mathcal{Q}(\mathbf{H})$ and $n = \mathsf{dim}\ \mathsf{ran}(P) < \infty$.
- If $\sin\Theta < 1$ (as defined by (1.7)), then there exist n eigenvalues of the operator $\mathbf{H}$ which are approximated by the n Ritz values from the subspace $\mathsf{ran}(P)$ in the sense of (1.5).
- If $\frac{\sin\Theta}{1-\sin\Theta} < \frac{\lambda_{n+1}-\mu_n}{\lambda_{n+1}+\mu_n}$, then the Ritz values from the subspace $\mathsf{ran}(P)$ approximate the first n eigenvalues of $\mathbf{H}$ (counting the eigenvalues according to their multiplicities), and we have an eigenvector estimate. (Analogous estimates hold for any other contiguous spectral interval.)

2. The notation and preliminaries. The environment in this article will be a Hilbert space $\mathcal{H}$, with the scalar product $(\cdot,\cdot)$. The scalar product is antilinear in the first variable and linear in the second. We start with a closed symmetric form $h(\cdot,\cdot)$, which is additionally assumed to be *nonnegative*:

$$h[u] = h(u,u) \ge 0, \qquad u \in \mathcal{Q}(h). \tag{2.1}$$

Here $\mathcal{Q}(h)$ denotes the domain of the form h. In what follows, when we say the nonnegative form h, we shall always mean the closed symmetric form h that satisfies (2.1). The form h shall be called *positive definite* when it is closed symmetric and there exists $m_h > 0$ such that

$$h[u] = h(u,u) \ge m_h \|u\|^2, \qquad u \in \mathcal{Q}(h).$$

There is also an equivalent operator version of these definitions. The self-adjoint operator $\mathbf{H}$ is called *nonnegative* if

$$(u, \mathbf{H}u) \ge 0, \qquad u \in \mathcal{D}(\mathbf{H}).$$

[1]Here we have assumed that we are approximating the lower end of the spectrum. Analogous procedures can be formulated for other contiguous spectral intervals.

Subsequently, $\mathbf{H}$ is called *positive definite* if there exists $m_{\mathbf{H}} > 0$ such that

$$(u, \mathbf{H}u) \geq m_{\mathbf{H}} \|u\|^2, \qquad u \in \mathcal{D}(\mathbf{H}).$$

In this chapter we assume $\overline{\mathcal{Q}}^{\,\mathcal{H}} = \mathcal{H}$, but later we shall also allow $\overline{\mathcal{Q}}^{\,\mathcal{H}}$ to be any nontrivial subspace of $\mathcal{H}$. For nonnegative self-adjoint operators one defines, with the help of the spectral theorem, the usual functional calculus. We write the spectral decomposition of the self-adjoint operator $\mathbf{H}$ as

$$\mathbf{H} = \int \lambda \, \mathrm{d}E_{\mathbf{H}}(\lambda),$$

where $E_{\mathbf{H}}(\lambda)$ is the right continuous *spectral family* associated with the operator $\mathbf{H}$. When there can be no confusion we simply write $E(\lambda)$.

The representation theorem for nonnegative forms [13, p. 331] implies that there exists a self-adjoint operator $\mathbf{H}$ such that $\mathcal{D}(\mathbf{H}^{1/2}) = \mathcal{Q}(h)$ and

$$h(u, v) = (\mathbf{H}^{1/2}u, \mathbf{H}^{1/2}v), \qquad u, v \in \mathcal{Q}(h).$$

Following [7], we call $\mathcal{D}(\mathbf{H})$ the *operator domain* of $\mathbf{H}$ and $\mathcal{Q}(\mathbf{H}) = \mathcal{D}(\mathbf{H}^{1/2})$ the *quadratic form domain* of $\mathbf{H}$. We write $\mathcal{D}$ and $\mathcal{Q}$ when there can be no confusion. With the help of the spectral theorem we see that

$$\mathcal{D}(\mathbf{H}) = \left\{ u \in \mathcal{H} : \|\mathbf{H}u\|^2 = \int \lambda^2 \, d(E(\lambda)u, u) < \infty \right\},$$
$$\mathcal{Q}(\mathbf{H}) = \left\{ u \in \mathcal{H} : h[u] = \|\mathbf{H}^{1/2}u\|^2 = \int \lambda \, d(E(\lambda)u, u) < \infty \right\}.$$

In general, when dealing with the forms in a Hilbert space we shall follow the terminology of Kato; cf. [13]. In one point we will depart from the conventions in [13]. A nonnegative form

$$h(u, v) = (\mathbf{H}^{1/2}u, \mathbf{H}^{1/2}v)$$

will be called *nonnegative definite* when $\lambda_e(\mathbf{H}) := \inf \sigma_{\mathrm{ess}}(\mathbf{H}) > 0$. Analogously, a nonnegative operator $\mathbf{H}$ such that $\lambda_e(\mathbf{H}) > 0$ will be also called *nonnegative definite*. We will often say nonnegative, meaning the nonnegative definite. Now, we give definitions of some terms that will frequently be used; cf. [7, 13].

DEFINITION 2.1. *A bounded operator $A : \mathcal{H} \to \mathcal{U}$ is called* degenerate *if* $\mathsf{ran}(A)$ *is finite dimensional.*

DEFINITION 2.2. *Let $\mathbf{H}$ and $\mathbf{A}$ be nonnegative operators. We define the* order relation $\leq$ *between the nonnegative operators by saying that*

$$\mathbf{A} \leq \mathbf{H}$$

if $\mathcal{Q}(\mathbf{H}) \subset \mathcal{Q}(\mathbf{A})$ *and*

$$\|\mathbf{A}^{1/2}u\| \leq \|\mathbf{H}^{1/2}u\|, \qquad u \in \mathcal{Q}(\mathbf{H}),$$

or equivalently if

$$a[u] \leq h[u], \qquad u \in \mathcal{Q}(h),$$

when a and h are nonnegative forms defined by the operators $\mathbf{A}$ and $\mathbf{H}$ and $\mathbf{A} \leq \mathbf{H}$.

A main principle that we shall use to develop the perturbation theory will be the *monotonicity of the spectrum* with regard to the order relation between nonnegative operators. This principle can be expressed in many ways. The relevant results, which are scattered over the monographs [7, 13], are summed up in the following theorem; see also [15, Corollary A.1].

THEOREM 2.3. *Let* $\mathbf{A} = \int \lambda \, dE_{\mathbf{A}}(\lambda)$ *and* $\mathbf{H} = \int \lambda \, dE_{\mathbf{H}}(\lambda)$ *be nonnegative operators in* $\mathcal{H}$*, and let* $\mathbf{A} \leq \mathbf{H}$*. By* $0 \leq \mu_1 \leq \mu_2 \leq \cdots < \lambda_e(\mathbf{A})$ *and* $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots < \lambda_e(\mathbf{H})$ *denote the discrete eigenvalues of* $\mathbf{A}$ *and* $\mathbf{H}$*; then*

1. $\lambda_e(\mathbf{A}) \leq \lambda_e(\mathbf{H})$,
2. $\mathsf{dim}\ E_{\mathbf{H}}(\gamma) \leq \mathsf{dim}\ E_{\mathbf{A}}(\gamma)$, *for every* $\gamma \in \mathbb{R}$,
3. $\mu_k \leq \lambda_k$, $k = 1, 2, \ldots$.

We close this introductory section with the well-known theorem about the perturbation of the *essential spectrum*.

THEOREM 2.4. *Let* $\mathbf{H}$ *and* $\mathbf{A}$ *be positive definite operators. If the operator*

$$\mathbf{H}^{-1} - \mathbf{A}^{-1}$$

is compact, then $\sigma_{ess}(\mathbf{H}) = \sigma_{ess}(\mathbf{A})$.

3. The generalized inverse and angle between the subspaces. There are many ways to express that $u \in \mathcal{Q}(h)$ is an eigenvector of the operator $\mathbf{H}$. We will give a geometric characterization of this property. Assume that $\|u\| = 1$ and $\mu = h[u]$. An elementary trigonometric argument yields

$$\|\mathbf{H}^{1/2}u - \mu\mathbf{H}^{-1/2}u\| = 0 \Leftrightarrow \sin\Theta(\mathbf{H}^{1/2}u, \mathbf{H}^{-1/2}u) = 0. \tag{3.1}$$

Equation (3.1) implies that u is an eigenvector of $\mathbf{H}$ if and only if $\mathsf{sin}\Theta(\mathbf{H}^{1/2}u, \mathbf{H}^{-1/2}u) = 0$. The ability to assess the size of $\mathsf{sin}\Theta(\mathbf{H}^{1/2}u, \mathbf{H}^{-1/2}u)$ will be central to the analysis of the Rayleigh–Ritz method in this paper.

In this section we give the background information on the angles between two finite dimensional subspaces of a Hilbert space, as given in [3, 13, 21]. Basic results on generalized inverses of (unbounded) operators defined between two Hilbert spaces will be presented as well. These results will be applied to the problem of computing $\mathsf{sin}\Theta(\mathbf{H}^{1/2}\mathcal{X}, \mathbf{H}^{-1/2}\mathcal{X})$ for the given positive definite $\mathbf{H}$ and some finite dimensional $\mathcal{X} \subset \mathcal{Q}(\mathbf{H})$.

Closed subspaces of the Hilbert space $\mathcal{H}$ can be represented as images of the corresponding orthogonal projections. We shall freely speak about the dimension of the projection P, meaning the dimension of the range of the projection P. In the case in which P is finite dimensional, we have another representation for the subspace $\mathsf{ran}(P)$. For a given n-dimensional subspace $\mathsf{ran}(P) \subset \mathcal{Q}$ there exists an isometry $X : \mathbb{C}^n \to \mathcal{H}$ such that $\mathsf{ran}(P) = \mathsf{ran}(X)$, where $P = XX^*$. Therefore, $\mathsf{ran}(X)$ is an alternative representation of the n-dimensional subspace $\mathsf{ran}(P)$. The isometry X will be called the basis of the subspace $\mathsf{ran}(P)$. We shall freely use both representation of the finite dimensional subspace. $P_X = XX^*$ will generically denote the orthogonal projection on the space $\mathsf{ran}(X)$ (for some isometry $X : \mathbb{C}^n \to \mathcal{H}$).

Let $\mathsf{ran}(P)$ and $\mathsf{ran}(Q)$ be two finite dimensional subspaces of the Hilbert space $\mathcal{H}$. The function $\angle$ that measures the separation of the pair of subspaces $\mathsf{ran}(P)$ and $\mathsf{ran}(Q)$ will be called an *angle function* if it satisfies the following properties:

1. $\angle(P, Q) \geq 0$, and $\angle(P, Q) = 0$ if and only if $\mathsf{ran}(P) \subset \mathsf{ran}(Q)$ or $\mathsf{ran}(Q) \subset \mathsf{ran}(P)$;
2. $\angle(P, Q) = \angle(Q, P)$;

3. $\angle(P,Q) \le \angle(P,R) + \angle(R,Q)$ if $\dim(\mathrm{ran}(P)) \le \dim(\mathrm{ran}(R)) \le \dim(\mathrm{ran}(Q))$ or $\dim(\mathrm{ran}(P)) \ge \dim(\mathrm{ran}(R)) \ge \dim(\mathrm{ran}(Q))$;
4. $\angle(UP,UQ) = \angle(P,Q)$, for any unitary U.

In what follows we will use the following angle functions (see [21]):

$$\Theta(P,Q) = \arcsin\max\{\|P(\mathbf{I}-Q)\|, \|Q(\mathbf{I}-P)\|\}, \tag{3.2}$$

$$\Theta_p(P,Q) = \arcsin\min\{\|P(\mathbf{I}-Q)\|, \|Q(\mathbf{I}-P)\|\}. \tag{3.3}$$

The function $\Theta(P,Q)$ from (3.2) will be called the *maximal canonical angle* between the subspaces P and Q. The function $\Theta_p(P,Q)$ from (3.3) will be called the *maximal principal angle* between the subspaces P and Q.

The following lemma, which is a consequence of [13, Theorem I-6.34], gives an insight into the behavior of the canonical and the principal angles which were defined by (3.2) and (3.3).

LEMMA 3.1. *Let P and Q be two orthogonal projections such that* $\dim(\mathrm{ran}(P)) \le \dim(\mathrm{ran}(Q))$, *and let*

$$\|P(\mathbf{I}-Q)\| < 1;$$

then we have the following alternative. Either

1. $\dim(\mathrm{ran}(P)) = \dim(\mathrm{ran}(Q))$ *and*

$$\sin\Theta(P,Q) = \sin\Theta_p(P,Q) = \|P-Q\| < 1, \qquad \textit{or}$$

2. $\dim(\mathrm{ran}(P)) < \dim(\mathrm{ran}(Q))$ *and*

$$\sin\Theta_p(P,Q) = \|P(\mathbf{I}-Q)\| < 1.$$

For most of our needs, Lemma 3.1 describes the relation between the finite dimensional subspaces $\mathrm{ran}(P)$ and $\mathrm{ran}(Q)$ in sufficient detail. However, sometimes it will be necessary to analyze the structure of the finite dimensional projections $P_V = VV^*$ and $P_U = UU^*$ in further detail. To this end we define the *canonical angles* $\theta_1, \dots, \theta_n$ between the spaces $\mathrm{ran}(U)$ and $\mathrm{ran}(V)$ as

$$\theta_i = \arccos\sigma_i, \quad i = 1, \dots, n, \tag{3.4}$$

where $\sigma_1, \dots, \sigma_n$ are the singular values of the matrix

$$V^*U \in \mathbb{C}^{m\times n}.$$

We have assumed that $m = \dim\ \mathrm{ran}(V)$, $n = \dim\ \mathrm{ran}(U)$, and $m \le n$. The canonical angles are related to the angle function (3.2) through the formula (see [21])

$$\sin\Theta(P_V, P_U) = \max_i \sin\theta_i.$$

We also define the *acute principal angles* $\theta_1^p \le \theta_2^p \le \cdots \le \theta_k^p$, where $k \le n$, as those canonical angles θ_i which satisfy the condition $0 < \theta_i < \pi/2$. Subsequently, we obtain a connection to the angle function (3.3) through the formula

$$\sin\Theta_p(P_V, P_U) = \max_i \sin\theta_i^p.$$

In dealing with the projections and degenerate operators it is useful to have a notion of the generalized inverse. We will use the definition of the generalized inverse of a closed densely defined operator in $\mathcal{H}$ from [18]; see also [13, Chapter IV.5].

DEFINITION 3.2. *Let $\mathbf{T}:\mathcal{H}\to\mathcal{U}$ be a closed operator such that $\overline{\mathcal{D}(\mathbf{T})}=\mathcal{H}$. The operator $\mathbf{T}^\dagger:\mathcal{U}\to\mathcal{H}$ is defined by*

$$\begin{aligned}\mathcal{D}(\mathbf{T}^\dagger)&=\mathsf{ran}(\mathbf{T})\oplus\mathsf{ran}(\mathbf{T})^\perp,\\ \mathbf{T}^\dagger u&=(\,\mathbf{T}\,|_{\mathsf{ker}(\mathbf{T})^\perp})^{-1}P_{\mathsf{ran}(\mathbf{T})}u,\qquad u\in\mathcal{D}(\mathbf{T}^\dagger),\end{aligned}$$

and it is called the Moore–Penrose generalized inverse *of T.*

The properties of the generalized inverse[2] are analyzed in the monograph [18]. In particular we use the following characterization.

THEOREM 3.3 (see [18, Theorem I.5.7]). *Let $\mathbf{T}:\mathcal{H}\to\mathcal{U}$ be the closed operator, and let $\overline{\mathcal{D}(\mathbf{T})}=\mathcal{H}$; then $\mathbf{T}^\dagger$ is the unique closed operator such that*

$$\begin{aligned}\mathbf{T}^\dagger\mathbf{T}\mathbf{T}^\dagger&=\mathbf{T}^\dagger\qquad\text{on }\mathcal{D}(\mathbf{T}^\dagger),\\ \mathbf{T}\mathbf{T}^\dagger&=P_{\mathsf{ran}(\mathbf{T})}\big|_{\mathcal{D}(\mathbf{T}^\dagger)},\\ \mathbf{T}^\dagger\mathbf{T}&=P_{\mathsf{ker}(\mathbf{T})^\perp}\big|_{\mathcal{D}(\mathbf{T})},\end{aligned}$$

where $P_{\mathcal{M}}$ is the orthogonal projection on $\mathcal{M}$. The operator $\mathbf{T}^\dagger$ is bounded if and only if $\mathbf{T}$ has a closed range.

The nonnegative operator $\mathbf{H}^\dagger$ has the spectral decomposition

$$\mathbf{H}^\dagger=\int\frac{1}{\lambda}\,dE(\lambda),\qquad \mathcal{D}(\mathbf{H}^\dagger)=\left\{u\in\mathcal{H}:\int\frac{1}{\lambda^2}\,d(E(\lambda)u,u)<\infty\right\},$$

and the functional calculus implies

$$\mathbf{H}^{\dagger 1/2}=\mathbf{H}^{1/2\dagger}.$$

Theorem 3.3 shows a relation between the Moore–Penrose generalized inverses and orthogonal projections in a Hilbert space. This is precisely the reason why the generalized inverses will be useful in our study.

A bounded operator $W:\mathcal{H}\to\mathcal{U}$ is called *partially isometric* if there exists a closed subspace $\mathcal{M}\subset\mathcal{H}$ such that

$$\|Wu\|=\|P_{\mathcal{M}}u\|,\qquad u\in\mathcal{H}.$$

This is equivalent to

$$W^*W=P_{\mathcal{M}}.$$

The set $\mathcal{M}=\mathsf{ran}(W^*)\subset\mathcal{H}$ is called the *initial set* of the partial isometry W, and $\mathsf{ran}(W)\subset\mathcal{U}$ is called the *final set*. Since $\mathsf{ker}(W^*)\oplus\mathsf{ran}(W)$ we see

$$WW^*=P_{\mathsf{ran}(W)},$$

and so W^* is also the partial isometry with the initial set $\mathsf{ran}(W)$. We shall also use the notation

$$W^*W=P_{W^*},\qquad WW^*=P_W.$$

[2]The generalized inverses can also be defined in more general settings. Their properties are also analyzed in [18].

It is obvious that

$$W^* = W^\dagger,$$

and we have the following lemma.

LEMMA 3.4. *A bounded operator $W : \mathcal{H} \to \mathcal{U}$ is partially isometric if and only if*

$$WW^*W = W.$$

For the proof, see [13].

LEMMA 3.5. *Let V and W be two partial isometries; then*

$$\|P_V P_W\| = \|VP_W\| = \|V^*W\|.$$

Proof. Using Lemma 3.4, we compute

$$\begin{aligned}\|P_V P_W\|^2 &= \operatorname{spr}(P_W P_V P_W) = \operatorname{spr}(WW^*VV^*WW^*)\\ &= \operatorname{spr}(W^*VV^*WW^*W) = \operatorname{spr}(W^*VV^*W) = \|V^*W\|.\end{aligned}$$

In this computation we have used the identity

$$\operatorname{spr}(ABC) = \operatorname{spr}(CAB),$$

which holds for bounded operators A, B, C. ∎

4. Geometrical properties of the Ritz value perturbation. In this section we will present a perturbation approach to the Rayleigh–Ritz approximation of the spectrum of a positive definite operator. The nonnegative definite case is technically more complex and warrants a separate section. Although this chapter is devoted to the positive definite case, some of the statements and definitions will be given in full generality in which they will be later used in the text.

Let $0 \le h$ be a nonnegative form, and let $\mathsf{ran}(X) \subset \mathcal{Q}(h)$ be the n-dimensional test space. The matrix

$$\Xi_{\mathbf{H},X} = (\mathbf{H}^{1/2}X)^*\mathbf{H}^{1/2}X \in \mathbb{C}^{n\times n}$$

will be called the *Rayleigh quotient* associated with the basis X. When there can be no confusion, we shall denote the Rayleigh quotient by Ξ and drop the indices. The eigenvalues of the matrix Ξ will be numbered in the ascending order

$$\mu_1 \le \mu_2 \le \cdots \le \mu_n. \tag{4.1}$$

We call the numbers μ_i the Ritz values of the operator $\mathbf{H}$ (form h) from the subspace $\mathsf{ran}(X)$. This definition is correct since the eigenvalues of the matrix Ξ do not depend on the choice of the basis X. In the rest of this chapter we will use $P = XX^*$ to denote the projection onto the range of the isometry $X : \mathbb{C}^n \to \mathcal{H}$.

For the given h and $\mathsf{ran}(X) \subset \mathcal{Q}(h)$, $P = XX^*$, we define the symmetric forms δh and h' using the formulae

$$\delta h(u,v) = h(Pu,(\mathbf{I}-P)v) + h((\mathbf{I}-P)u,Pv), \qquad u,v \in \mathcal{Q}(h), \tag{4.2}$$

$$h'(u,v) = h(Pu,Pv) + h((\mathbf{I}-P)u,(\mathbf{I}-P)v), \qquad u,v \in \mathcal{Q}(h). \tag{4.3}$$

Obviously, (4.2) and (4.3) imply

$$h'(u,v) = h(u,v) - \delta h(u,v), \qquad u,v \in \mathcal{Q}(h). \tag{4.4}$$

Before we can proceed we need the following definition.

DEFINITION 4.1. *If* $\mathbf{H}$ *is a self-adjoint operator and* P *a projection, to say that* P *commutes with* $\mathbf{H}$ *means that* $u \in \mathcal{D}(\mathbf{H})$ *implies* $Pu \in \mathcal{D}(\mathbf{H})$ *and*

$$\mathbf{H}Pu = P\mathbf{H}u, \qquad u \in \mathcal{D}(\mathbf{H}).$$

In what follows we will describe the properties of the symmetric form h' and of the operator $\mathbf{H}'$ that it generates.

LEMMA 4.2. *Let the nonnegative definite form* h *and the subspace* $\mathsf{ran}(X) \subset \mathcal{Q}$ *be given. Let* $\mathbf{H}$ *be the nonnegative definite operator defined by the form* h*. The form* h' *from* (4.3) *is closed and positive, and it defines the self-adjoint operator* $\mathbf{H}'$*. Furthermore,* $\mathbf{H}'$ *is positive definite if* $\mathbf{H}$ *is positive definite,* $\sigma_{ess}(\mathbf{H}) = \sigma_{ess}(\mathbf{H}')$*, and*

$$\mathbf{H}'X = X\Xi \tag{4.5}$$

for $\Xi = (\mathbf{H}^{1/2}X)^*\mathbf{H}^{1/2}X \in \mathbb{C}^{n\times n}$.

Proof. The operators $\mathbf{H}^{1/2}P$ and $\mathbf{H}^{1/2}(\mathbf{I}-P)$ are closed and so is the form

$$h'(u,v) = h(Pu,Pv) + h((\mathbf{I}-P)u,(\mathbf{I}-P)v).$$

This form is obviously nonnegative, so it defines a nonnegative self-adjoint operator $\mathbf{H}'$. We will now show that the subspace $\mathsf{ran}(X)$ reduces $\mathbf{H}'$. Indeed, for $y \in \mathcal{Q}$, $x \in \mathbb{C}^n$ we have

$$\begin{aligned} h'(y,Xx) &= (\mathbf{H}^{1/2}y, \mathbf{H}^{1/2}Xx) - (\mathbf{H}^{1/2}(\mathbf{I}-P)y, \mathbf{H}^{1/2}Xx) \\ &= (\mathbf{H}^{1/2}XX^*y, \mathbf{H}^{1/2}Xx) \\ &= (\Xi X^*y, x). \end{aligned}$$

This is equivalent to

$$(\mathbf{H}'^{1/2}y, \mathbf{H}'^{1/2}Xx) = (y, X\Xi x), \quad y \in \mathcal{Q},\ x \in \mathbb{C}^n.$$

It implies $\mathsf{ran}(X) \subset \mathcal{D}(\mathbf{H}')$ and

$$(y, \mathbf{H}'Xx - X\Xi x) = 0$$

for all $y \in \mathcal{H}$, $x \in \mathbb{C}^n$. Hence,

$$\mathbf{H}'X = X\Xi, \tag{4.6}$$

which is equivalent to the statement that P commutes with $\mathbf{H}'$ (see Definition 4.1). We now prove that $\sigma_{ess}(\mathbf{H}) = \sigma_{ess}(\mathbf{H}')$. Assume that h is a positive definite form; then h' from (4.3) is positive definite, too. From (4.4) we obtain

$$\delta h(\mathbf{H}^{-1}u, \mathbf{H}'^{-1}v) = (\mathbf{H}'^{-1}u - \mathbf{H}^{-1}u, v), \qquad u,v \in \mathcal{H}.$$

On the other hand,

$$\delta h(\mathbf{H}^{-1}u, \mathbf{H}'^{-1}v) = (\mathbf{H}^{1/2}P\mathbf{H}^{-1}u, \mathbf{H}^{1/2}P_\perp\mathbf{H}'^{-1}v) + (\mathbf{H}^{1/2}P_\perp\mathbf{H}^{-1}u, \mathbf{H}^{1/2}P\mathbf{H}'^{-1}v)$$

defines a compact operator. Theorem 2.4 implies $\sigma_{ess}(\mathbf{H}) = \sigma_{ess}(\mathbf{H}')$, and the statement of the theorem is proved for a positive definite h. In the general case, take $\alpha > 0$. The form $\widetilde{h}(u,v) = h(u,v) + \alpha(u,v)$ is positive definite. Furthermore, we establish

$$\widetilde{h}'(u,v) = \alpha(u,v) + h'(u,v),$$
$$\delta\widetilde{h}(u,v) = \delta h(u,v),$$

and so $\sigma_{ess}(\widetilde{\mathbf{H}}) = \sigma_{ess}(\widetilde{\mathbf{H}}')$. The conclusion $\sigma_{ess}(\mathbf{H}) = \sigma_{ess}(\mathbf{H}')$ follows by the spectral mapping theorem. □

COROLLARY 4.3. *Let the nonnegative definite form h and the subspace $\mathsf{ran}(X) \subset \mathcal{Q}$ be given. The projections $P = XX^*$ and $P_{\mathsf{ran}(\mathbf{H}')}$ commute, and $\mathsf{ker}(\mathbf{H}') \subset \mathsf{ker}(\mathbf{H})$.*

Remark 4.4. For positive definite h Lemma 4.2 describes the operator $\mathbf{H}'$ in sufficient detail. For a general nonnegative h the operator $\mathbf{H}'$ has somewhat more complex structure. Further properties of the operator $\mathbf{H}'$, constructed in the case in which h is a general nonnegative form, will be discussed in section 4.1.

We now concentrate on the positive definite case.

THEOREM 4.5. *Let the subspace $\mathsf{ran}(X) \subset \mathcal{Q}$ be given, and let h be positive definite. Assume $\mathsf{sin}\Theta := \mathsf{sin}\Theta(\mathbf{H}^{1/2}X, \mathbf{H}^{-1/2}X) < 1$; then*

$$(1 - \sin\Theta)h'[u] \le h[u] \le (1 + \sin\Theta)h'[u], \qquad u \in \mathcal{Q}(h), \tag{4.7}$$

$$\left(1 - \frac{\sin\Theta}{1 - \sin\Theta}\right) h[u] \le h'[u] \le \left(1 + \frac{\sin\Theta}{1 - \sin\Theta}\right) h[u], \qquad u \in \mathcal{Q}(h). \tag{4.8}$$

Proof. The product $\mathbf{H}^{1/2}\mathbf{H}'^{-1/2}$ is well defined since $\mathcal{Q} = \mathcal{D}(\mathbf{H}^{1/2}) = \mathcal{D}(\mathbf{H}'^{1/2})$. This implies that the form

$$\delta h_s(x,y) = \delta h(\mathbf{H}'^{-1/2}x, \mathbf{H}'^{-1/2}y)$$

defines the bounded operator δH_s. After the substitutions $u = \mathbf{H}'^{-1/2}x$, $v = \mathbf{H}'^{-1/2}y$ we obtain

$$\max_{u,v \in \mathcal{Q}(h)} \frac{|\delta h(u,v)|}{\sqrt{h'[u]h'[v]}} = \|\delta H_s\|. \tag{4.9}$$

We now show $\|\delta H_s\| = \mathsf{sin}\Theta$. Set

$$V = \mathbf{H}^{1/2}P\mathbf{H}'^{-1/2}, \tag{4.10}$$

$$W = \mathbf{H}^{1/2}P_\perp\mathbf{H}'^{-1/2}, \tag{4.11}$$

with $P_\perp = \mathbf{I} - P$. Relation (4.4) implies

$$\begin{aligned}\delta h(\mathbf{H}'^{-1/2}u, \mathbf{H}'^{-1/2}v) &= h(P_\perp\mathbf{H}'^{-1/2}u, P\mathbf{H}'^{-1/2}v) + h(P\mathbf{H}'^{-1/2}u, P_\perp\mathbf{H}'^{-1/2}v)\\ &= (Wu, Vv) + (Vu, Wv),\end{aligned} \tag{4.12}$$

which can be written as

$$\delta H_s = V^*W + W^*V. \tag{4.13}$$

The equations (4.10)–(4.13) yield

$$VW^* = WV^* = 0, \tag{4.14}$$

$$\|\delta H_s\| = \|W^*VV^*W + V^*WW^*V\| = \|V^*W\|. \tag{4.15}$$

As the next step we establish that V and W are partial isometries such that

$$\mathsf{ran}(V) = \mathsf{ran}(\mathbf{H}^{1/2}P), \tag{4.16}$$

$$\mathsf{ran}(W)^{\perp} = \mathsf{ran}(\mathbf{H}^{-1/2}P). \tag{4.17}$$

The proof will follow from Lemma 4.2. It runs along the same lines in both cases, so we will present the proof only for W. Take some $u, v \in \mathcal{H}$; then

$$\begin{aligned}(Wu, Wv) &= (\mathbf{H}^{1/2}P_{\perp}\mathbf{H}'^{-1/2}u, \mathbf{H}^{1/2}P_{\perp}\mathbf{H}'^{-1/2}v)\\ &= h(P_{\perp}\mathbf{H}'^{-1/2}u, P_{\perp}\mathbf{H}'^{-1/2}v) = h'(P_{\perp}\mathbf{H}'^{-1/2}u, P_{\perp}\mathbf{H}'^{-1/2}v) = (P_{\perp}u, v),\end{aligned}$$

and so $W^*W = P_{\perp}$. This proves that W is a partial isometry.

Relation (4.16) is obvious, since

$$\mathsf{ran}(\mathbf{H}^{1/2}P\mathbf{H}^{'-1/2}) = \mathsf{ran}(\mathbf{H}^{1/2}P)$$

is guaranteed by the assumption $\mathsf{ran}(P) \subset \mathcal{Q}(h)$ and the injectivity of $\mathbf{H}^{'-1/2}$.

The proof of (4.17) requires a bit more work. One computes

$$W^*\mathbf{H}^{-1/2}P = \mathbf{H}^{'-1/2}P_{\perp}\mathbf{H}^{1/2}\mathbf{H}^{-1/2}P = 0,$$

which implies

$$\mathsf{ran}(\mathbf{H}^{-1/2}P) \subset \mathsf{ker}(W^*) = \mathsf{ran}(W)^{\perp}.$$

On the other hand,

$$W^* = P_{\perp}A, \tag{4.18}$$

where $A = \overline{\mathbf{H}'^{-1/2}\mathbf{H}^{1/2}} : \mathcal{H} \to \mathcal{H}$ is a homeomorphism (of linear topological vector spaces), and so

$$\mathsf{dim\ ker}(W^*) = \mathsf{dim\ ker}(P_{\perp}) = \mathsf{dim\ ran}(P) = \mathsf{dim\ ran}(\mathbf{H}^{-1/2}P),$$

and (4.17) is established. The assumption $\mathsf{sin}\Theta < 1$ and Lemma 3.5 guarantee

$$\sin\Theta = \|V^*W\|.$$

Finally, using (4.9), we establish

$$(1 - \sin\Theta)h'[u] \leq h[u] \leq (1 + \sin\Theta)h'[v],$$

which is the statement (4.7).

It is a well-known fact that, given some $0 < \lambda, \mu$ and $0 < \eta < 1$, the implication

$$\frac{|\lambda - \mu|}{\mu} \leq \eta \Rightarrow \frac{|\lambda - \mu|}{\lambda} \leq \frac{\eta}{1 - \eta} \tag{4.19}$$

holds. Since h and h' are positive definite forms, the relation (4.8) is proved. ∎

Example 4.6. Let $-\partial_{xx}$ be considered as the self-adjoint operator with

$$\mathcal{D}(-\partial_{xx}) = \{u \in H^2[0,1] :\ u(0) = u(1) = 0\}.$$

The partial integration establishes that $-\partial_{xx}$ is defined by the positive definite form

$$h(u,v) = \int_0^1 \partial_x u \; \partial_x v \; dx, \qquad u, v \in \mathcal{Q}(-\partial_{xx}) = H_0^1[0,1]. \tag{4.20}$$

The operator $\partial_x u$, $u \in H_0^1[0,1]$ is closed but not self-adjoint; therefore (4.20) is an alternative operator representation (factorization) to the "square root" representation (4.21) of the form h (the operator $-\partial_{xx}$).

Take any positive definite form h; then

$$h(u,v) = (\mathbf{H}^{1/2}u, \mathbf{H}^{1/2}v) \tag{4.21}$$

is only one of the possible operator representations of the form h. All of the preceding results are independent of the choice of the operator representation $h(u,v) = (\mathbf{R}u, \mathbf{R}v)$, since

$$\sin\Theta = \max_{u,v\in\mathcal{Q}} \frac{|\delta h(u,v)|}{\sqrt{h'[u]h'[v]}} \tag{4.22}$$

and h' depends only on h and $\mathsf{ran}(P)$.

Furthermore, all of the representations of the form h are in a sense equivalent. Let $\mathbf{R} : \mathcal{H} \to \mathcal{H}'$ be a closed operator such that

$$h(x,y) = (\mathbf{R}x, \mathbf{R}y) = \left(\mathbf{H}^{1/2}x, \mathbf{H}^{1/2}y\right) \tag{4.23}$$

and $\mathcal{Q} = \mathcal{D}(\mathbf{R}) = \mathcal{D}(\mathbf{H}^{1/2})$; then by [13, Chapter VI.7]

$$\mathbf{R} = U\mathbf{H}^{1/2}, \quad \mathbf{R}^* = \mathbf{H}^{1/2}U^*, \tag{4.24}$$

where U is the isometry from $\mathcal{H}'$ onto $\mathsf{ran}(\mathbf{R})$. Independence of the estimate (4.7) from the representation (4.23) could have also been proved by the unitary invariance of the canonical angle and (4.24). Formula (4.22) is an important corollary of Theorem 4.5. In the next theorem we prove that also

$$\frac{\sin\Theta}{1-\sin\Theta} = \max_{u,v\in\mathcal{Q}} \frac{|\delta h(u,v)|}{\sqrt{h[u]h[v]}} \tag{4.25}$$

holds. Equations (4.22) and (4.25) demonstrate that the constants $\mathsf{sin}\Theta$ and $\frac{\mathsf{sin}\Theta}{1-\mathsf{sin}\Theta}$ in (4.7) and (4.8) cannot be improved upon.

The following lemma is taken out of the joint paper [9]; cf. [5]. We present it here without a proof.

LEMMA 4.7. *Let the form h be positive definite, and let the forms h' and δh be as in* (4.4)*; then*

$$\max_{u,v\in\mathcal{Q}} \frac{|\delta h(u,v)|}{\sqrt{h[u]h[v]}} = \frac{\sin\Theta}{1-\sin\Theta} \tag{4.26}$$

holds. Here $\mathsf{sin}\Theta = \mathsf{sin}\Theta(\mathbf{H}^{1/2}X, \mathbf{H}^{-1/2}X)$*, where* $\mathsf{ran}(X) \subset \mathcal{Q}$ *was the subspace used to define h' and δh.*

4.1. The nonnegative definite case. In the nonnegative case we have to provide an alternative definition for a subspace that will play the role of $\mathsf{ran}(\mathbf{H}^{-1/2}X)$. We have shown $W = \mathbf{H}^{1/2}P_\perp\mathbf{H}'^{-1/2}$ to be a partial isometry such that

$$\mathcal{W} = \mathsf{ran}(\mathbf{H}^{1/2}P_\perp)^\perp = \mathsf{ran}(W)^\perp = \mathsf{ran}(\mathbf{H}^{-1/2}X).$$

The left part of the equality is also well defined in the case in which $\mathbf{H}^{1/2}$ is not invertible, so we set

$$\mathcal{W} = \mathsf{ran}(\mathbf{H}^{1/2}P_\perp)^\perp.$$

The construction (4.4) was performed with the assumption that h is nonnegative definite and $\mathsf{ran}(X) \subset \mathcal{Q}$. Lemma 4.2 says $\sigma_{ess}(\mathbf{H}) = \sigma_{ess}(\mathbf{H}')$, and so $\mathbf{H}^{'\dagger 1/2}$ is a bounded operator and

$$V = \mathbf{H}^{1/2}P\mathbf{H}^{'\dagger 1/2}, \tag{4.27}$$

$$W = \mathbf{H}^{1/2}P_\perp\mathbf{H}^{'\dagger 1/2} \tag{4.28}$$

are everywhere defined. Corollary 4.3 enables us to conclude that $\mathsf{ran}(V) = \mathsf{ran}(\mathbf{H}^{1/2}P)$ and $\mathsf{ran}(W) = \mathsf{ran}(\mathbf{H}^{1/2}P_\perp)$, so we set

$$\mathcal{V} = \mathsf{ran}(V), \qquad \mathcal{W} = \mathsf{ran}(W)^\perp. \tag{4.29}$$

Lemma 4.2 states that, given a positive definite $\mathbf{H}$, the constructed operator $\mathbf{H}'$ must always be positive definite. In the general nonnegative situation we have only the result of Corollary 4.3. It establishes that $\mathbf{H}'$ is a nonnegative definite operator and that $\mathsf{ker}(\mathbf{H}') \subset \mathsf{ker}(\mathbf{H})$. This does not give sufficient information on the structure of $\mathbf{H}'$. Formulae like (4.7)–(4.8) are meaningful in the nonnegative definite case, too. They, however, invariably imply $\mathsf{ker}(\mathbf{H}) = \mathsf{ker}(\mathbf{H}')$. We, therefore, proceed in two steps. First, we establish a general (theoretical) condition on the subspace $\mathcal{X} = \mathsf{ran}(P)$, which guarantees that $\mathsf{ker}(\mathbf{H}) = \mathsf{ker}(\mathbf{H}')$. Second, we give a practical computational formula.

The subspaces $\mathcal{W}$ and $\mathcal{V}$ need not have the same dimension, so we will have to use the principal angle to compare them; cf. Lemma 3.1. In what follows we show that

$$\sin\Theta_p(\mathcal{V},\mathcal{W})$$

takes the role of $\mathsf{sin}\Theta(\mathbf{H}^{1/2}X, \mathbf{H}^{-1/2}X)$ in the nonnegative version of Theorem 4.5. In the case when $\mathbf{H}^{1/2}$ is invertible (4.17) implies $\mathcal{V} = \mathsf{ran}(\mathbf{H}^{1/2}X)$ and $\mathcal{W} = \mathsf{ran}(\mathbf{H}^{-1/2}X)$. The subspaces $\mathbf{H}^{-1/2}X$ and $\mathbf{H}^{1/2}X$ have the same dimension, so Corollary 3.1 yields

$$\sin\Theta_p(\mathcal{V},\mathcal{W}) = \sin\Theta(\mathbf{H}^{1/2}X, \mathbf{H}^{-1/2}X).$$

We establish the properties of V and W and give a characterization of the subspace $\mathcal{W}$ in the following lemma.

LEMMA 4.8. *Let $\mathcal{X} = \mathsf{ran}(P)$, $V = \mathbf{H}^{1/2}P\mathbf{H}^{'\dagger 1/2}$, and $W = \mathbf{H}^{1/2}P_\perp\mathbf{H}^{'\dagger 1/2}$; then*

$$V^*V = P_{\mathsf{ran}(\mathbf{H}'P)}, \tag{4.30}$$

$$W^*W = P_{\mathsf{ran}(\mathbf{H}'P_\perp)}, \tag{4.31}$$

$$WV^* = VW^* = 0, \tag{4.32}$$

$$\mathcal{W} = \mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X}, \tag{4.33}$$

where $\mathcal{W}$ *is from* (4.29) *and*

$$\mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X} = \{x : \mathbf{H}^{1/2}x \in \mathcal{X}\}$$

denotes the inverse image of the subspace $\mathcal{X}$ *under the mapping* $\mathbf{H}^{1/2}$.

Proof. The relations (4.30)–(4.32) follow analogously as in the proof of Theorem 4.5. It remains only to prove (4.33).

We first show that $\mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X} \subset \mathcal{W} = \mathsf{ran}(W)^{\perp}$. Take any $u \in \mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X}$; then

$$\mathbf{H}^{1/2}u = z \in \mathcal{X}.$$

This implies

$$0 = (z, P_{\perp}\mathbf{H}'^{\dagger 1/2}v) = (u, \mathbf{H}^{1/2}P_{\perp}\mathbf{H}'^{\dagger 1/2}v), \qquad v \in \mathcal{H},$$

which proves $u \in \mathsf{ran}(W)^{\perp} = \mathcal{W}$.

The other inclusion follows in two steps. Take $u \in \mathcal{W}$; then

$$(u, \mathbf{H}^{1/2}P_{\perp}\mathbf{H}'^{\dagger 1/2}v) = 0, \qquad v \in \mathcal{H}.$$

On the other hand, the subspace

$$\mathsf{ran}(P_{\perp}\mathbf{H}'^{\dagger 1/2})^{\perp} = \mathsf{ran}(P_{\perp}P_{\mathsf{ran}(\mathbf{H}')})^{\perp} \subset \mathcal{Q}(\mathbf{H})$$

is finite dimensional, so we conclude $u \in \mathcal{Q}(\mathbf{H})$. Corollary 4.3 implies

$$0 = (\mathbf{H}^{1/2}u, P_{\perp}P_{\mathsf{ran}(\mathbf{H}')}v) = (\mathbf{H}^{1/2}u, P_{\mathsf{ran}(\mathbf{H}')}P_{\perp}v) = (\mathbf{H}^{1/2}u, P_{\perp}v), \qquad v \in \mathcal{H},$$

which proves $\mathbf{H}^{1/2}u \in \mathcal{X}$. With this conclusion we have established (4.33). ∎

As a direct consequence of Corollary 3.1 and (4.33) we obtain the following result.

COROLLARY 4.9. *Let* $\mathcal{X} = \mathsf{ran}(P)$, $V = \mathbf{H}^{1/2}P\mathbf{H}'^{\dagger 1/2}$, *and* $W = \mathbf{H}^{1/2}P_{\perp}\mathbf{H}'^{\dagger 1/2}$; *then*

$$\|P_{\mathcal{V}}P_{\mathcal{W}^{\perp}}\| \le \|P_{\mathcal{V}^{\perp}}P_{\mathcal{W}}\|,$$

and so

$$\sin\Theta_p(\mathbf{H}^{1/2}\mathcal{X}, \mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X}) = \|V^*W\|. \tag{4.34}$$

It would be pleasing to use $\mathbf{H}^{1/2\dagger}$ in the place of $\mathsf{inv}(\mathbf{H}^{1/2})$. This is possible only under additional restrictions on the subspace $\mathsf{ran}(P)$. To get a better feeling for the meaning of $\mathsf{sin}\Theta_p(\mathbf{H}^{1/2}\mathcal{X}, \mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X})$ consider the following example.

Example 4.10. Take

$$H = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \qquad \mathcal{X} = \begin{bmatrix} 1 \\ 0 \end{bmatrix};$$

then

$$H' = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

is, unlike H, a positive definite matrix. Now,

$$H^{1/2} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}, \qquad H^{1/2\dagger} = \begin{bmatrix} \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \end{bmatrix},$$

and we compute

$$\mathsf{ran}(V) = \mathrm{span}\{[1 \quad 1]^*\}, \qquad \mathsf{ran}(W)^\perp = \mathrm{span}\{[-1 \quad 1]^*\},$$

which proves that in this case $\mathsf{sin}\Theta_p(\mathsf{ran}(V), \mathsf{ran}(W)^\perp) = 1$ and

$$\mathsf{ran}(W)^\perp = \mathsf{ker}(H) \neq \mathsf{ran}(H^{1/2\dagger}P).$$

Instead of advocating the use of the general formula (4.33) we will establish a "compatibility condition" under which we may use the generalized inverse of $\mathbf{H}^{1/2}$ to check the statement of the theorems.

The next result is a nonnegative analogue of Theorem 4.5. It will enable us to, in effect, "deflate away" the kernel of the nonnegative form h and reduce the problem to the positive definite case.

THEOREM 4.11. *Let the subspace $\mathcal{X} = \mathsf{ran}(P) \subset \mathcal{Q}$ be given, and let h be a nonnegative form. Assume $\mathsf{sin}\Theta_p(\mathbf{H}^{1/2}\mathcal{X}, \mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X}) = \mathsf{sin}\Theta_p < 1$. Then*

$$(1 - \sin\Theta_p)h'[u] \leq h[u] \leq (1 + \sin\Theta_p)h'[u], \qquad u \in \mathcal{Q}(h), \tag{4.35}$$

$$\left(1 - \frac{\sin\Theta_p}{1 - \sin\Theta_p}\right) h[u] \leq h'[u] \leq \left(1 + \frac{\sin\Theta_p}{1 - \sin\Theta_p}\right) h[u], \qquad u \in \mathcal{Q}(h). \tag{4.36}$$

Proof. The proof is similar to the proof of Theorem 4.5. Let h' and δh be as in (4.4). Set δH_s to be the operator defined by the form

$$\delta h_s(x, y) = \delta h(\mathbf{H}'^{\dagger 1/2}x, \mathbf{H}'^{\dagger 1/2}y), \qquad x, y \in \mathcal{H}.$$

The form δh_s is closed and everywhere defined, and thus δH_s is a bounded operator. We obviously have $\mathsf{ker}(\mathbf{H}'^{\dagger 1/2}) = \mathsf{ker}(\mathbf{H}') \subset \mathsf{ker}(\delta H_s)$, and so $P_{\mathsf{ran}(\mathbf{H}')}$ commutes with the operator δH_s. With the use of Corollary 4.3 one computes, analogously as in Theorem 4.5,

$$\begin{aligned}\delta h(\mathbf{H}'^{\dagger 1/2}x, \mathbf{H}'^{\dagger 1/2}y) &= h(P_\perp \mathbf{H}'^{\dagger 1/2}x, P\mathbf{H}'^{\dagger 1/2}y) + h(P\mathbf{H}'^{\dagger 1/2}x, P_\perp \mathbf{H}'^{\dagger 1/2}y) \\ &= (Wx, Vy) + (Vx, Wy),\end{aligned}$$

and so

$$\delta H_s = V^*W + W^*V. \tag{4.37}$$

Since $\mathbf{H}'^{1/2}\mathbf{H}'^{\dagger 1/2} = P_{\mathsf{ran}(\mathbf{H}')}$ we obtain

$$\max_{u,v \in \mathsf{ran}(\mathbf{H}') \cap \mathcal{Q}} \frac{|\delta h(u, v)|}{\sqrt{h'[u]h'[v]}} = \|\delta H_s\| = \|V^*W\|. \tag{4.38}$$

Corollary 4.9 implies that the assumption $\mathsf{sin}\Theta_p < 1$, in fact, reads

$$\sin\Theta_p = \|V^*W\| < 1.$$

With this in hand, we have established

$$(1 - \sin\Theta_p)h'[u] \leq h[u] \leq (1 + \sin\Theta_p)h'[u], \qquad u \in \mathcal{Q}(h),$$

which implies $\mathsf{ker}(\mathbf{H}') = \mathsf{ker}(\mathbf{H})$. The relation (4.36) follows by the same argument used in Theorem 4.5. ∎

The main insight into the structure of the operator $\mathbf{H}'$, gained from Theorem 4.11, is summed up in the following corollary.

COROLLARY 4.12. *Take a nonnegative form h and a subspace $\mathcal{X} = \mathsf{ran}(P) \subset \mathcal{Q}$. If $\mathsf{sin}\Theta_p(\mathbf{H}^{1/2}\mathcal{X}, \mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X}) < 1$, then $\mathsf{ran}(\mathbf{H}') = \mathsf{ran}(\mathbf{H})$.*

Corollary 4.12 gives precise meaning to the statement "deflate away." Set $\mathcal{R} = \mathsf{ran}(\mathbf{H}) = \mathsf{ran}(\mathbf{H}')$ and $\mathcal{N} = \mathsf{ker}(\mathbf{H}) = \mathsf{ker}(\mathbf{H}')$. The projections $P_{\mathcal{N}}$ and P commute, and so

$$P_{\mathcal{N}\cap\mathsf{ran}(P)} = P_{\mathcal{N}}P, \qquad \widetilde{P} = P - P_{\mathcal{N}\cap\mathsf{ran}(P)}$$

are both orthogonal projections. A direct calculation shows

$$\widetilde{\mathcal{X}} := \mathsf{ran}(\widetilde{P}) = \mathsf{ran}(P) \ominus (\mathcal{N} \cap \mathsf{ran}(P)) = \mathsf{ran}(\mathbf{H}') \cap \mathsf{ran}(P) = \mathsf{ran}(\mathbf{H}'P).$$

The form

$$\widetilde{h}(u, v) = h(P_{\mathcal{R}}u, P_{\mathcal{R}}v)$$

is positive definite in $\mathcal{R}$ and $\mathsf{ran}(\widetilde{P}) \subset \mathcal{Q}(\widetilde{h}) \cap \mathcal{R}$. Now, apply the construction (4.2)–(4.4) to the form $\widetilde{h}$ and the projection $\widetilde{P}$. By $\widetilde{\mathbf{H}} : \mathcal{R} \to \mathcal{R}$ denote the operator defined by the form $\widetilde{h}$ in $\mathcal{R}$; then $\mathsf{ran}(\widetilde{P}) \subset \mathcal{R}$ and

$$\widetilde{h}'(u, v) = h'(P_{\mathcal{R}}u, P_{\mathcal{R}}v).$$

We conclude that

$$\sin\Theta(\widetilde{\mathbf{H}}^{1/2}\widetilde{\mathcal{X}}, \widetilde{\mathbf{H}}^{-1/2}\widetilde{\mathcal{X}}) = \sin\Theta_p(\mathbf{H}^{1/2}\mathcal{X}, \mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X}) < 1,$$

and $\widetilde{h}$ and $\widetilde{P}$ satisfy the assumptions of Theorem 4.5. If we were to a priori assume $\mathsf{ran}(\mathbf{H}') = \mathsf{ran}(\mathbf{H})$, then this argument would give an alternative proof of Theorem 4.11. "Deflate away" means that we assume that we were given $\widetilde{h}$ and $\widetilde{P}$ as input.

Remark 4.13. Another consequence of Corollary 4.12 is that we can invoke Lemma 4.7 to conclude that the constant $\frac{\mathsf{sin}\Theta_p}{1-\mathsf{sin}\Theta_p}$ (in (4.36)) cannot be sharpened. Furthermore, Example 4.10 shows that the assumption

$$\sin\Theta_p(\mathbf{H}^{1/2}\mathcal{X}, \mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X}) < 1$$

is a necessary requirement to establish the inequalities (4.35) and (4.36) as well as to guarantee that $\mathsf{ran}(\mathbf{H}) = \mathsf{ran}(\mathbf{H}')$ (equivalently, $\mathsf{ker}(\mathbf{H}) = \mathsf{ker}(\mathbf{H}')$).

Remark 4.14. Let $\mathcal{X} = \mathsf{ran}(P)$, and let the forms h and h' be as in Theorem 4.11. Set

$$h_\varepsilon(u, v) := h(u, v) + \varepsilon^2(u, v), \qquad u, v \in \mathcal{Q}(h);$$

then $h'_\varepsilon(u, v) = h'(u, v) + \varepsilon^2(u, v)$. Now,

$$\lim_{\varepsilon\to 0} \sup_{u\in\mathcal{Q}(h)} \frac{|h[u] - h'[u]|}{h'[u] + \varepsilon^2\|u\|^2} = \lim_{\varepsilon\to 0} \sup_{u\in\mathcal{Q}(h)} \frac{|h_\varepsilon[u] - h'_\varepsilon[u]|}{h'_\varepsilon[u]} \le 1,$$

and we define

$$\mathtt{R} := \lim_{\varepsilon\to 0} \sup_{u\in\mathcal{Q}(h)} \frac{|h[u] - h'[u]|}{h'[u] + \varepsilon^2\|u\|^2}$$

and obtain an inequality like (4.35). This in turn implies that $\mathtt{R} < 1$ is equivalent to $\mathsf{ker}(\mathbf{H}) = \mathsf{ker}(\mathbf{H}')$. An optimality argument yields

$$\mathsf{sin}\Theta_p(\mathbf{H}^{1/2}\mathcal{X}, \mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X}) = \lim_{\varepsilon\to 0} \sup_{u\in\mathcal{Q}(h)} \frac{|h[u] - h'[u]|}{h'[u] + \varepsilon^2\|u\|^2} = \mathtt{R}$$

as an alternative analytical formula for $\mathsf{sin}\Theta_p$. We have opted for a geometrical argument, since it sets the scene for a complete analysis of the singular values of δH_s; cf. [10]. Theorem 4.11 illustrates (when compared with Theorem 4.5) the natural limits of our geometric/operator-block-matrix approach—nonnegative definite operators have a different structure than positive definite operators and should not be seen, within this theory, as ε^2 close to a positive definite ones; cf. Remark 4.16. Furthermore, numerical experience from the matrix case, as well as formula (1.7) in this paper, show that $\mathsf{sin}\Theta_p(\mathbf{H}^{1/2}\mathcal{X}, \mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X})$ is the quantity which is more accessible to computation than

$$\lim_{\varepsilon\to 0} \sup_{u\in\mathcal{Q}(h)} \frac{|h[u] - h'[u]|}{h'[u] + \varepsilon^2\|u\|^2},$$

which requires the solution of a double optimization problem.

4.1.1. Important special case. The assumption that P and $P_{\mathsf{ker}(\mathbf{H})}$ commute, together with Corollary 4.3, yields $\mathsf{ker}(\mathbf{H}) = \mathsf{ker}(\mathbf{H}')$ and $\mathsf{ran}(\mathbf{H}) = \mathsf{ran}(\mathbf{H}')$. This implies

$$\mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X} = \mathbf{H}^{1/2\dagger}\mathcal{X}. \tag{4.39}$$

The projections P and $P_{\mathsf{ker}(\mathbf{H})}$ certainly commute when $\mathsf{ker}(\mathbf{H}) \perp \mathsf{ran}(P)$ or when[3] $\mathsf{ker}(\mathbf{H}) \subset \mathsf{ran}(P)$. This discussion is summed up in the following corollary.

COROLLARY 4.15. *Assume that $P = XX^*$ and $P_{\mathsf{ker}(\mathbf{H})}$ commute, and let*

$$\mathsf{sin}\Theta_p(\mathbf{H}^{1/2}X, \mathbf{H}^{1/2\dagger}X) < 1;$$

then

$$(1 - \sin\Theta_p)h'[u] \le h[u] \le (1 + \sin\Theta_p)h'[u], \qquad u \in \mathcal{Q}(h), \tag{4.40}$$

$$\left(1 - \frac{\sin\Theta_p}{1 - \sin\Theta_p}\right) h[u] \le h'[u] \le \left(1 + \frac{\sin\Theta_p}{1 - \sin\Theta_p}\right) h[u], \qquad u \in \mathcal{Q}(h). \tag{4.41}$$

Remark 4.16. To assess the restriction that P and $P_{\mathsf{ker}(\mathbf{H})}$ should commute, consider the definition of the relatively accurate approximation of the number $\lambda \in \mathbb{R}_+$. $\mu \in \mathbb{R}_+$ is a relatively accurate approximation of $\lambda \in \mathbb{R}_+$ if

1. $\lambda = \mu$ when $\lambda = 0$,
2. $\frac{|\lambda-\mu|}{\mu} < 1$ when $\lambda \neq 0$.

This implies that we can expect to compute a "relatively accurate" Ritz value approximation of the spectrum of the nonnegative definite operator $\mathbf{H}$ only in the case when we have computed a basis for $\mathsf{ker}(\mathbf{H})$; cf. [1].

[3]The other situation when P and $P_{\mathsf{ker}(\mathbf{H})}$ commute is when $\mathsf{ran}(P) \subset \mathsf{ker}(\mathbf{H})$; this situation is, however, trivial and we have tacitly left it out.

Remark 4.13 implies that we may assume that the condition of Corollary 4.15 is $\mathsf{ker}(\mathbf{H}) \perp \mathsf{ran}(P)$. To compute the basis of the set $\mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X}$ we need to repeatedly solve the equation

$$\mathbf{H}^{1/2}u = x_i, \qquad i = 1, \ldots, \mathsf{dim}(\mathcal{X}).$$

The vectors x_i are assumed to be a basis for $\mathcal{X}$. The restriction that $\mathsf{ker}(\mathbf{H}) \perp \mathsf{ran}(P)$ amounts to nothing more than imposing a compatibility condition on x_i (e.g., think of the Laplacian with Neumann boundary conditions).

4.2. A first approximation estimate. Theorem 2.3 and Lemma 4.2 yield the first eigenvalue estimates. The next theorem will give an eigenvalue estimate with the minimum restrictions on the subspace $\mathsf{ran}(X) \subset \mathcal{Q}$. Sharper bounds are possible when we impose additional assumptions on $\mathsf{ran}(X)$. Even this (first order) estimate will compare favorably with other higher order bounds that can be found in the literature; cf. [9].

THEOREM 4.17. *Let $0 \leq h$, and let the n-dimensional subspace $\mathsf{ran}(P) \subset \mathcal{Q}$, $P = XX^*$, be given. Define*

$$\Xi = (\mathbf{H}^{1/2}X)^*\mathbf{H}^{1/2}X, \qquad \Xi \in \mathbb{C}^{n\times n},$$

and assume $\mu_n < \lambda_e(\mathbf{H})$. Here, the Ritz values are numbered as in (4.1). *If $\mathsf{ran}(P)$ is such that $\mathsf{sin}\Theta_p < 1$, then there are n eigenvalues of the operator $\mathbf{H}$, counting the eigenvalues according to their multiplicities, such that*

$$|\lambda_{i_j} - \mu_j| \leq \mu_j \sin\Theta_p, \qquad j = 1, \ldots, n, \tag{4.42}$$

$$|\lambda_{i_j} - \mu_j| \leq \lambda_{i_j} \frac{\sin\Theta_p}{1 - \sin\Theta_p}, \qquad j = 1, \ldots, n, \tag{4.43}$$

where $i_{(\cdot)} : \mathbb{N} \to \mathbb{N}$ is a permutation.

Proof. Corollary 4.12 readily implies the conclusion (4.42) for the Ritz values $\mu_j = 0$, $j = 1, \ldots, \mathsf{dim}(\mathsf{ker}(\Xi))$. Therefore, we may safely assume that h is a positive definite form. Lemma 4.2 implies $\sigma_{ess}(\mathbf{H}) = \sigma_{ess}(\mathbf{H}')$, and thus the assumption $\mu_n < \lambda_e(\mathbf{H})$ guarantees that μ_n is a discrete eigenvalue of $\mathbf{H}'$. Theorem 4.11 established

$$(1 - \sin\Theta_p)h'[u] \leq h[u] \leq (1 + \sin\Theta_p)h'[u], \qquad u \in \mathcal{Q}(h),$$

$$\left(1 - \frac{\sin\Theta_p}{1 - \sin\Theta_p}\right) h[u] \leq h'[u] \leq \left(1 + \frac{\sin\Theta_p}{1 - \sin\Theta_p}\right) h[u], \qquad u \in \mathcal{Q}(h).$$

The conclusion follows directly from Theorem 2.3. ∎

For the numerical evidence concerning the performance of the estimate (4.43), see the numerical tests from [9].

5. Localizing the approximated eigenvalues. There are many ways to match the computed Ritz values to a part of the spectrum of the operator $\mathbf{H}$ of the same multiplicity. These approaches usually differ with regard to the amount of additional information allowed about the spectrum of the operator $\mathbf{H}$. Here, we present two possible answers to that problem.

Theorem 4.17 can be interpreted as a first localization result. It gives an estimate of the infimum of

$$\max_{j=1,\ldots,n} \frac{|\lambda_{i_j} - \mu_j|}{\mu_j}$$

over all of the permutations $i_{(\cdot)} : \mathbb{N} \to \mathbb{N}$. So, we would be correct in stating that the Ritz values are approximating the eigenvalues of $\mathbf{H}$ that are closest to $\sigma(\Xi)$.

Having only limited additional information, we got a limited answer. We know that there is a collection of eigenvalues of operator $\mathbf{H}$, having the joint multiplicity n, that is being approximated by the Ritz values from the subspace $\mathsf{ran}(X)$. The information we have on the location of those eigenvalues in the spectrum of $\mathbf{H}$ is only that they are the eigenvalues closest to computed Ritz values.

Only when we have additional information about the location of the part of the spectrum that *we do not want to approximate* can we guarantee that we are approximating the part of the spectrum we are interested in. A best known example of such estimates is the Temple–Kato inequality. Assume $\lambda_1 < \lambda_2$ and let $u \in \mathcal{D}(\mathbf{H})$ be a unit vector such that $(u, \mathbf{H}u) < \gamma \leq \lambda_2$; then

$$(u, \mathbf{H}u) \geq \lambda_1 \geq (u, \mathbf{H}u) - \frac{(\mathbf{H}u, \mathbf{H}u) - (u, \mathbf{H}u)^2}{\gamma - (u, \mathbf{H}u)}. \tag{5.1}$$

For a proof see [19]. The estimate (5.1) is valid for a general self-adjoint operator $\mathbf{H}$. In the following we shall formulate another assumption with the same effect, namely to separate the "unwanted" component of the spectrum from the Ritz values. Our result, however, does not need the regularity constraint $u \in \mathcal{D}(\mathbf{H})$. Moreover, we will obtain sharp bounds for the matching cluster of eigenvalues. In the last section of this chapter we will demonstrate that on some examples our bound considerably outperforms the estimate (5.1).

We now give a theorem that determines those eigenvalues of the operator $\mathbf{H}$, given by a symmetric form h, which are approximated by the Ritz values associated with the test subspace $\mathsf{ran}(X) \subset \mathcal{Q}$. Before we proceed with the formulation of the theorem we state a well-known fact that, given $0 < \lambda, \mu$ and $\sin\Theta_p < 1$, the relation

$$\frac{|\lambda - \mu|}{\mu} \leq \sin\Theta_p < 1$$

implies the relation

$$\frac{|\lambda - \mu|}{\lambda} \leq \frac{\sin\Theta_p}{1 - \sin\Theta_p} \leq 2\sin\Theta_p. \tag{5.2}$$

THEOREM 5.1. *Set* $\gamma_r = \min\{(\lambda_p - \mu_k)(\lambda_p + \mu_k)^{-1} | k = 1, \ldots, n;\ p = n+1, \ldots, \infty\}$ *and* $\eta_{\Theta_p} = \mathsf{sin}\Theta_p(1 - \mathsf{sin}\Theta_p)^{-1}$. *Take a nonnegative form* h *and the subspace* $\mathsf{ran}(X) \subset \mathcal{Q}$. *Assume* $r = \dim(\mathsf{ker}(\mathbf{H})) \leq n$, *set* $P = XX^*$, *and let* h' *be as in* (4.3). *By* $\mu_1 \leq \cdots \leq \mu_n$, *denote the eigenvalues of the matrix* $\Xi = (\mathbf{H}^{1/2}X)^*\mathbf{H}^{1/2}X \in \mathbb{C}^{n\times n}$. *If* $\gamma_r \geq 0$ *and* $\eta_{\Theta_p} < \min\{\gamma_r, 1\}$, *then*

$$|\lambda_i - \mu_i| \leq \mu_i \sin\Theta_p, \quad i = 1, \ldots, n. \tag{5.3}$$

Proof. The assumption $\eta_{\Theta_p} < \min\{\gamma_r, 1\}$ and Theorem 4.11 imply $\mathsf{ker}(\mathbf{H}) \subset \mathsf{ran}(X)$. Also, by Theorem 4.11 we have $\mathsf{ker}(\mathbf{H}) = \mathsf{ker}(\mathbf{H}')$, so we are allowed to "deflate away" the kernel of $\mathbf{H}$. Therefore, set $P_1 = P_{\mathsf{ran}(\mathbf{H}'P)}$ and proceed as if h were positive definite and $P = P_1$.

The rest of the proof is completely analogous to the proof of [9, Theorem 5.1]. The only difference is that in the place of $\eta = \mathsf{sin}\Theta_p/(1 - \mathsf{sin}\Theta_p)$ from [9, Theorem 5.1] one uses a sharper quantity $\mathsf{sin}\Theta_p$. ∎

If we are provided with the information that

$$\eta_{\Theta_p} = \frac{\sin\Theta_p}{1-\sin\Theta_p} < \gamma_c := \min\left\{ \min_{\substack{k=1,\dots,n\\ p=1,\dots,q-1}} \frac{\mu_k-\lambda_p}{\lambda_p+\mu_k}, \min_{\substack{k=1,\dots,n\\ p=q+n,\dots,\infty}} \frac{\lambda_p-\mu_k}{\lambda_p+\mu_k}, 1 \right\}, \tag{5.4}$$

then $\mu_1 \le \cdots \le \mu_n$ approximate the "inner" eigenvalues

$$\lambda_q \le \lambda_{q+2} \cdots \le \lambda_{q+n-1} \ .$$

This statement is made precise in the following theorem.

THEOREM 5.2. *Take a nonnegative definite form h and a subspace $\mathsf{ran}(X) \subset \mathcal{Q}$. By $\mu_1 \le \cdots \le \mu_n$ denote the eigenvalues of the matrix $\Xi = (\mathbf{H}^{1/2}X)^*\mathbf{H}^{1/2}X \in \mathbb{C}^{n\times n}$. If $\eta_{\Theta_p} < \gamma_c$, where γ_c is as in* (5.4)*, then $\mathsf{ran}(P) \subset \mathsf{ran}(\mathbf{H}')$ and*

$$|\lambda_{i+q-1} - \mu_i| \le \mu_i \sin\Theta_p, \quad i = 1,\dots,n.$$

Proof. The assumption (5.4) and Theorem 4.11 and Corollary 4.12 imply

$$\mathsf{ran}(\mathbf{H}') = \mathsf{ran}(\mathbf{H}) \qquad \text{and} \qquad \mathsf{ran}(P) \subset \mathsf{ran}(\mathbf{H}).$$

The rest of the proof follows analogously as in the proof of Theorem 5.1. ∎

Remark 5.3. Theorems 5.1 and 5.2 imply that the spectrum of the operator $\mathbf{H}$ can stably (sensibly) be divided in two disjoint parts: the part that is being approximated by the $\sigma(\Xi)$ and the rest of the spectrum. To understand this statement assume that the conditions of Theorem 5.1 hold. In this case both of the "block diagonal" forms

$$h(u,v) = h(E(\lambda_n)u, E(\lambda_n)v) + h(E(\lambda_n)_\perp u, E(\lambda_n)_\perp v) \simeq \begin{bmatrix} \Lambda & \\ & \Lambda_c \end{bmatrix},$$
$$h'(u,v) = h(Pu,Pv) + h(P_\perp u, P_\perp v) \simeq \begin{bmatrix} \Xi & \\ & \Xi_c \end{bmatrix}$$

have "diagonal blocks" with disjoint spectra. We have assumed $\Lambda = \mathrm{diag}(\lambda_1,\dots,\lambda_n)$ and $\Xi = \mathrm{diag}(\mu_1,\dots,\mu_n)$ and that Ξ_c and Λ_c were unbounded operators defined by the forms h' and h in the spaces $\mathsf{ran}(P_\perp)$ and $\mathsf{ran}(E(\lambda_n)_\perp)$. In fact, we will colloquially call h' the *block diagonal part of the operator* $\mathbf{H}$ *with respect to the subspace* $\mathsf{ran}(P)$. We will use the notation h_P to denote h' in situations when it is not clear with respect to which test space $\mathsf{ran}(P)$ this construction was performed.

6. Eigenvector approximation estimates. For the computed Ritz values

$$0, 0, \dots, 0, \mu_{r+1}, \mu_{r+2}, \dots, \mu_n$$

Theorem 4.17 guarantees the existence of the eigenvalues

$$\lambda_{i_1} \le \lambda_{i_2} \le \cdots \le \lambda_{i_n},$$

which are being approximated by the Ritz values (provided $\mathsf{sin}\Theta_p < 1$) in the sense of

$$|\lambda_{i_j} - \mu_j| \le \mu_j \sin\Theta_p, \qquad j = 1,\dots,n.$$

Assume that $v_1,\dots,v_n$ are mutually orthogonal eigenvectors that belong to the eigenvalues $\lambda_{i_1} \le \lambda_{i_2} \le \cdots \le \lambda_{i_n}$. If the conditions of Theorems 5.1 and 5.2 are satisfied, Remark 5.3 assures us that

$$\mathrm{span}\{v_1,\dots,v_n\} = \mathsf{ran}(E(\{\lambda_{i_1},\lambda_{i_2},\dots,\lambda_{i_n}\})).$$

Here we have assumed that $\mathbf{H} = \int \lambda \, \mathrm{d}E(\lambda)$. To ease the presentation we generically use

$$\widehat{E} = E(\{\lambda_{i_1}, \lambda_{i_2}, \ldots, \lambda_{i_n}\})$$

to denote the projection on the subspace that is selected by a result like Theorem 5.1.

The central role in the analysis of the eigenvector approximations will be played by the following lemma.

LEMMA 6.1. *Let h be a nonnegative form, and let $\mathbf{H}^\dagger$ be bounded. Take $\mathsf{ran}(P) \subset \mathcal{Q}$ such that $\mathsf{sin}\Theta_p < 1$, and define*

$$s(x, y) = \delta h(\mathbf{H}^{\dagger 1/2} x, \mathbf{H}'^{\dagger 1/2} y), \qquad x, y \in \mathcal{H}.$$

The form s defines a bounded operator S and

$$S = \mathbf{H}^{1/2} \mathbf{H}'^{\dagger 1/2} - \overline{\mathbf{H}^{\dagger 1/2} \mathbf{H}'^{1/2}}, \tag{6.1}$$

$$|(x, Sy)| = |s(x, y)| \leq \frac{\sin \Theta_p}{\sqrt{1 - \sin \Theta_p}} \|x\| \|y\|, \quad x, y \in \mathcal{H}. \tag{6.2}$$

Proof. The closed graph theorem implies that the operator

$$S = \mathbf{H}^{1/2} \mathbf{H}'^{\dagger 1/2} - \overline{\mathbf{H}^{\dagger 1/2} \mathbf{H}'^{1/2}}$$

is bounded. Also, $\mathsf{ker}(\mathbf{H}) = \mathsf{ker}(\mathbf{H}') = \mathsf{ker}(S)$ and $P_{\mathsf{ker}(S)}$ commutes with S. It is sufficient to prove the estimate for $x, y \in \mathsf{ran}(\mathbf{H})$. The inequality (4.38) gives

$$|\delta h(\mathbf{H}^{\dagger 1/2} x, \mathbf{H}'^{\dagger 1/2} y)| \leq \sin \Theta_p \|y\| \; h'[\mathbf{H}^{\dagger 1/2} x]^{1/2}.$$

Analogously, (4.35) implies

$$\|\mathbf{H}'^{1/2} \mathbf{H}^{\dagger 1/2}\| \leq \frac{1}{\sqrt{1 - \sin \Theta_p}} \, . \tag{6.3}$$

Altogether, the estimate (6.2) follows. □

The operator S has the special structure. Assume $\mathbf{H}'u = \mu u$ and $\mathbf{H}v = \lambda v$; then

$$\begin{aligned} (v, Su) &= \lambda^{1/2} (v, u) \mu^{1/2} - \lambda^{-1/2} (v, u) \mu^{1/2} \\ &= \frac{\lambda - \mu}{\sqrt{\lambda \mu}} (v, u) \, . \end{aligned} \tag{6.4}$$

The equation (6.4) introduces the distance function

$$\frac{\lambda - \mu}{\sqrt{\lambda \mu}},$$

which measures the distance between the Ritz values and the spectrum of the operator $\mathbf{H}$. This distance function will play an important role in the estimates that follow. The next theorem extends the scope, as well as strengthens the eigenvector estimate from [9, 16], and is even new in the matrix case. It can be seen as the eigenvector companion result of Theorem 4.17.

THEOREM 6.2. *Let h be a nonnegative form, and let $\mathsf{ran}(P) \subset \mathcal{Q}$ be such that it satisfies the assumptions of Theorem* 4.17. *Let $u_1, \ldots, u_n$ be the mutually orthogonal*

eigenvectors belonging to the eigenvalues $\mu_1, \dots, \mu_n$ of $\mathbf{H}'P$; then there exist mutually orthogonal eigenvectors $v_1, \dots, v_n$ of $\mathbf{H}$, belonging to the eigenvalues $\lambda_{i_1}, \dots \lambda_{i_n}$, and

$$\|v_j - u_j\| \leq \frac{\sqrt{2}\ \sin\Theta_p}{\sqrt{1-\sin\Theta_p}} \max_{k\neq j} \frac{\sqrt{\mu_j \lambda_{i_k}}}{|\lambda_{i_k} - \mu_j|}. \tag{6.5}$$

The eigenvalues λ_{i_j}, $j = 1, \dots, r$, are numbered in the ascending order as given by Theorem 4.17.

Proof. Assume $\mu_1 = \cdots = \mu_r = 0$. Corollary 4.12 implies that $u_i \in \mathsf{ker}(\mathbf{H})$ for $i = 1, \dots, r$, and so we take

$$v_i = u_i, \quad i = 1, \dots, r.$$

For v_j, $j = r+1, \dots n$, take any orthonormal set of eigenvectors belonging to the eigenvalues λ_{i_j}, $j = r+1, \dots, n$. Since both u_i and v_i, for $i = r+1, \dots, n$, are perpendicular to $\mathsf{ker}(\mathbf{H})$ we may assume that $\mathbf{H}$ is positive definite and we are only given u_i, $i = r+1, \dots n$, as test vectors. Take s from Lemma 6.1 and use (6.1) to compute

$$\begin{aligned} s(v_k, u_j) &= \delta h(\mathbf{H}^{-1/2} v_k, \mathbf{H}^{'-1/2} u_j) = \left(v_k, \mathbf{H}^{1/2}\mathbf{H}^{'-1/2} u_j\right) - \left(\mathbf{H}^{'1/2}\mathbf{H}^{-1/2} v_k, u_j\right) \\ &= \left(\lambda_{i_k}^{1/2}\mu_j^{-1/2} - \lambda_{i_k}^{-1/2}\mu_j^{1/2}\right)(v_k, u_j) \end{aligned}$$

and

$$\begin{aligned} \sum_{k\neq j} |\,(v_k, u_j)\,|^2 &\leq \max_{k\neq j} \frac{\lambda_{i_k}\mu_j}{(\lambda_{i_k} - \mu_j)^2} \sum_{k\neq j} |s(v_k, u_j)|^2 \leq \max_{k\neq j} \frac{\lambda_{i_k}\mu_j}{(\lambda_{i_k} - \mu_j)^2}\ \|S^* u_j\|^2 \\ &\leq \max_{k\neq j} \frac{\lambda_{i_k}\mu_j}{(\lambda_{i_k} - \mu_j)^2} \frac{\sin^2\Theta_p}{1-\sin\Theta_p}. \end{aligned}$$

Scaling v_j, u_j so that $(v_j, u_j) \geq 0$, we obtain

$$\begin{aligned} \|v_j - u_j\| = \sqrt{2}\Big[1 - (v_j, u_j)\Big]^{1/2} &= \sqrt{2}\left[1 - \left[1 - \sum_{k\neq j} |\,(v_k, u_j)\,|^2\right]^{1/2}\right]^{1/2} \\ &\leq \sqrt{2}\left[1 - \left[1 - \max_{k\neq j} \frac{\lambda_{i_k}\mu_j}{(\lambda_{i_k} - \mu_j)^2} \frac{\sin^2\Theta_p}{1-\sin\Theta_p}\right]^{1/2}\right]^{1/2} \\ &\leq \frac{\sqrt{2}\ \sin\Theta_p}{\sqrt{1-\sin\Theta_p}} \max_{k\neq j} \frac{\sqrt{\mu_j\lambda_{i_k}}}{|\lambda_{i_k} - \mu_j|}. \end{aligned}$$

This proves the lemma in the case in which $\sigma_{ess}(\mathbf{H}) = \emptyset$. In the general case we use the formula

$$\frac{\sqrt{\lambda_e \mu_j}}{\lambda_e - \mu_j} \left|\left(E_{\mathbf{H}^{1/2}}([\sqrt{\lambda_e}, \infty\rangle) u_j, S u_j\right)\right| \geq \left|\left(E_{\mathbf{H}^{1/2}}([\sqrt{\lambda_e}, \infty\rangle) u_j, u_j\right)\right|$$

and analogous argument. ∎

6.1. Alternative approaches to the vector perturbation problem. The eigenvalue (invariant subspace) perturbation problem has been the focus of attention of many researchers. A comparison between various approaches is not easy, since at the heart of each subspace estimate lies an approximation problem which in general cannot be solved explicitly. As a consequence a compromise—which is dictated by the structure of a particular problem (operator) under study—has to be made prior to optimization so that we can establish a computational formula. Subsequently, each of these estimates is a tool which is keyed (e.g., sharp) on the class of problems for which it was designed.

The results of this article are designed so that the obtained formulae perform well on the class of nonnegative operators. Particular attention is paid to robustness of the estimates in the presence of singularities. On the other hand, section 4.1 shows that this theory, unlike the theory of [17], cannot be extended to indefinite operators without serious alterations.

To illustrate this issue let us consider the matrices

$$H_\eta = \begin{bmatrix} \frac{1}{101} & 0 & -\frac{1}{101} \\ 0 & \frac{1}{100} & 0 \\ -\frac{1}{101} & 0 & 1+\eta^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{101}} & 0 & 0 \\ 0 & \frac{1}{10} & 0 \\ -\frac{1}{\sqrt{101}} & 0 & \sqrt{\frac{100}{101}+\eta^2} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{101}} & 0 & -\frac{1}{\sqrt{101}} \\ 0 & \frac{1}{10} & 0 \\ 0 & 0 & \sqrt{\frac{100}{101}+\eta^2} \end{bmatrix},$$

and let $w = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^*$ be given. The numerically sharpest estimate[4] of $\sin\angle(w, v_1(H_\eta))$ from [3] is the $\tan 2\Theta$-Theorem (cf. [14, 17]), and it reads

$$\sin\angle(w, v_1(H_\eta)) \le \sin\left(\frac{1}{2}\arcsin\left(2\frac{\|H_\eta w - (w, H_\eta w)w\|}{\lambda_2(H) - (w, H_\eta w)}\right)\right). \tag{6.6}$$

On the other hand, Theorem 6.2 yields

$$\sin\angle(w, v_1(H_\eta)) \le \|v_1(H_\eta) - w\| \le \frac{\sqrt{2}\,\sin\Theta_\eta}{\sqrt{1-\sin\Theta_\eta}} \frac{\sqrt{(w, H_\eta w)\lambda_2(H_\eta)}}{|\lambda_2(H_\eta) - (w, H_\eta w)|}. \tag{6.7}$$

An important difference between (6.6) and (6.7) is in the way in which the singularity of H_η, as $\eta \to \infty$, is handled. Estimate (6.6) yields

$$\sin\angle(w, v_1(H_\eta)) \le \sin\left(\frac{1}{2}\arctan\left(\frac{2}{\frac{1}{100}-\frac{1}{101}}\frac{1}{101}\right)\right),$$

which remains constant as $\eta \to \infty$. On the other hand, (6.7) yields

$$\sin\angle(w, v_1(H_\eta)) \le \frac{\sqrt{\frac{1}{100}\frac{1}{101}}}{\frac{1}{100}-\frac{1}{101}} \frac{\sqrt{\frac{2}{101+101\eta^2}}}{\sqrt{1-\sqrt{\frac{1}{101+101\eta^2}}}},$$

which tends to 0 as $\eta \to \infty$. This is realistic since

$$\lim_{\eta\to\infty} H_\eta^{-1} = \begin{bmatrix} 101 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

[4] We assume $H_\eta v_i(H_\eta) = \lambda_i(H_\eta)v_i(H_\eta)$, for eigenvalues and eigenvectors of H_η.

and $\sin\angle(v_1(H_\eta), w) = \frac{1}{101\,\eta^2} - \frac{100}{10201\,\eta^4} + \frac{19997}{2060602\,\eta^6} + \mathrm{O}(\frac{1}{\eta^8})$ as $\eta \to \infty$. On this example, for $\eta \geq 50$, estimate (6.7) considerably outperforms (6.6). As η grows larger, (6.7) gets sharper in comparison with (6.6), which remains constant. For a comparison of the estimates from [3] and the estimates from this article on an example of Sturm–Liouville eigenvalue problems with coupled boundaries, see [9]. This matrix example also shows that there is still a possibility of improvement of the estimates, by perhaps a change of the eigenvector-perturbation metric $\sin\angle(v_1(H_\eta), w)$. This is a subject of ongoing research, and the results will be published elsewhere.

7. A simple model problem. We will now present an application of our theory to the singularly perturbed Sturm–Liouville eigenvalue problem (1.9). Estimates (5.1) and [2, Theorem 6.21](Kato–Temple eigenvector estimate) do not apply due to overly stringent regularity assumptions on the test vector; cf. (1.9)–(1.10).

Consider the family of positive definite forms

(7.1)
$$h_\eta(u,v) = h_b(u,v) + \eta^2 h_e(u,v) = \int_0^2 u'v'\ dx + \eta^2 \int_1^2 u'v'\ dx, \quad u, v \in H_0^1[0,2].$$

By $\mathbf{H}_\eta$ denote the positive definite operator which is defined by the form h_η from (7.1). We are interested in eigenvalues and eigenvectors of the operator $\mathbf{H}_\eta$ for large η. Here, $H_0^1[0,2]$ denotes the first order Sobolev space with zero trace on the boundary.

This is the eigenvalue problem for the vibration of a highly inhomogeneous string. We are considering only an academic example where we can efficiently compute all the information we need. For more realistic applications, see [8].

If we identify the functions from $H_0^1[0,\alpha]$, $\alpha > 0$, with their extension by zero to the whole of $[0,\beta]$ for $\beta \geq \alpha$, then we can write

$$H_0^1[0,\alpha] \subset H_0^1[0,\beta], \qquad 0 < \alpha < \beta. \tag{7.2}$$

Let $\chi_{[0,1]}$ be the characteristic function of the interval $[0,1]$, and let $\chi_{[0,1]^c} = 1 - \chi_{[0,1]}$. Keeping (7.2) in mind, we conclude that

$$\mathbf{H}_\eta = -\partial_x(1 + \eta^2 \chi_{[0,1]^c})\partial_x, \qquad \mathcal{D}(\mathbf{H}_\eta) = H^2[0,2] \cap H_0^1[0,2].$$

It is known that the forms h_η converge to the form

$$h_\infty(u,v) = \int_0^1 u'v'\ dx, \quad u, v \in H_0^1\,[0,1]\,,$$

in the norm resolvent sense.[5] Operators $\mathbf{H}_\eta$ and $\mathbf{H}_\infty$ have discrete spectra, and all the eigenvalues are nondegenerate; cf. [23]. Since we will be considering the whole family of operators $\mathbf{H}_\eta$, additional notation will be introduced to ease the understanding. By

$$\lambda_1^\eta < \cdots < \lambda_n^\eta < \cdots$$

we denote the increasingly ordered eigenvalues of the operator $\mathbf{H}_\eta$, and by

$$\lambda_1^\infty < \cdots < \lambda_n^\infty < \cdots$$

[5]More on the properties of this convergence can be found in [11, 22].

the eigenvalues of the operator $\mathbf{H}_\infty$.

The eigenpairs of the operator $\mathbf{H}_\infty$—which is defined in $L^2[0,1]$—are $(n^2\pi^2, \sqrt{2}\mathsf{sin}(n\pi x))$, $n \in \mathbb{N}$. The functions

$$u_n(x) = \begin{cases} \sqrt{2}\sin(n\pi x), & 0 \le x \le 1, \\ 0, & 1 \le x, \end{cases} \qquad n \in \mathbb{N}, \tag{7.3}$$

are in $H_0^1[0,1]$ and also in $H_0^1[0,2]$. Therefore, they can be used as test functions for an approximation of the eigenvalues of $\mathbf{H}_\eta$ (for large η). Furthermore, according to (1.7) we obtain

$$\sin^2\Theta_\eta(u_i) := \sin^2\Theta(\mathbf{H}_\eta^{-1/2}u_i, \mathbf{H}_\eta^{1/2}u_i) = \frac{(u_i, \mathbf{H}_\eta^{-1}u_i) - (u_i, \mathbf{H}_\infty^\dagger u_i)}{(u_i, \mathbf{H}_\eta^{-1}u_i)}.$$

Let us now concentrate on the approximation of the lowest eigenvalue. We compute the Ritz value

$$h_\eta(u_1, u_1) = \pi^2.$$

When $\mathsf{sin}\Theta_\eta(u_1) < 1$, Theorem 4.17 guarantees the existence of an eigenvalue $\lambda_{i_1}^\eta$ such that

$$\frac{|\lambda_{i_1}^\eta - \lambda_1^\infty|}{\lambda_1^\infty} \le \sin\Theta_\eta(u_1)\ .$$

A direct computation shows that

$$\begin{aligned}
&(u_1, \mathbf{H}_\eta^{-1}u_1 - \mathbf{H}_\infty^\dagger u_1) \\
&\quad = \int_0^1 \left[\int_0^x 2\left(\frac{y\ \big(1 + (1+\eta^2)\ (1-x)\big)}{2+\eta^2} - y\ (1-x) \right) \sin(\pi\, y)\sin(\pi\, x)\ dy \right. \\
&\qquad \left. + \int_x^1 2\left(\frac{x\ \big(1 + (1+\eta^2)\ (1-y)\big)}{2+\eta^2} - x\ (1-y) \right) \sin(\pi\, y)\sin(\pi\, x)\, dy \right] dx \\
&\quad = \frac{2}{(2+\eta^2)\pi^2} = O(\eta^{-2}).
\end{aligned} \tag{7.4}$$

This establishes that $\mathsf{sin}\Theta_\eta(u_1) \to 0$, and thus Theorem 4.17 will be applicable for $\eta \ge 1$ such that

$$\frac{(u_1, \mathbf{H}_\eta^{-1}u_1) - (u_1, \mathbf{H}_\infty^\dagger u_1)}{(u_1, \mathbf{H}_\eta^{-1}u_1)} = \frac{2}{4+\eta^2} < 1.$$

Furthermore, based on [11] and [22], we conclude that the assumptions of Theorem 6.2 must be satisfied for η large. We will now investigate this claim further.

The eigenvalues of the operator $\mathbf{H}_\eta$ satisfy the equation

$$\sqrt{1+\eta^2}\cot(\sqrt{\lambda^\eta}) + \cot\left(\sqrt{\frac{\lambda^\eta}{1+\eta^2}}\right) = 0, \tag{7.5}$$

and the nonnormalized eigenvectors are

$$\widehat{v}_i^{\eta}(x) = \begin{cases} \sin(\sqrt{\lambda_i^{\eta}}x), & 0 \le x \le 1, \\ \dfrac{\sin(\sqrt{\lambda_i^{\eta}})}{\sin\left(\sqrt{\frac{\lambda_i^{\eta}}{1+\eta^2}}\right)} \ \sin\left(\sqrt{\dfrac{\lambda_i^{\eta}}{1+\eta^2}}x\right), & 1 \le x. \end{cases}$$

Set $v_i^{\eta} = \|\widehat{v}_i^{\eta}\|^{-1}\ \widehat{v}_i^{\eta}$; then Theorems 5.1 and 6.2 imply

$$\sin\angle(v_1^{\eta}, u_1) = \|v_1^{\eta} - u_1\| \le \frac{\pi\sqrt{\lambda_2^{\eta}}}{\lambda_2^{\eta} - \pi^2}\ \frac{2}{\sqrt{4+\eta^2-\sqrt{8+2\eta^2}}}.$$

From (**??**) we establish the uniform estimate

$$\|v_1^{\eta} - u_1\| \le\ \frac{1.333334}{\sqrt{4+\eta^2-\sqrt{8+2\eta^2}}}, \qquad \eta \ge 2.$$

This illustrates a way to obtain rigorous eigenvector estimates. First, we have localized the approximated eigenvalue by an application of Theorem 5.1. This has selected the approximated eigenvector. Theorem 6.2 then yields an accuracy of that approximation.

Let us note that

$$h_{\infty}(u_n, u_n) = h_{\eta}(u_n, u_n) = n^2\pi^2,$$
$$\sin\Theta_{\eta}(u_n) = \frac{(u_n, \mathbf{H}_{\eta}^{-1}u_n) - (u_n, \mathbf{H}_{\infty}^{\dagger}u_n)}{(u_n, \mathbf{H}_{\eta}^{-1}u_n)} = \frac{2}{4+\eta^2}.$$

This implies that we can get estimates for all λ_i^{η} and v_i^{η} by an analogous procedure. In establishing the convergence results for higher eigenvalues and eigenvectors it was important that we a priori knew that all λ^{η} were nondegenerate. Our theory has successfully been applied to similar singularly perturbed operators which were defined in $L^2(\Omega)$, $\Omega \subset \mathbb{R}^n$; see [8]. For those operators such a claim does not hold. There it is important to generalize the subspace results from [9] as well as to obtain higher order estimates (in $\sin\Theta_{\eta}$) for eigenvalues. These results were obtained in the Ph.D. thesis [8] and will be reported elsewhere.

Remark 7.1. Note that neither the results from [3] nor the results from [17] apply to this problem, since both $\mathcal{D}(\mathbf{H}_{\eta}) \not\subset \mathcal{D}(\mathbf{H}'_{\eta})$ and $\mathcal{D}(\mathbf{H}'_{\eta}) \not\subset \mathcal{D}(\mathbf{H}_{\eta})$. This prevents a direct application of the $\tan\Theta$-theorems from [3, 17]; cf. [17, Theorem 1]. In view of the remarks from section 6.1 we feel that our compromise between the computability of the estimates and the applicability of the theory to the class of operators we are interested in is well suited to the form-theoretic approach.

8. Conclusion. A method to compute an estimate of the accuracy of the subspace approximation method is presented. It can also be used to obtain accurate lower estimates of a desired group of eigenvalues. The bounds have to be viewed as a combination of the Ritz value bound, which gives the existence of the matching of the Ritz values and eigenvalues, and the subspace bound, which describes the nature of that matching.

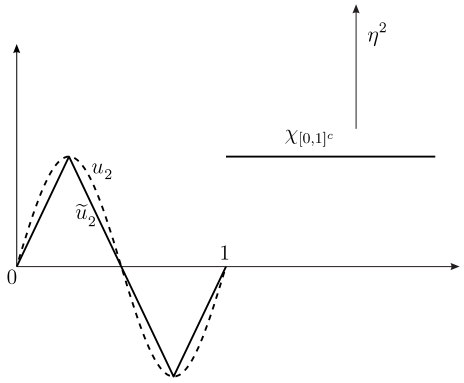


FIG. 8.1. *Various test functions for* $\mathbf{H}_\infty$ *and* $\mathbf{H}_\eta$, η *large.*

The case study that was just performed can be described as leading to a "pseudospectral" method. We have used the completely solvable ("well-behaved") operator

$$(\mathbf{H}_\infty^{1/2}u, \mathbf{H}_\infty^{1/2}v) = h_\infty(u,v) = \int_0^1 u'v'\ dx, \quad u, v \in H_0^1\,[0,1]\,,$$

to analyze the singularly perturbed operator $\mathbf{H}_\eta$. Since the eigenvalue problem for the operator $\mathbf{H}_\infty$ was completely solvable, we have used the eigenfunctions of the operator $\mathbf{H}_\infty$ to define a test space for the operator $\mathbf{H}_\eta$. Analogously, we could have used other test functions from $H_0^1[0,1]$ to analyze the operator $\mathbf{H}_\eta$. For instance, assume that we have used the linear finite elements to compute an approximation $\widetilde{u}_i$ of the function u_i; see Figure 8.1. Theorem 4.17 can be invoked if we find a way to estimate $\sin\Theta(\mathbf{H}_\eta^{-1/2}\widetilde{u}_i, \mathbf{H}_\eta^{1/2}\widetilde{u}_i)$. The study of singularly perturbed eigenvalue problems and finite element spectral approximations has been performed in [8]. The results will be presented in subsequent reports.

Acknowledgments. The author would like to thank Prof. Dr. Krešimir Veselić, Hagen, for helpful discussions and support during the research and the preparation of this manuscript. The author also thanks an anonymous referee for a stimulating report and for pointing out reference [17]. Special thanks go to Dr. Josip Tambača, Zagreb, for many stimulating discussions.

REFERENCES

[1] P. ARBENZ AND Z. DRMAČ, *On positive semidefinite matrices with known null space*, SIAM J. Matrix Anal. Appl., 24 (2002), pp. 132–149.
[2] F. CHATELIN, *Spectral Approximation of Linear Operators*, Academic Press (Harcourt Brace Jovanovich Publishers), New York, 1983.
[3] C. DAVIS AND W. M. KAHAN, *The rotation of eigenvectors by a perturbation.* III, SIAM J. Numer. Anal., 7 (1970), pp. 1–46.
[4] 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.
[5] Z. DRMAČ, *On relative residual bounds for the eigenvalues of a Hermitian matrix*, Linear Algebra Appl., 244 (1996), pp. 155–163.
[6] 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.
[7] W. G. FARIS, *Self-Adjoint Operators*, Lecture Notes in Math. 433, Springer-Verlag, Berlin, 1975.

[8] L. Grubišić, *Ritz Value Estimates and Applications in Mathematical Physics*, Ph.D. thesis, Fernuniversität in Hagen, dissertation.de Verlag im Internet (ISBN: 3-89825-998-6), 2005.

[9] L. Grubišić and K. Veselić, *On Ritz approximations for positive definite operators* I *(theory)*, Linear Algebra Appl., to appear.

[10] L. Grubišić and K. Veselić, *On weakly formulated Sylvester equation and applications*, Integral Equations Operator Theory, submitted; preprint available from http://arxiv.org/abs/math/0507532.

[11] R. Hempel and O. Post, *Spectral gaps for periodic elliptic operators with high contrast: An overview*, in Progress in Analysis, Vols., I, II (Berlin, 2001), World Scientific Publishing, River Edge, NJ, 2003, pp. 577–587.

[12] W. Kahan, *Inclusion Theorems for Clusters of Eigenvalues of Hermitian Matrices*, Technical report, Computer Science Department, University of Toronto, Toronto, ON, 1967.

[13] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Grundlehren Math. Wiss. 132, Springer-Verlag, Berlin, 1976.

[14] V. Kostrykin, K. A. Makarov, and A. K. Motovilov, *On the existence of solutions to the operator Riccati equation and the* $\tan\Theta$ *theorem*, Integral Equations Operator Theory, 51 (2005), pp. 121–140.

[15] S. Levendorskiĭ, *Asymptotic Distribution of Eigenvalues of Differential Operators*, Math. Appl. (Soviet Series) 53 (translated from the Russian), Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990.

[16] R. Mathias and K. Veselić, *A relative perturbation bound for positive definite matrices*, Linear Algebra Appl., 270 (1998), 315–321.

[17] A. K. Motovilov and A. V. Selin, *Some Sharp Norm Estimates in the Subspace Perturbation Problem*, preprint, http://arxiv.org/abs/math.SP/0409558 (2004).

[18] Z. M. Nashed, *Perturbations and approximations for generalized inverses and linear operator equations*, in Generalized Inverses and Applications (Madison, WI, 1973), Publication of the Mathematical Research Center of the University of Wisconsin 32, Academic Press, New York, 1976, pp. 325–396.

[19] M. Reed and B. Simon, *Methods of Modern Mathematical Physics.* I–IV, Academic Press (Harcourt Brace Jovanovich Publishers), New York, 1978.

[20] K. Veselić and I. Slapničar, *Floating-point perturbations of Hermitian matrices*, Linear Algebra Appl., 195 (1993), pp. 81–116.

[21] P.-Å. Wedin, *On angles between subspaces of a finite dimensional inner product space*, in Matrix Pencils (Proceedings of the Conference Held at Pite Havsbad, March 22–24, 1982), Springer-Verlag, New York, Berlin, 1983, pp. 263–286.

[22] J. Weidmann, *Stetige Abhängigkeit der Eigenwerte und Eigenfunktionen elliptischer Differentialoperatoren vom Gebiet*, Math. Scand., 54 (1984), pp. 51–69.

[23] J. Weidmann, *Spectral Theory of Ordinary Differential Operators*, Lecture Notes in Math. 1258, Springer-Verlag, Berlin, 1987.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 4, pp. 1126–1156

ACCURATE SYMMETRIC RANK REVEALING AND EIGENDECOMPOSITIONS OF SYMMETRIC STRUCTURED MATRICES*

FROILÁN M. DOPICO† AND PLAMEN KOEV‡

Abstract. We present new $O(n^3)$ algorithms that compute eigenvalues and eigenvectors to high relative accuracy in floating point arithmetic for the following types of matrices: symmetric Cauchy, symmetric diagonally scaled Cauchy, symmetric Vandermonde, and symmetric totally nonnegative matrices when they are given as products of nonnegative bidiagonal factors. The algorithms are divided into two stages: the first stage computes a symmetric rank revealing decomposition of the matrix to high relative accuracy, and the second stage applies previously existing algorithms to this decomposition to get the eigenvalues and eigenvectors. Rank revealing decompositions are also interesting in other problems, such as the numerical determination of the rank and the approximation of a matrix by a matrix with smaller rank.

Key words. eigenvalue, eigenvector, high relative accuracy, symmetric rank revealing factorization, Cauchy matrix, Vandermonde matrix, totally nonnegative matrix

AMS subject classifications. 65F15, 65F30

DOI. 10.1137/050633792

1. Introduction.

When traditional algorithms are used to compute the eigenvalues and eigenvectors of ill-conditioned *real symmetric matrices* in floating point arithmetic, only the eigenvalues with largest absolute values are computed with guaranteed relative accuracy. The tiny eigenvalues may be computed with no relative accuracy at all—and even with the wrong sign. The eigenvectors are computed with small error with respect to the absolute eigenvalue gap. This means that if ϵ is the machine precision, and v_i and $\hat{v}_i$ are, respectively, the exact and computed eigenvectors corresponding to an eigenvalue λ_i, then the acute angle between these vectors is bounded as $\theta(v_i, \hat{v}_i) \leq O(\epsilon)/\mathrm{gap}_i$, where $\mathrm{gap}_i = (\min_{j \neq i} |\lambda_i - \lambda_j|)/\max_k |\lambda_k|$. This implies that if there is more than one tiny eigenvalue, then the corresponding eigenvectors are computed with large errors, even if the tiny eigenvalues are well separated in the relative sense. See [1, section 4.7] for a survey on errors bounds for the symmetric eigenproblem.

Our goal is to derive algorithms for computing eigenvalues and eigenvectors of some structured $n \times n$ symmetric matrices *to high relative accuracy by respecting the symmetry of the problem*, and with cost $O(n^3)$, i.e., roughly the same cost as traditional algorithms for dense symmetric matrices. By *high relative accuracy* we mean that the eigenvalues λ_i, the eigenvectors v_i, and their computed counterparts

*Received by the editors June 16, 2005; accepted for publication (in revised form) by I. C. F. Ipsen April 19, 2006; published electronically December 18, 2006. This research was partially supported by the Ministerio de Educación y Ciencia of Spain through grant BFM-2003-00223, by the PRICIT Program of the Comunidad de Madrid through SIMUMAT Project grant S-0505/ESP/0158 (Froilán M. Dopico), and by NSF grant DMS-0314286 (Plamen Koev). This research was partially performed while P. Koev was visiting the Universidad Carlos III de Madrid in December 2004–January 2005 and was supported by the Ph.d. Program of Ingeniería Matemática.

http://www.siam.org/journals/simax/28-4/63379.html

†Departamento de Matemáticas, Universidad Carlos III de Madrid, Avda. Universidad 30, 28911 Leganés, Spain (dopico@math.uc3m.es).

‡Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139 (plamen@math.mit.edu).

$\hat{\lambda}_i$ and $\hat{v}_i$ will satisfy

$$(1) \qquad |\hat{\lambda}_i - \lambda_i| \leq O(\epsilon)|\lambda_i| \quad \text{and} \quad \theta(v_i, \hat{v}_i) \leq \frac{O(\epsilon)}{\min\limits_{j \neq i} \left| \frac{\lambda_i - \lambda_j}{\lambda_i} \right|} \quad \text{for} \quad i = 1, \ldots, n.$$

These conditions guarantee that the new algorithms compute *all* eigenvalues, including the tiniest ones, with correct sign and leading digits. Moreover, the eigenvectors corresponding to relatively well separated tiny eigenvalues are accurately computed. In the case of a multiple eigenvalue, or a cluster of very close eigenvalues in the relative sense, the previous bound for $\theta(v_i, \hat{v}_i)$ becomes infinite or very large. In this case, we understand by high relative accuracy that the sines of the canonical angles between the unperturbed and the perturbed invariant subspaces corresponding to the cluster of eigenvalues are bounded by $O(\epsilon)$ over the relative gap between the eigenvalues inside the cluster and those outside the cluster [27]. This means that the new algorithms compute accurate bases of invariant subspaces corresponding to cluster of eigenvalues well separated in the relative sense from the rest of the eigenvalues.

In this work, we focus on the following classes of symmetric matrices: diagonally scaled Cauchy matrices (this class includes usual symmetric Cauchy matrices), Vandermonde matrices, and nonsingular totally nonnegative (TN) matrices. Symmetric diagonally scaled Cauchy matrices are defined through two ordered sets of real numbers, $\{x_1, x_2, \ldots, x_n\}$ and $\{s_1, s_2, \ldots, s_n\}$, and they are of the form

$$C = SC'S, \quad \text{where} \quad C'_{ij} = \frac{1}{x_i + x_j}, \quad 1 \leq i, j \leq n, \quad \text{and} \quad S = \operatorname{diag}(s_1, s_2, \ldots, s_n);$$

i.e., they are the two-sided product of a usual symmetric Cauchy matrix C' times a diagonal matrix S. *It should be noticed that if S is the identity matrix, then $C = C'$, and C is just a usual symmetric Cauchy matrix.* Symmetric Vandermonde matrices depend only on one real parameter a, and they are defined as

$$A = \left[a^{(i-1)(j-1)} \right]_{i,j=1}^{n}.$$

This is the only type of Vandermonde matrices that is symmetric. As far as we know, this is the first time that the class of symmetric Vandermonde matrices has been studied in the literature. TN matrices are the matrices with all minors nonnegative. For symmetric diagonally scaled Cauchy matrices, we assume that the parameters $\{x_i\}_{i=1}^n$ and $\{s_i\}_{i=1}^n$ are given, i.e., we are not given just the entries of the matrices. This is a very natural assumption in situations where Cauchy matrices appear, such as, for instance, in rational interpolation theory. For symmetric Vandermonde matrices, we adopt the (also natural) assumption that the parameter a is given. In the case of TN matrices, we assume that the TN structure is explicitly revealed; i.e., any TN matrix is represented as a product of nonnegative bidiagonal matrices [18, 19]. This bidiagonal decomposition is particularly attractive because its nontrivial entries determine the eigenvalues of the matrix with high relative accuracy, and it can be computed very accurately for many important classes of TN matrices [26]. To finish this short presentation of the type of matrices we are dealing with, we want to stress that the symmetric diagonally scaled Cauchy and the symmetric Vandermonde matrices are, in general, indefinite matrices, while the symmetric nonsingular TN matrices are positive definite.

There exist $O(n^3)$ algorithms for computing eigendecompositions of symmetric diagonally scaled Cauchy and symmetric Vandermonde matrices with high relative

accuracy, *but these algorithms do not respect the symmetry of the problem.* They are based on the idea of rank revealing decomposition (RRD): an RRD of $G \in \mathbb{R}^{m\times n}$, $m \geq n$, is a factorization $G = XDY^T$, where $D \in \mathbb{R}^{r\times r}$ is diagonal and nonsingular, and $X \in \mathbb{R}^{m\times r}$ and $Y \in \mathbb{R}^{n\times r}$ are well-conditioned matrices of full column rank (notice that this implies $r = \mathrm{rank}(G)$). Demmel et al. presented in [6] an algorithm for computing the singular value decomposition (SVD) of G with high relative accuracy when the computed factors $\widehat{X}, \widehat{D}$, and $\widehat{Y}$ of an RRD satisfy the following forward error bounds:

$$\begin{aligned} |D_{ii} - \widehat{D}_{ii}| &= O(\epsilon)|D_{ii}|, \\ \|X - \widehat{X}\|_2 &= O(\epsilon)\|X\|_2, \\ \|Y - \widehat{Y}\|_2 &= O(\epsilon)\|Y\|_2, \end{aligned} \tag{2}$$

where $\|\cdot\|_2$ is the spectral, or two-norm. Throughout this paper we will use the expression *accurate RRD* to mean an RRD that satisfies the error bounds (2). Algorithms for computing accurate RRDs of general diagonally scaled Cauchy and Vandermonde matrices were derived in [5], and therefore it is possible to compute the SVD of these matrices with high relative accuracy. Finally, an algorithm for computing a high relative accuracy eigendecomposition of a symmetric matrix, given an SVD computed with high relative accuracy, was developed in [11]. We note that when these algorithms are used, the symbols $O(\epsilon)$ appearing in (1) should be replaced with $O(\max\{\kappa_2(X), \kappa_2(Y)\}\,\epsilon)$, where $\kappa_2(X) \equiv \|X\|_2 \cdot \|X^{-1}\|_2$ is the spectral condition number of X.

The process outlined in the previous paragraph does not respect the symmetry of the problem in two stages. First, the RRDs of diagonally scaled Cauchy and Vandermonde matrices computed in [5] are not symmetric, i.e., $X \neq Y$, when G is symmetric. Second, even when G is symmetric and $X = Y$, the algorithm in [6] computes the SVD of G without respecting the symmetry of the problem. Respecting the symmetry is a very important property of eigenvalue algorithms (as well as other computations in the field of numerical linear algebra) because it often leads to increased speed, decreased storage requirements, and improved stability properties [3, 10, 21].

As two of our major contributions we present algorithms for computing accurate *symmetric* RRDs of symmetric diagonally scaled Cauchy matrices and symmetric Vandermonde matrices, i.e., decompositions $G = XDX^T$ with X well conditioned and D diagonal, which satisfy the bounds (2). In this context, it is important to stress that RRDs have been computed in practice as LDU factorizations provided by Gaussian elimination with complete pivoting (GECP) [6]. As can be seen in [21, section 4.4] and [22, Chapter 11], just preserving the symmetry of general dense symmetric indefinite matrices in a stable factorization of LU type requires much more complicated algorithms and pivoting strategies than the usual Gaussian elimination. In our algorithms, we need to preserve the symmetry and *also* attain the accuracy (2). This demands a careful exploitation of the structure of the problems that allows us to get important benefits from the point of view of operational cost. The algorithm we present for computing RRDs of symmetric diagonally scaled Cauchy matrices needs only half the operations required by the general nonsymmetric algorithm presented in [5]. In the case of symmetric Vandermonde matrices, the improvements are much more significant: the cost of the algorithm in [5] is $O(n^3)$ and requires complex arithmetic, and the cost of the algorithm we develop is $2n^2$ and requires only real arithmetic. We note, however, that for symmetric Vandermonde matrices our algorithm computes

accurate RRDs only if $|a| \leq \frac{2}{3}$ or $|a| \geq \frac{3}{2}$. For the rest of the values of the parameter, i.e., $\frac{2}{3} < |a| < \frac{3}{2}$, our algorithm computes LDL^T factorizations with componentwise relative errors of $O(\epsilon)$, but they are not RRDs because L may be ill conditioned. This means that the factorizations $A = LDL^T$ we compute of symmetric Vandermonde matrices cannot be used to compute accurate eigendecompositions for values of $|a|$ close to one. However, they can be potentially useful in other contexts such as, for instance, in fast solvers of systems of linear equations $Ax = b$, where A is a symmetric Vandermonde matrix. The operational savings we have just described may not be of primary interest for computing accurate eigendecompositions, because in that case an $O(n^3)$ algorithm with high cost has to be applied to the RRD, but they are very important in other applications of RRDs.

Once an accurate symmetric RRD of a symmetric indefinite matrix G is computed, the J-orthogonal algorithm, introduced in [35] and carefully analyzed in [33], can be used to compute an eigendecomposition of G to high relative accuracy, preserving the symmetry of the process. Also, the signed SVD algorithm of [11] may be used, but then the symmetry is lost in this second stage. It should be noticed that the error bounds for the J-orthogonal algorithm [33] are not exactly of type (1) because the $O(\epsilon)$ symbols are rigorously $\kappa\epsilon$, where κ is the maximum of the condition numbers of some intermediate matrices appearing in the algorithm, which has not been bounded by any moderate magnitude. The error bounds for the signed SVD algorithm [11] are exactly of type (1) because the error for the eigenvectors depends on a different, smaller eigenvalue relative gap than the one in (1). However, in practice, both the J-orthogonal and signed SVD algorithms compute the eigenvalues and eigenvectors to high relative accuracy.

Our third major contribution is to develop algorithms for computing accurate RRDs of a nonsingular TN matrix whenever its bidiagonal factors are given. RRDs of general, not necessarily symmetric, TN matrices can be computed by combining algorithms in [26] and in [6], but the computation of *symmetric RRDs* requires a new approach. It should be remarked that algorithms for computing eigenvalues and singular values of general nonsingular TN matrices already have been presented in [26]. If the TN matrix is symmetric, the techniques in [26] allow us to modify these algorithms to compute eigenvalues to high relative accuracy *respecting the symmetry.* However, the algorithms in [26] do not use RRDs computed by a finite process.

Nonsingular symmetric TN matrices are positive definite; thus a symmetric RRD $A = XDX^T$ has positive elements on the diagonal matrix D. In this case we can compute an accurate eigendecomposition of A starting from this RRD, using a simpler and more efficient approach than the J-orthogonal or signed SVD algorithms. To do so, we compute the singular values and left singular vectors of $XD^{1/2}$ by using the one-sided Jacobi method with the rotations applied on the left [10, section 5.4.3] (see also the seminal reference [9]). This yields eigenvalues and eigenvectors with high relative accuracy as in (1), where the $O(\epsilon)$ symbols are replaced with $O(\epsilon\,\kappa_2(X))$. Obviously, this process preserves the symmetry.

In the previous paragraphs we have stressed the essential role of accurate RRDs in computing spectral problems to high relative accuracy. However, the computation of accurate RRDs is an interesting problem in its own right that can be used in other problems, such as the numerical determination of the rank, and the approximation of a matrix by a matrix with smaller rank [34, Chapter 5]. This is one of the reasons why reducing the cost in computing accurate RRDs is an important issue.

The three classes of symmetric matrices we consider—diagonally scaled Cauchy, Vandermonde, and TN—require three very different techniques for computing their accurate symmetric RRDs. In this regard, in [30] accurate symmetric RRDs of total signed compound and diagonally scaled totally unimodular matrices are computed by using an approach related to the one we used for diagonally scaled Cauchy matrices, i.e., combining accurate computation of Schur complements with the Bunch–Parlett pivoting strategy for the diagonal pivoting method [4]. Two other interesting classes of structured matrices for which there are algorithms for computing accurate RRDs are weakly diagonally dominant M-matrices [7, 31] and polynomial Vandermonde matrices [8]. For *symmetric* weakly diagonally dominant M-matrices, the general algorithm presented in [7] for nonsymmetric matrices respects the symmetry because it performs only diagonal pivoting. The algorithm in [8] does not preserve the symmetry for symmetric matrices, but the symmetric polynomial Vandermonde matrices are nonsymmetric, except in very special cases.

The paper is organized as follows. In section 2 we study how the eigenvalues and eigenvectors of a symmetric matrix are changed by errors of type (2) in a symmetric RRD. In section 3 we present the algorithm, and its error analysis, for computing symmetric RRDs of symmetric diagonally scaled Cauchy matrices. The same is done in section 4 for symmetric Vandermonde matrices. Section 5 includes the algorithms for computing accurate RRDs (symmetric and nonsymmetric) of nonsingular TN matrices. We present numerical experiments in section 6. Finally, in the appendix the technical proof of Theorem 3.1 for the rounding error analysis of diagonally scaled Cauchy matrices is carefully developed in a more general setting.

2. Perturbation properties of symmetric RRDs. Let G be an $m\times n$ matrix, and let $G = XDY^T$ be an RRD of G. It was shown in [6, Theorem 2.1] that the RRD of G determines its SVD to high relative accuracy; i.e., small relative normwise perturbations of X and Y, and small relative componentwise perturbations of D, produce small relative changes in all singular values of G, and produce small changes in the singular vectors with respect to the singular value relative gap. Next we prove that a symmetric RRD of a symmetric matrix determines its eigenvalues and eigenvectors to high relative accuracy.

THEOREM 2.1. *Let $A = XDX^T$ and $\widetilde{A} = \widetilde{X}\widetilde{D}\widetilde{X}^T$ be RRDs of the real symmetric $n\times n$ matrices A and $\widetilde{A}$. Let $\lambda_1 \geq \cdots \geq \lambda_n$ be the eigenvalues of A and $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_n$ be the eigenvalues of $\widetilde{A}$. Let $q_1, \ldots, q_n$ and $\tilde{q}_1, \ldots, \tilde{q}_n$ be the corresponding orthonormal eigenvectors. Let us assume that*

$$\frac{\|\widetilde{X} - X\|_2}{\|X\|_2} \leq \beta,$$
$$\frac{|\widetilde{D}_{ii} - D_{ii}|}{|D_{ii}|} \leq \beta \quad \textit{for all } i,$$

where $0 \leq \beta < 1$. Let $\eta = \beta\,(2+\beta)\,\kappa_2(X)$ be smaller than 1; then

$$|\lambda_i - \tilde{\lambda}_i| \;\leq\; (2\eta + \eta^2)\,|\lambda_i|, \qquad 1 \leq i \leq n,$$

and

$$\sin\theta(q_i, \tilde{q}_i) \leq \frac{\eta}{1-\eta}\left(1 + \frac{2+\eta}{\min_{j\neq i}\frac{|\tilde{\lambda}_i - \lambda_j|}{|\lambda_j|}}\right), \qquad 1 \leq i \leq n,$$

where $\theta(q_i, \tilde{q}_i)$ is the acute angle between q_i and $\tilde{q}_i$. In the case of multiple eigenvalues, or clusters of very close eigenvalues in the relative sense, a similar bound holds for the sines of the canonical angles of the corresponding invariant subspaces.

Proof. The proof is similar to that of Theorem 2.1 in [6]. The main idea is to express $\widetilde{A}$ as a *symmetric* multiplicative perturbation of A, i.e., $\widetilde{A} = (I+E)A(I+E)^T$. This is combined with [12, Theorem 2.1] and [27, Theorem 3.1]. ∎

A more general version of Theorem 2.1, including similar perturbation results for invariant subspaces [27], can be developed. These bounds are useful when several eigenvalues form a tight cluster, well separated from the remaining eigenvalues, because in this case the invariant subspace is well conditioned, while the individual eigenvectors are very ill conditioned. It is also possible to present perturbation results for eigenvectors with the relative gap defined using exclusively eigenvalues of A, at the cost of bounding the sine of the double angle, i.e., $\sin 2\theta(q_i, \tilde{q}_i)$ [10, Theorem 5.7], [28, Theorem 2.2].

3. Symmetric diagonally scaled Cauchy matrices. For a real symmetric matrix A, the LU factorization computed using Gaussian elimination, with partial or complete pivoting, does not always preserve the symmetry of the problem. Symmetric pivoting strategies, i.e., permuting rows and columns in the same way, may be unstable or may not exist. A trivial instance is when all the entries on the main diagonal are zero. The most widely used factorization [1, 21, 22] for symmetric matrices is the following special block LU factorization:

$$PAP^T = L\, D_b\, L^T,$$

where P is a permutation matrix, L is unit lower triangular, and D_b is block diagonal with diagonal blocks of dimension 1 or 2. The 2×2 diagonal blocks are symmetric indefinite matrices, and the corresponding diagonal blocks of L are the 2×2 identity matrix. This method is sometimes called the *diagonal pivoting method* [22] and can be implemented with partial, complete, or rook pivoting. We are interested in computing a symmetric RRD; therefore we will focus on the *Bunch–Parlett complete pivoting strategy* [4], which in practice[1] produces a well-conditioned matrix L. Notice that $L\, D_b\, L^T$ is not an RRD because D_b is not diagonal. To get an RRD, we will perform a spectral factorization of each of the 2×2 blocks of D_b; thus $D_b = VDV^T$ with D diagonal and V orthogonal and block diagonal as D_b. Finally,

$$PAP^T = L\, D_b\, L^T = (LV)D(LV)^T \equiv XDX^T \tag{3}$$

is a symmetric RRD. This procedure has been essentially introduced in [32] to compute a symmetric indefinite decomposition GJG^T, where $J = \mathrm{diag}(\pm 1)$. Notice that a GJG^T factorization can be easily computed from XDX^T as $(X\sqrt{|D|})\, J\, (\sqrt{|D|}X^T)$. Moreover, if XDX^T is accurately computed, then GJG^T is also accurately computed, and vice versa. In the rest of the paper we will focus on RRDs XDX^T from the point of view of both algorithms and error analysis.

To be more specific, the method can be described as follows. Let Π be a permutation matrix such that

$$\Pi A \Pi^T = \begin{bmatrix} E & C^T \\ C & B \end{bmatrix}, \tag{4}$$

[1] It can be proven that $\kappa_\infty(L) < n\,(3.78)^n$, by using Theorem 8.12 and Problem 8.5 in [22]. This bound is similar to that appearing in GECP. Therefore, there exists a remote possibility of the Bunch–Parlett pivoting strategy failing to compute a well-conditioned factor L.

where E is a 1×1 or a 2×2 nonsingular matrix. The pivot E and the permutation Π are chosen by comparing the numbers $\mu_0 = \max_{i,j} |a_{ij}| \equiv |a_{rs}|$ $(r \geq s)$ and $\mu_1 = \max_i |a_{ii}| \equiv |a_{pp}|$. If $\mu_1 \geq \alpha\mu_0$, where α is a parameter $(0 < \alpha < 1)$, then $E = a_{pp}$, and if $\mu_1 < \alpha\mu_0$, then E has dimension 2 and $E_{21} = |a_{rs}|$. The classical value for the parameter is $\alpha = (1 + \sqrt{17})/8$ (≈ 0.64). Then we can factorize

$$\Pi A \Pi^T = \begin{bmatrix} I & 0 \\ CE^{-1} & I \end{bmatrix} \begin{bmatrix} E & 0 \\ 0 & B - CE^{-1}C^T \end{bmatrix} \begin{bmatrix} I & E^{-1}C^T \\ 0 & I \end{bmatrix}. \tag{5}$$

If E is a 2×2 matrix, let $E = U\Lambda U^T$ be its orthogonal spectral factorization computed by the Jacobi procedure [21, section 8.4]. Then

$$\Pi A \Pi^T = \begin{bmatrix} U & 0 \\ CU\Lambda^{-1} & I \end{bmatrix} \begin{bmatrix} \Lambda & 0 \\ 0 & B - CE^{-1}C^T \end{bmatrix} \begin{bmatrix} U^T & \Lambda^{-1}U^TC^T \\ 0 & I \end{bmatrix}. \tag{6}$$

The process is recursively repeated on the Schur complement $B - CE^{-1}C^T$.

In the case of diagonally scaled Cauchy matrices, it was shown in [5] how to compute all the Schur complements with an entrywise small relative error. Therefore, to compute an accurate symmetric RRD, the remaining task is to show that in (6) the orthogonal diagonalization $E = U\Lambda U^T$ of the 2×2 pivot and the matrix $CU\Lambda^{-1}$ can be accurately computed for each Schur complement.

Let us summarize some key results in [5]. The entries of an $n \times n$ symmetric diagonally scaled Cauchy matrix C are $C_{ij} = s_i s_j/(x_i + x_j)$, where the s_i and x_i, $1 \leq i \leq n$, are given real floating point numbers. Let $S^{(m)}$ be the mth Schur complement of C $(S^{(0)} \equiv C)$. We enumerate the elements of $S^{(m)}$ as the corresponding elements of C. The recurrence relation,

$$S_{rs}^{(m)} = S_{rs}^{(m-1)} \frac{(x_r - x_m)(x_s - x_m)}{(x_m + x_s)(x_r + x_m)} \qquad \text{for} \qquad m+1 \leq r, s \leq n, \tag{7}$$

allows us to compute accurately each Schur complement from the previous one. This is what we need when the Bunch–Parlett pivoting strategy selects a 1×1 pivot. If a 2×2 pivot is selected, we apply (7) twice to obtain

$$S_{rs}^{(m+1)} = S_{rs}^{(m-1)} \frac{(x_r - x_m)(x_s - x_m)}{(x_m + x_s)(x_r + x_m)} \cdot \frac{(x_r - x_{m+1})(x_s - x_{m+1})}{(x_{m+1} + x_s)(x_r + x_{m+1})}. \tag{8}$$

Combining (7) and (8) with (6), we get the following algorithm to compute a symmetric RRD of a symmetric diagonally scaled Cauchy matrix.[2]

ALGORITHM 1. Symmetric RRD of a symmetric diagonally scaled Cauchy matrix.

`Input:` $S = \{s_1, \ldots, s_n\}$; $x = \{x_1, \ldots, x_n\}$

`Output:`

D `is a` rank $\times$ `rank diagonal matrix, where rank is the rank of the diagonally scaled Cauchy matrix defined by` S `and` x.

X `is an` $n \times$ `rank block lower triangular matrix, with diagonal blocks of dimension` 1×1 `or` 2×2.

`IPIV is an` n`-dimensional vector containing a permutation of` $\{1, \ldots, n\}$ `such that, if` $Q = I_n$ `and` $P = Q(\texttt{IPIV}, :)$, `then`

$$P \left[\frac{s_i s_j}{x_i + x_j} \right]_{i,j=1}^{n} P^T = XDX^T.$$

[2] We will use MATLAB [29] notation for submatrices, e.g., $A(i : j, k : l)$ will indicate the submatrix of A consisting of rows i through j and columns k through l, and $A(:, k : l)$ will indicate the submatrix of A consisting of columns k through l.

```
% Initializing variables
  α = (1 + √17)/8 ≈ 0.64
  rank = n
  IPIV = 1 : n
  D = zeros(n)
  for p = 1 : n and q = 1 : p
      A(p, q) = s_p s_q/(x_p + x_q)
      A(q, p) = A(p, q)
  endfor
% Main loop
  k = 1
  while k ≤ n
      μ_0 = maximum entry of |A(k : n, k : n)| ≡ |A(r, s)|  (r ≥ s)
      μ_1 = maximum entry of diag(|A(k : n, k : n)|) ≡ |A(p, p)|
      if μ_1 ≥ αμ_0
          if μ_1 = 0
              rank = k − 1
              k = n + 1
          else
              swap entries k ↔ p in IPIV
              swap entries k ↔ p in x
              swap rows k ↔ p and swap columns k ↔ p in A
              for r = k + 1 : n and s = k + 1 : r
                  A(r, s) = A(r, s) (x_r − x_k)(x_s − x_k)/((x_k + x_s)(x_r + x_k))
                  A(s, r) = A(r, s)
              endfor
              D(k, k) = A(k, k)
              A(k : n, k) = A(k : n, k)/A(k, k)
              A(k, k + 1 : n) = zeros(1, n − k)
              k = k + 1
          endif
      else
          swap entries k ↔ s and swap entries k + 1 ↔ r in IPIV
          swap entries k ↔ s and swap entries k + 1 ↔ r in x
          swap rows k ↔ s and swap rows k + 1 ↔ r in A
          swap columns k ↔ s and swap columns k + 1 ↔ r in A
          for r = k + 2 : n and s = k + 2 : r
              A(r, s) = A(r, s) (x_r − x_k)(x_s − x_k)(x_r − x_{k+1})(x_s − x_{k+1})/((x_k + x_s)(x_r + x_k)(x_{k+1} + x_s)(x_r + x_{k+1}))
              A(s, r) = A(r, s)
          endfor
% Orthogonal diagonalization of the 2 × 2 pivot A(k : k + 1, k : k + 1)
          z = (A(k + 1, k + 1) − A(k, k))/A(k + 1, k)/2
          if z = 0
              t = 1
          else
              t = sign(z)/ (abs(z) + √(1 + z^2))
          endif
```

$cs = 1/\sqrt{1+t^2}$
$sn = t \cdot cs$
$U = \left[\begin{smallmatrix} cs & sn \\ -sn & cs \end{smallmatrix}\right]$
$D(k,k) = A(k,k) - t \cdot A(k+1,k)$
$D(k+1,k+1) = A(k+1,k+1) + t \cdot A(k+1,k)$
$A(k:k+1, k:k+1) = U$
$A(k+2:n, k:k+1) = A(k+2:n, k:k+1) \cdot U \cdot \mathrm{diag}[\frac{1}{D(k,k)}, \frac{1}{D(k+1,k+1)}]$
$A(k:k+1, k+2:n) =$ `zeros(2, n-k-1)`
$k = k+2$
endif
endwhile
$X = A(:, 1:\texttt{rank})$
$D = D(1:\texttt{rank}, 1:\texttt{rank})$
$Q =$ `eye`(n)
$P = Q(IPIV, :)$

The cost of Algorithm 1 is $4\,n^3/3 + O(n^2)$ flops, or $2\,n^3/3 + O(n^2)$ if all n^2 possible values of $(x_r - x_m)$ and $1/(x_r + x_m)$ are precomputed. Next, we show that the computed symmetric RRD is accurate.

THEOREM 3.1. *Let*

$$C = \left[\frac{s_i s_j}{x_i + x_j}\right]_{i,j=1}^{n}$$

be a real symmetric diagonally scaled Cauchy matrix, where $s_1, \ldots, s_n$ and $x_1, \ldots, x_n$ are floating point numbers. Let P, $\widehat{X}$, and $\widehat{D}$ be the matrices of the factorization (3) *computed by Algorithm* 1 *applied to C in floating point arithmetic with machine precision ϵ. Let us apply Algorithm* 1 *in exact arithmetic to C, but choosing the same dimensions and positions for the pivots as those selected in floating point arithmetic. Let X and D be the exact factors; thus $PCP^T = XDX^T$. If*

$$\frac{648\,(n+2)\,\epsilon}{1 - 648\,(n+2)\,\epsilon} < 1,$$

then

1.
$$|\widehat{D}_{ii} - D_{ii}| \le \frac{146\,(n+4)\,\epsilon}{1 - 146\,(n+4)\,\epsilon} |D(i,i)| \qquad \textit{for all} \qquad i = 1, \ldots, n.$$

2.
$$||\widehat{X} - X||_F \le 13\,\frac{684\,(n+2)\,\epsilon}{1 - 684\,(n+2)\,\epsilon}\,||X||_F.$$

If, moreover,

$$\frac{12481\,n\,\epsilon}{1 - 12481\,n\,\epsilon} < \frac{1}{2},$$

then

3.

$$||\widehat{X}(:,j) - X(:,j)||_2 \leq 144\,\sqrt{n}\,\frac{684\,(n+2)\,\epsilon}{1-684\,(n+2)\,\epsilon}\,||X(:,j)||_2 \quad \textit{for all} \quad j=1,\ldots,n.$$

According to Theorem 2.1 and (2), the third item in Theorem 3.1 is not necessary for computing accurate eigenvalues and eigenvectors. It is included for the sake of completeness and because it allows us to state error bounds for the column scaling of X with minimum condition number. We remark that the numerical constants appearing in the bounds above are not optimal: we have sometimes overestimated the constants to get simpler bounds. However, the order of magnitude is correct up to a factor smaller than 10. Theorem 3.1 remains valid if the rank, say ρ, of the matrix is less that n. In this case, the last $n-\rho$ diagonal elements of D are exactly computed to be zero, and the corresponding columns of X are just the $n-\rho$ columns of the identity matrix, and they are also exactly computed.

The proof of Theorem 3.1 is technical and is presented in the appendix. However, the argument explaining why Algorithm 1 accurately computes a symmetric rank revealing factorization of the diagonally scaled Cauchy matrix C can be easily understood. In the first place, the recurrence relation (7) allows us to compute the entries of the Schur complements with a relative error bounded by $8n\epsilon/(1-8n\epsilon)$. Therefore, the elements of D and the entries of the columns of X corresponding to 1×1 pivots are also computed with small relative errors. For the quantities corresponding to 2×2 pivots, the error analysis heavily depends on the properties of these pivots. As we will prove in the appendix, the 2×2 pivots selected by the Bunch–Parlett complete pivoting strategy are very well-conditioned indefinite matrices (with a spectral condition number less than 4.6 for the value $\alpha = 0.64$ used in Algorithm 1), and the entries of their unitary eigenvectors are greater than 0.47 (again for $\alpha = 0.64$). Therefore, the Jacobi algorithm computes with small relative error the eigenvalues (i.e., the elements of D) and the entries of the eigenvectors of the 2×2 pivots. According to (6), the upper 2×2 block of the corresponding two columns of X is just the eigenvector matrix U, and therefore its entries are accurately computed. The rest of the elements of these two columns of X are obtained through multiplying by U and by Λ^{-1}, but these two matrices are well conditioned and all their entries have been computed with small relative error. This last step does not guarantee small entrywise relative errors but it does guarantee small normwise relative errors for X.

4. Symmetric Vandermonde matrices. A symmetric Vandermonde matrix is defined as

$$A = \left[a^{(i-1)(j-1)}\right]_{i,j=1}^{n} = \begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & a & a^2 & \ldots & a^{n-1} \\ 1 & a^2 & a^4 & \ldots & a^{2(n-1)} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & a^{n-1} & a^{2(n-1)} & \ldots & a^{(n-1)^2} \end{bmatrix}, \tag{9}$$

where a is a real number. The class of symmetric Vandermonde matrices depends only on one parameter, and it is the only class of matrices which are, simultaneously, symmetric and of Vandermonde type. Symmetric Vandermonde matrices with $n > 2$ are singular when $a = 0$, $a = 1$, and $a = -1$. In these cases they have only, 2, 1, and 2, respectively, nonzero eigenvalues that can be accurately computed by any standard symmetric eigenvalue algorithm because they are of similar magnitudes. In fact, when

$a = 1$, the only nonzero eigenvalue is equal to n. We assume that a is different from $0, 1$ and -1. The matrix A is positive definite if $a > 1$ and, in this case, A is also totally positive. Therefore, when $a > 1$, an accurate bidiagonal factorization of A can be computed [26, section 3], and its eigenvalues can be obtained to high relative accuracy with the method presented in [26]. The algorithm we introduce in section 5 for computing accurate symmetric RRDs of symmetric totally positive matrices can also be applied to symmetric Vandermonde matrices with $a > 1$.

In this section, we present a method for computing an accurate RRD of A, in the sense of (2), by respecting the symmetry of A. This allows us to compute eigenvalues and eigenvectors to high relative accuracy, as explained in the introduction.

The method we present to compute an accurate RRD of A is very different from the one we used for diagonally scaled Cauchy matrices. The Schur complement of a Vandermonde matrix does not inherit the Vandermonde structure. Moreover, row and column permutations coming from any pivoting strategy also destroy the symmetric Vandermonde structure. Our approach avoids the computation of the Schur complements and, also, avoids pivoting. To be more precise, in the case $|a| < 1$, we use exact formulas for the elements of the LDL^T factorization of A, where L is unit lower triangular and D is diagonal, and we prove that the condition number of L in the 1-norm is $O(n^2)$ when $|a| \leq \frac{2}{3}$. In the case $|a| > 1$, we use exact formulas for the elements of the $\bar{L}\bar{D}\bar{L}^T$ factorization of the *converse* of A, i.e., $A^{\#} \equiv \left[A_{n-i+1,n-j+1}\right]_{i,j=1}^{n}$, and we will prove that $\kappa_1(\bar{L}) = O(n^2)$ when $|a| \geq \frac{3}{2}$. Note that in both cases $|a| \leq \frac{2}{3}$ and $|a| \geq \frac{3}{2}$, we are dealing with matrices whose elements vary widely and in which the largest elements are in the first positions. This is the reason why we are able to get RRDs without using pivoting strategies. The formulas we use allow us to compute accurate LDL^T factorizations for any value of a, but only when $|a| \leq \frac{2}{3}$ or $|a| \geq \frac{3}{2}$ can we guarantee that they are RRDs. These limits are somewhat arbitrary since we can consider values of a closer to $|a| = 1$ at the cost of increasing the bound for $\kappa_1(L)$. However, it should be stressed that we cannot consider values of a as close as we want to $|a| = 1$ because $\kappa_1(L)$ approaches 2^n as $|a|$ approaches 1.

In plain words, there are three limits for which the matrix A is extremely ill conditioned and that have eigenvalues that can vary widely: $|a|$ small enough, $|a|$ large enough, and $|a|$ close enough to 1. We are able to compute eigenvalues and eigenvectors of A to high relative accuracy *by respecting the symmetry* only in the first two cases, i.e., when A contains elements with very different magnitudes. Eigenvalues and eigenvectors *for any value of* a can be computed to high relative accuracy by combining the algorithm presented in [5] to compute a nonsymmetric RRD of A with the signed SVD (SSVD) algorithm in [11], *at the cost of not respecting the symmetry* of the problem.

Consider first the case $|a| < 1$. We start with the LDU decomposition $A = LDL^T$. The entries of L and D are quotient of minors of A [15, section 1.II]:

$$d_i = \frac{\det A(1:i, 1:i)}{\det A(1:i-1, 1:i-1)} = a^{\frac{1}{2}(i-2)(i-1)} \cdot \prod_{t=1}^{i-1}(a^t - 1), \tag{10}$$

$$l_{ij} = \frac{\det A([1:j-1, i], 1:j)}{\det A(1:j, 1:j)} = \prod_{t=1}^{j-1} \frac{1 - a^{i-j+t}}{1 - a^t}. \tag{11}$$

Next, we prove that when $|a| \leq \frac{2}{3}$, the entries of L and L^{-1} are bounded by e^6; thus L is well conditioned.

LEMMA 4.1. *If* $0 \le x \le \frac{2}{3}$ *and* $j \ge 1$, *then*

$$\prod_{t=1}^{j-1} \frac{1}{1-x^t} \le e^6.$$

Proof. We start by observing that $\log(1-x^t) \ge -3x^t$ for $t \ge 1$: If $f(z) = \log(1-z)+3z$, then $f'(z) = \frac{1}{z-1}+3 = \frac{3z-2}{z-1} \ge 0$, meaning $f(z)$ is increasing on $[0, \frac{2}{3}]$ and $f(z) \ge f(0) = 0$ on the same interval. Therefore,

$$\log\left(\prod_{t=1}^{\infty}(1-x^t)\right) \ge -3\sum_{t=1}^{\infty} x^t = \frac{-3x}{1-x} \ge -6$$

and

$$\prod_{t=1}^{j-1} \frac{1}{1-x^t} \le \prod_{t=1}^{\infty} \frac{1}{1-x^t} \le e^6. \qquad \square$$

Next, we bound the entries l_{ij} of L: If $0 < a \le \frac{2}{3}$, or $-\frac{2}{3} \le a < 0$ and $i-j$ is even, we have

$$\frac{1-a^{i-j+t}}{1-a^t} \le \frac{1}{1-|a|^t},$$

and using (11) and Lemma 4.1 we get

$$l_{ij} = \prod_{t=1}^{j-1} \frac{1-a^{i-j+t}}{1-a^t} \le \prod_{t=1}^{j-1} \frac{1}{1-|a|^t} \le e^6.$$

Otherwise, if $-\frac{2}{3} \le a < 0$ and $i-j$ is odd, we again have

$$l_{ij} = \prod_{t=1}^{j-1} \frac{1-a^{i-j+t}}{1-a^t} = \frac{1-a^{i-1}}{1-a^{i-j}} \cdot \prod_{t=1}^{j-1} \frac{1-a^{i-j-1+t}}{1-a^t} \le \frac{1+|a|^{i-1}}{1+|a|^{i-j}} \cdot e^6 \le e^6.$$

Either way, $l_{ij} \le e^6$ and $\|L\|_1 \le e^6 n$.

The entries of L^{-1} are also quotients of minors of A, as we now describe. From the LDU decomposition $A = LDU$ we get $A^{-T\#} = L^{-T\#} D^{-T\#} U^{-T\#}$. Therefore, by formula (1.31) in [2],

$$\begin{aligned}
\left(L^{-1}\right)_{ij} &= \left(L^{-T\#}\right)_{n-j+1,n-i+1} \\
&= \frac{\det A^{-T\#}([1:n-i, n-j+1], 1:n-i+1)}{\det A^{-T\#}(1:n-i+1, 1:n-i+1)} \\
&= \frac{\det A^{-1}(i:n, [j, i+1:n])}{\det A^{-1}(i:n, i:n)} \\
&= (-1)^{i+j} \cdot \frac{\det A([1:j-1, j+1:i], 1:i-1)}{\det A(1:i-1, 1:i-1)} \\
&= (-1)^{i+j} \cdot a^{\frac{1}{2}(i-j-1)(i-j)} \cdot \prod_{t=1}^{j-1} \frac{1-a^{i-j+t}}{1-a^t}. \qquad (12)
\end{aligned}$$

Similarly, $\left|(L^{-1})_{ij}\right| \leq e^6$ when $|a| \leq \frac{2}{3}$, and

$$\kappa_1(L) = \|L\|_1 \cdot \|L^{-1}\|_1 \leq e^{12} n^2;$$

i.e., L is well conditioned when $|a| \leq \frac{2}{3}$. The constant e^{12} and the factor n^2 in the previous bound are pessimistic, and the true values of $\kappa_1(L)$ are much smaller. They are shown in the following table for some values of a in 30×30 Vandermonde matrices:

a	$-2/3$	-0.5	-0.3	-0.05	0.05	0.3	0.5	$2/3$
$\kappa_1(L)$	92.12	79.25	69.83	61.50	64.16	126.98	379.12	2694.99

When $|a| > 1$ we consider the *converse* of A:

$$A^{\#} \equiv \left[A_{n-i+1,n-j+1}\right]_{i,j=1}^{n} = \left[a^{(n-i)(n-j)}\right]_{i,j=1}^{n}.$$

The matrices A and $A^{\#}$ are similar via an orthogonal similarity transformation,

$$A = JA^{\#}J,$$

where the matrix $J = [\delta_{n-i+1,j}]_{i,j=1}^{n}$ is the *reverse identity* (which is orthogonal and involutary: $J = J^T = J^{-1}$). Therefore, it suffices to compute an accurate RRD of $A^{\#}$. Consider the LDU decomposition $A^{\#} = \bar{L}\bar{D}\bar{L}^T$. The entries of $\bar{L}$ and $\bar{D}$ are quotients of minors of $A^{\#}$; thus, after some long but elementary manipulations, we get

$$\bar{d}_i = a^{(n-i)^2 - \frac{i(i-1)}{2}} \cdot \prod_{t=1}^{i-1}(a^t - 1), \tag{13}$$

$$\bar{l}_{ij} = a^{(n-1)(j-i)} \prod_{t=1}^{j-1} \frac{a^{i-j+t} - 1}{a^t - 1}. \tag{14}$$

For $|a| \geq \frac{3}{2}$, the entries $\bar{l}_{ij}$ are bounded as

$$\bar{l}_{ij} \leq \prod_{t=1}^{j-1} \frac{1}{1 - |a|^{-t}} \leq e^6.$$

For the entries of $\bar{L}^{-1}$, we obtain analogously to (12),

$$\left(\bar{L}^{-1}\right)_{ij} = (-1)^{i+j} \cdot a^{(j-i)(n-\frac{1}{2}(i-j+1))} \cdot \prod_{t=1}^{j-1} \frac{a^{i-j+t} - 1}{a^t - 1}.$$

Finally, since $n - \frac{1}{2}(i-j-1) \geq j-1$ we have

$$\left|\left(\bar{L}^{-1}\right)_{ij}\right| = a^{(j-i)(n-\frac{1}{2}(i-j+1))} \cdot \prod_{t=1}^{j-1} \frac{a^{i-j+t} - 1}{a^t - 1} \leq \prod_{t=1}^{j-1} \frac{1}{1 - |a|^{-t}} \leq e^6.$$

Again, $\kappa_1(\bar{L}) \leq e^{12}n^2$. Therefore, $\bar{L}$ is well conditioned when $|a| \geq \frac{3}{2}$, and $A = (J\bar{L})\bar{D}(J\bar{L})^T$ is an RRD of A. The true values of $\kappa_1(\bar{L})$ are much smaller than the bound—in particular, for 30×30 matrices $\kappa_1(\bar{L}) = 13.37$ for $a = \frac{3}{2}$ and $\kappa_1(\bar{L}) = 2.35$ for $a = -\frac{3}{2}$. We have observed that $\kappa_1(\bar{L})$ decreases as $|a|$ increases.

In order to guarantee high relative accuracy in each computed entry of L, $\bar{L}$, D, and $\bar{D}$, we compute all expressions $a^i - 1$ to high relative accuracy as $a^i - 1$ when $a^i < 0$ and as $(|a| - 1)(|a|^{i-1} + |a|^{i-2} + \cdots + 1)$ when $a^i > 0$.

The cost of computing factorizations with the formulas (10) and (11), or (13) and (14), is $O(n^2)$. We need n^2 flops to compute a^i for $i = 1, 2, \ldots, n^2$, and n flops to compute $\sum_{p=0}^{j} |a|^p$ for $j = 1, 2, \ldots, n$. With this, at most n extra flops are needed to compute $a^i - 1$ for $i = 1, 2, \ldots, n$. All the diagonal elements d_i, $i = 1, 2, \ldots, n$, are computed in $6n$ flops. If $i - j = k$, the $n - k$ off-diagonal elements l_{ij} are computed in $2(n - k)$ flops. Taking into account that $k = 1, 2, \ldots, n - 1$, $n^2 + O(n)$ flops are needed to compute all off-diagonal elements l_{ij}. The total cost of computing the LDL^T factorization using (10) and (11) is $2n^2 + O(n)$ flops. A similar argument shows that the total cost of computing the $\bar{L}\bar{D}\bar{L}^T$ factorization using (13) and (14) is $2n^2 + O(n)$ flops. This extremely fast performance is important in its own right, but for the purpose of computing eigenvalues and eigenvectors to high relative accuracy the cost of applying the J-orthogonal or SSVD algorithms to the RRD is $O(n^3)$, and the cost $O(n^2)$ in the RRD computation does not significantly improve the total cost.

5. Computing an RRD of a TN matrix. The matrices with all minors nonnegative are called *totally nonnegative* (*TN*). They appear in a wide range of problems and applications (see [2, 14, 17, 24, 26] and references therein). One of the most important application is to one-dimensional oscillatory problems [16].

It has been recently shown [26, 25] that many accurate computations with nonsingular $n \times n$ TN matrices are possible when these matrices are appropriately represented as products of nonnegative bidiagonal matrices:

$$A = L^{(1)} \cdot L^{(2)} \cdots L^{(n-1)} \cdot D \cdot U^{(n-1)} \cdots U^{(2)} \cdot U^{(1)}, \tag{15}$$

where D is diagonal. This decomposition was introduced in [18, 19], and it is a unique, intrinsic representation for any nonsingular TN matrix A. This *bidiagonal decomposition* will be denoted by $\mathcal{BD}(A)$. We refer to [26, section 2.2] for a detailed explanation of the structure of the factorization (15), and also for the essential relationship between this factorization and *Neville elimination*, an alternative process to Gaussian elimination that allows one to compute (15) and to check whether a matrix is TN or not.

The numerical virtues of $\mathcal{BD}(A)$ are discussed at length in [26, 25]. This decomposition reveals the TN structure of A, and its nontrivial entries accurately determine the eigenvalues, the SVD, the inverse, and other properties of a nonsingular TN matrix. Starting with the representation (15), one can perform many highly accurate matrix computations with nonsingular TN matrices [26, 25], and, in particular, the SVD of a TN matrix A can certainly be computed given (15) (see Algorithm 6.1 from [26]). The SVD is, of course, an RRD. This approach, however, relies on the convergence properties of an algorithm for computing the SVD of a bidiagonal matrix.

Our goal in this section is to design algorithms that compute an accurate RRD of a nonsingular TN matrix given its bidiagonal factorization (15) in $O(n^3)$ time by using a *finite* process and respecting the symmetry; i.e., a symmetric TN matrix will

have a symmetric RRD. This last requirement forces us to develop two algorithms: one for general TN matrices and another specifically for symmetric TN matrices.

5.1. RRD of a nonsymmetric TN matrix. Given the bidiagonal decomposition (15) of a nonsingular TN matrix A, we can accurately compute a decomposition $A = QBH^T$, where Q and H are orthogonal and B is bidiagonal, using the first part of Algorithm 6.1 from [26]. All entries of B are computed with relative errors of order ϵ, while Q and H are computed by accumulating Givens rotations with normwise errors of order ϵ, i.e., $\|Q - \widehat{Q}\|_2 = O(\epsilon)$. A similar bound holds for H. If $B = \bar{D}\bar{U}$, where $\bar{D}$ is diagonal and $\bar{U}$ is unit upper bidiagonal, then $B = \bar{D}\bar{U}$ need not be an RRD of B.

How do we compute an RRD of B? We can simply run GECP on B. Since B is acyclic (the bipartite graph of B does not have any cycles), the process of Gaussian elimination with complete pivoting will not involve any subtractions and will therefore be highly accurate (see section 6 and Algorithm 10.1 in [6]). More precisely, if P_1 and P_2^T are the permutation matrices coming from the complete pivoting strategy and $B = P_1 LDU P_2^T$, with L unit lower triangular, U unit upper triangular, and D diagonal, then all the entries of the L, D, and U factors are computed with relative errors of order ϵ.

Once we have $B = P_1 LDU P_2^T$, we obtain an RRD of A:

$$A = (QP_1L) \cdot D \cdot (UP_2^T H^T) \equiv XDY^T.$$

A direct and standard error analysis shows that the computed factors satisfy the error bounds (2). The cost of computing B is at most $\frac{16}{3}n^3 + O(n^2)$ flops [26], and forming Q and H requires not more than $6n^3 + O(n^2)$ flops. The cost of GECP on B does not exceed $\frac{2}{3}n^3 + O(n^2)$ flops and $\frac{n^3}{3}$ comparisons. Finally, the last two matrix multiplications require not more than $2n^3$ flops. The total cost does not exceed $14\,n^3 + O(n^2)$ flops and $\frac{n^3}{3}$ comparisons.

5.2. RRD of a symmetric TN matrix. The techniques of section 5.1 can certainly be used to compute an RRD of a nonsingular symmetric TN matrix given its bidiagonal decomposition. This approach does not, however, respect the symmetry of the matrix. In this subsection we present a different RRD algorithm, which does respect the symmetry.

Let the bidiagonal decomposition of a symmetric and nonsingular TN matrix A be given. Then in (15) we have $L^{(i)} = (U^{(i)})^T$. We can use the techniques of [26] to apply highly accurate Givens rotations to A and reduce A to tridiagonal form T:

$$A = QTQ^T,$$

where Q is orthogonal and $T = LDL^T$ is TN. All entries in the lower unit bidiagonal factor L and in the diagonal factor D are computed with relative errors of order ϵ, while the error in Q is $\|Q - \widehat{Q}\|_2 = O(\epsilon)$. Notice that the previous process computes $\mathcal{BD}(T)$ and Q starting from $\mathcal{BD}(A)$, and that the decomposition $T = LDL^T$ need not reveal the rank of T since L need not be well conditioned.

The remaining task in getting an accurate symmetric RRD is to compute, given $\mathcal{BD}(T)$, an accurate RRD of T by using symmetric GECP:

$$T = P\bar{L}\bar{D}\bar{L}^T P^T,$$

where P is a permutation matrix, $\bar{L}$ is unit lower triangular, and $\bar{D}$ is diagonal. Then the symmetric RRD of A is

$$A = (QP\bar{L})\bar{D}(QP\bar{L})^T.$$

We will show how to compute P and all the entries of $\bar{L}$ and $\bar{D}$ with relative errors of order ϵ. Our approach is based on two key ideas: the first is that T is positive definite, and thus the pivoting strategy in GECP will be diagonal, and the second is that the elements of $\bar{L}$ and $\bar{D}$ are signed quotients of minors of T. We will proceed in three steps as follows: (a) The bidiagonal factorization of a principal submatrix of T is accurately computed starting from $\mathcal{BD}(T)$ in Algorithm 3; (b) this is used in Algorithm 4 to compute accurate minors of T; and (c) the elements of $\bar{L}$ and $\bar{D}$ are computed as quotients of minors in Algorithm 5, together with P.

We can summarize the algorithm to compute a symmetric RRD of a nonsingular symmetric TN matrix A as follows.

ALGORITHM 2. Computing a symmetric RRD $A = XDX^T$ of a symmetric nonsingular TN matrix A given $\mathcal{BD}(A)$.

1. Apply Givens rotations as in [26, section 4.3] to compute an orthogonal matrix Q and $\mathcal{BD}(T)$ of a symmetric TN tridiagonal matrix T such that $A = QTQ^T$.
2. Compute a symmetric RRD of $T = P\bar{L}\bar{D}\bar{L}^TP^T$ using Algorithm 5.
3. Multiply to get $A = (QP\bar{L})\bar{D}(QP\bar{L})^T \equiv XDX^T$.

Step 1 requires not more than $\frac{8}{3}n^3 + O(n^2)$ flops to get $\mathcal{BD}(T)$ (see [26]) and not more than $3n^3 + O(n^2)$ additional flops to compute Q. We will see that the cost of step 2 does not exceed $14\frac{1}{3}n^3 + O(n^2)$. Finally, the cost of step 3 does not exceed n^3. The total cost of Algorithm 2 does not exceed $21n^3 + O(n^2)$ flops.

We will show that the computation of the symmetric RRD $T = P\bar{L}\bar{D}\bar{L}^TP^T$ is subtraction free. Combining this with the errors in Q, $\mathcal{BD}(T)$, and matrix multiplication, it can be easily shown that the computed RRD satisfies (2).

Once a symmetric RRD, $A = XDX^T$, of the TN matrix A is computed, the eigenvalues and eigenvectors of A can be accurately computed by using the one-sided Jacobi algorithm to get the singular values and left singular vectors of $XD^{1/2}$ [10, section 5.4.3], [9]. The Jacobi rotations in this procedure have to be applied on the left, and the whole process respects the symmetry. The techniques introduced in [26] allow us to develop another symmetric method to compute accurate eigenvalues of a nonsingular symmetric TN matrix A: First, step 1 of Algorithm 2 is performed to get $T = LDL^T$; next, the differential quotient-difference algorithm with shifts (dqds) [13] is applied on the Cholesky factor $LD^{1/2}$ to compute its accurate singular values. This approach does not use RRDs.

5.2.1. The bidiagonal decomposition of a principal submatrix of a TN tridiagonal symmetric matrix. Let T be a nonsingular symmetric TN tridiagonal matrix[3] and S be a principal submatrix of T. The purpose of this section is to accurately compute $\mathcal{BD}(S)$ given $\mathcal{BD}(T)$.

Consider first the simple special case when the principal submatrix S is obtained

[3]The results of this section remain valid for positive definite tridiagonal matrices because a positive definite tridiagonal matrix is TN if and only if its off-diagonal elements are nonnegative [16, p. 81]. Therefore, any positive definite tridiagonal matrix is similar to a TN matrix through a diagonal similarity transformation with elements ± 1.

by erasing the ith row and the ith column of T:

$$T = \begin{bmatrix} t_{11} & t_{12} & & & \\ t_{21} & \ddots & \ddots & & \\ & \ddots & \ddots & t_{n-1,n} & \\ & & t_{n,n-1} & t_{nn} & \end{bmatrix};$$

$$S = T([1:i-1, i+1:n], [1:i-1, i+1:n])$$

$$= \left[\begin{array}{ccccc|ccccc} t_{11} & t_{12} & & & & & & & & \\ t_{21} & \ddots & \ddots & & & & & & & \\ & \ddots & \ddots & t_{i-2,i-1} & & & & & & \\ & & t_{i-1,i-2} & t_{i-1,i-1} & & & & & & \\ \hline & & & & & t_{i+1,i+1} & t_{i+1,i+2} & & & \\ & & & & & t_{i+2,i+1} & \ddots & \ddots & & \\ & & & & & & \ddots & \ddots & t_{n-1,n} & \\ & & & & & & & t_{n,n-1} & t_{nn} & \end{array}\right].$$

Once we figure out how to compute $\mathcal{BD}(S)$ from $\mathcal{BD}(T)$, we can proceed by induction and erase other rows and columns of S to obtain the bidiagonal decomposition of any principal submatrix of T.

Since the process of Neville elimination of S and T does not differ for the first $i-1$ rows and columns, we have $\mathcal{BD}(S(1:i-1, 1:i-1)) = \mathcal{BD}(T(1:i-1, 1:i-1))$, and we need only compute $\mathcal{BD}(S(i+1:n, i+1:n))$. Therefore, we may assume that $i = 1$ without any loss of generality.

Let $\mathcal{BD}(T)$ and $\mathcal{BD}(S)$ be given as

$$T = LDL^T \quad \text{and} \quad S = T(2:n, 2:n) = \bar{L}\bar{D}\bar{L}^T,$$

where $D = \operatorname{diag}(d_i)_{i=1}^n$, $\bar{D} = \operatorname{diag}(\bar{d}_i)_{i=2}^n$, and the unit lower bidiagonal matrices L and $\bar{L}$ have off-diagonal elements l_i, $i = 1, 2, \ldots, n-1$, and $\bar{l}_i$, $i = 2, 3, \ldots, n-1$, respectively. From $T = LDL^T$ we have

$$t_{11} = d_1; \;\; t_{ii} = l_{i-1}^2 d_{i-1} + d_i; \;\; t_{i-1,i} = l_{i-1} d_{i-1}, \quad i = 2, 3, \ldots, n, \tag{16}$$

and from $T(2:n, 2:n) = \bar{L}\bar{D}\bar{L}^T$ we get

$$t_{22} = \bar{d}_2; \;\; t_{ii} = \bar{l}_{i-1}^2 \bar{d}_{i-1} + \bar{d}_i; \;\; t_{i-1,i} = \bar{l}_{i-1}\bar{d}_{i-1}, \;\; i = 3, 4, \ldots, n. \tag{17}$$

By comparing (16) and (17), we obtain

$$\begin{aligned} \bar{d}_2 &= l_1^2 d_1 + d_2, \\ \bar{d}_i &= d_i + l_{i-1}^2 d_{i-1} - \bar{l}_{i-1}^2 \bar{d}_{i-1}, \quad i = 3, 4, \ldots, n, \\ \bar{l}_i \bar{d}_i &= l_i d_i, \quad i = 2, 3, \ldots, n-1. \end{aligned} \tag{18}$$

We introduce auxiliary variables $z_i \equiv \bar{d}_i - d_i$ and get rid of the subtraction in (18):

$$\begin{aligned} z_2 &= l_1^2 d_1, \\ \bar{d}_2 &= z_2 + d_2, \\ \bar{l}_i &= l_i d_i / \bar{d}_i, \quad i = 2, \ldots, n-1, \\ z_{i+1} &= \bar{d}_{i+1} - d_{i+1} = l_i^2 d_i - \bar{l}_i^2 \bar{d}_i = (\bar{d}_i - d_i) l_i^2 d_i / \bar{d}_i = l_i \bar{l}_i z_i, \qquad i = 2, \ldots, n-1, \\ \bar{d}_{i+1} &= z_{i+1} + d_{i+1}, \quad i = 2, \ldots, n-1. \end{aligned} \tag{19}$$

The iterations (19) need only be performed for those $i \geq 2$ for which $l_i \neq 0$. These iterations therefore cost $5(j-1)$, where $j < n$ is the smallest index such that $l_j = 0$ (or $j = n$ if $l_k \neq 0$ for $k = 1, 2, \ldots, n-1$). In the general case, when we remove the ith row and the ith column of T, the cost is $5(j-i)$, where $j \geq i$ is defined as above.

We now implement the recurrences (19).

ALGORITHM 3. Let $T = LDL^T$ be a nonsingular symmetric TN tridiagonal matrix, where $D = \operatorname{diag}(d_i)_{i=1}^n$, $d_i > 0$, $i = 1, 2, \ldots, n$, and L is a unit lower bidiagonal matrix with off-diagonal entries $l_i \geq 0, i = 1, 2, \ldots, n-1$. Let $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_r\}$, $1 \leq \alpha_1 < \alpha_2 < \cdots < \alpha_r \leq n$ be a subset of indices. Given the vectors $d = (d_1, d_2, \ldots, d_n)$ and $l = (l_1, \ldots, l_{n-1})$, the following subtraction-free algorithm computes the decomposition $T(\alpha, \alpha) = \bar{L}\bar{D}\bar{L}^T$ in at most $5r$ time:

```
function [d̄, l̄] = TNTridiagSubmatrix(d, l, α)
n = length(d)
d̄ = d; l̄ = l; l̄_n = 0
Let β be the complement of α in the set {1, 2, ..., n}
(In MATLAB notation: β = [1 : n]; β(α) = 0; β = β(β > 0))
for k = length(β) : −1 : 1
   if β_k < n
      z = d_{β_k} l²_{β_k}
      j = β_k + 1
      d̄_j = z + d_j
      while l̄_j ≠ 0
         l̄_j = l_j d_j / d̄_j
         z = l_j l̄_j z
         d̄_{j+1} = z + d_{j+1}
         j = j + 1
      end
   end
   l̄_{β_k − 1} = 0; l̄_{β_k} = 0
end
d̄ = d̄(α); l̄ = l̄(α); l̄ = l̄(1 : (r − 1))
```

5.2.2. A minor of a TN tridiagonal symmetric matrix. Next we consider the problem of accurately computing the value of any minor of a nonsingular symmetric TN tridiagonal matrix T:

$$T(\alpha, \beta) = T([i_1, \ldots, i_k], [j_1, \ldots, j_k]),$$

where $\alpha = [i_1, i_2, \ldots, i_k]$, $1 \leq i_1 < i_2 < \ldots < i_k \leq n$, and $\beta = [j_1, j_2, \ldots, j_k]$, $1 \leq j_1 < j_2 < \ldots < j_k \leq n$.

Let $1 \leq k_1 < k_2 < \cdots < k_r \leq k$ be all indices such that $i_{k_s} \neq j_{k_s}$, $s = 1, 2, \ldots, r$, and let $\gamma = \{i_1, i_2, \ldots, i_k\} \backslash \{i_{k_1}, \ldots, i_{k_r}\}$. Then [16, p. 80]

$$\det T(\alpha, \beta) = \det T(\gamma, \gamma) t_{i_{k_1} j_{k_1}} \cdots t_{i_{k_r} j_{k_r}}. \tag{20}$$

The minor $\det T(\gamma, \gamma)$ can be computed by first computing the bidiagonal decomposition of $T(\gamma, \gamma)$ using Algorithm 3 (then $\det T(\gamma, \gamma) = \bar{d}_1 \bar{d}_2 \cdots \bar{d}_{k-r}$). Any entry $t_{i_{k_s} j_{k_s}}$, $i_{k_s} \neq j_{k_s}$, equals either zero, $t_{m,m+1}$, or $t_{m+1,m}$. The latter two are easily computed from $T = LDL^T$: $t_{m,m+1} = t_{m+1,m} = d_m l_m$. The total cost of computing any minor $\det T(\alpha, \beta)$ following this procedure does not exceed $6k$ flops.

Remark 1. A set of indices $\mathtt{z} \subset \{1, 2, \ldots, n\}$ can be sorted in increasing order in $4n$ time by using the following MATLAB commands:

```
x=1:n; x(z)=0; y=1:n; z=y(x==0);
```

therefore, we can sort the index sets in $T(\alpha, \beta)$ in $8n$ time and allow index sets in arbitrary order in Algorithm 4 below.

ALGORITHM 4 (minor of a TN tridiagonal matrix). Let $T = LDL^T$ be a nonsingular symmetric TN tridiagonal matrix with notation as in Algorithm 3. Given the vectors d and l, and two sets of indices α and β, the following subtraction-free algorithm computes $|\det T(\alpha, \beta)|$ to high relative accuracy in at most $14n$ time:

```
function f = TNTridiagMinor(d, l, α, β)
    ... first sort α and β in increasing order (see Remark 1 above)...
f = 1; γ = [ ]
for i = 1 : length(α)
   if α_i = β_i
      γ = [γ, α_i]
   elseif |α_i − β_i| = 1
      f = f d_s l_s,  where s = min(α_i, β_i)
   else
      f = 0; return
   end
end
[d̄, l̄] = TNTridiagSubmatrix(d, l, γ)
f = f d̄_1 d̄_2 ··· d̄_s,  where s = length(γ)
```

5.2.3. Computing an RRD of a TN tridiagonal symmetric matrix. In this section we present an $O(n^3)$ algorithm which, given the factorization $T = LDL^T$ of a nonsingular symmetric TN tridiagonal matrix T, computes an accurate, symmetric RRD of T. The RRD in question is the LDU decomposition of T resulting from GECP, with L (resp., U) being a unit lower (resp., upper) triangular matrix. We compute each entry of this LDU decomposition as a quotient of minors of T. We compute each minor of T accurately using Algorithm 4.

Since T is positive definite, the pivoting in GECP will be diagonal. The pivot order is determined by comparing the diagonal entries in the Schur complements; if $\gamma = [\gamma_1, \gamma_2, \ldots, \gamma_k]$ is the current pivot order at step k, and $\alpha = \{1, 2, \ldots, n\} \backslash \gamma$, then the diagonal entries of the kth Schur complement are[4]

$$\frac{\det T([\gamma, \alpha_j], [\gamma, \alpha_j])}{\det T(\gamma, \gamma)}, \qquad j = 1, 2, \ldots, n-k. \tag{21}$$

We need only compare the numerators in (21) and we compute those using Algorithm 4.

[4]These expressions for the entries of the Schur complements are valid if for each step of Gaussian elimination the row and column containing the chosen pivot are moved to the first positions in the corresponding Schur complement and the rows and columns between the first and the ones containing the pivot are displaced down one position. This is not the usual implementation of pivoting in Gaussian elimination, which simply interchanges the first row and the first column with the pivot row and the pivot column, respectively [21, 22]. Obviously both implementations produce similar bounds on the elements of L and U, and, therefore, they are equivalent from the point of view of computing RRDs.

Once we obtain the pivot order γ, the entries of the LDU decomposition $T = P\bar{L}\bar{D}\bar{L}^T P^T$ resulting from GECP are computed as

$$\bar{D}_{ii} = \frac{\det T(\gamma(1:i), \gamma(1:i))}{\det T(\gamma(1:i-1), \gamma(1:i-1))}; \tag{22}$$

$$\bar{L}_{ji} = \frac{\det T(\gamma(1:i), \gamma([1:i-1, j]))}{\det T(\gamma(1:i), \gamma(1:i))}, \qquad j > i, \tag{23}$$

with each minor in (22) and (23) computed using Algorithm 4.

The sign of the minor $\det T(\gamma(1:i), \gamma([1:i-1, j]))$ equals $\mathrm{sgn}(\gamma(1:i)) \cdot \mathrm{sgn}(\gamma([1:i-1, j]))$. Here $\mathrm{sgn}(\delta)$ is the *sign* of $[\delta_1, \delta_2, \dots]$ as a permutation of the ordered set $\{\delta_1, \delta_2, \dots\}$, defined as $\mathrm{sgn}(\delta) \equiv (-1)^t$, where $t \equiv \#\{(k,l) | k < l \text{ and } \delta_k > \delta_l\}$; i.e., t is the minimum number of transpositions necessary to sort the elements of δ in increasing order. The first $i-1$ entries of $\gamma(1:i)$ and $\gamma([1:i-1, j])$ coincide; therefore the sign of $\det T(\gamma(1:i), \gamma([1:i-1, j]))$ equals $(-1)^s$, $s = \sum_{k=1}^{i-1} \mathtt{xor}(\gamma_i < \gamma_k, \gamma_j < \gamma_k)$.

ALGORITHM 5 (GECP on a TN tridiagonal matrix). Let $T = LDL^T$ be a nonsingular symmetric TN tridiagonal matrix. Given the vectors d and l (defined in Algorithm 3), the following subtraction-free algorithm computes the decomposition of $T = P\bar{L}\bar{D}\bar{L}^T P^T$ resulting from Gaussian elimination with complete pivoting. Every entry of $\bar{D}$ and $\bar{L}$ is computed to high relative accuracy, and the total cost does not exceed $14\frac{1}{3}n^3 + O(n^2)$.

```
function [P, L̄, D̄] = TNTridiagGECP(d, l)
n = length(d)
L̄ = eye(n); P = eye(n); D̄ = eye(n);
α = 1 : n; γ = [ ]
            ... First, determine the pivot order...
for i = 1 : n
   for j = 1 : n − i + 1
      z_j = TNTridiagMinor(d, l, [γ, α_j], [γ, α_j])
   end
   Let m be such that z_m = max_{1≤j≤n−i+1} z_j
   γ_i = α_m
   α = α([1 : m − 1, m + 1 : n − i + 1])
   t_i = z_m
end
            ... Next, compute the entries of D̄ and L̄ using (22) and (23)...
for i = 1 : n
   D̄_ii = t_i/t_{i−1}   ( ... assume t_0 = 1)
   for j = i + 1 : n
               ... Compute the sign of L̄_ji ...
      s = 1
      for k = 1 : i − 1
         s = s · (−1)^xor(γ_i < γ_k, γ_j < γ_k)
      end
      L̄_ji = s · TNTridiagMinor(d, l, γ(1 : i), γ([1 : i − 1, j]))/t_i
   end
end
P = P(:, γ)
```

6. Numerical experiments. We performed extensive numerical tests and confirmed the accuracy and cost of our algorithms. More precisely, we combined Algorithm 1 and the one-sided J-orthogonal algorithm [33, Algorithm 3.3.1, page 66] to compute, preserving the symmetry, accurate eigenvalues and eigenvectors of symmetric diagonally scaled Cauchy matrices with different dimensions and several distributions of random Cauchy parameters. The output was compared with that of another $O(n^3)$ accurate algorithm (nonsymmetric RRD computed as in [5] combined with the SSVD algorithm from [11]), and also with the output from the MATLAB `eig` function in variable precision arithmetic (with precision set to $\log_{10}\kappa_2(C)+20$ decimal digits, guaranteeing at least 16 correct significant digits in each eigenvalue). The output of all three algorithms agreed to at least 14 digits for all the eigenvalues, including the ones with tiniest absolute values. The computed eigenvectors also satisfied the bounds (1). Most test matrices had condition numbers well in excess of 10^{16}, so conventional eigenvalue algorithms (e.g., the MATLAB function `eig` in double [23] precision) failed to get any correct digits in the eigenvalues with tiniest absolute values and in the direction of the eigenvectors corresponding to these eigenvalues (when at least two tiny eigenvalues λ_i such that $|\lambda_i| \le 10^{-16}\|C\|_2$ were present). We performed similar tests on symmetric Vandermonde matrices for several dimensions and choices of the parameter a, and also for symmetric TN matrices. In the case of symmetric Vandermonde matrices, we also tested matrices with $\frac{2}{3} < |a| < \frac{3}{2}$ and verified that the factorizations obtained with the approach in section 4 are not RRDs when $|a|$ is close to one ($\kappa_2(L) \to 2^n$ as $|a| \to 1$). For these matrices, eigenvalues and eigenvectors to high relative accuracy can be obtained, at present, only through the nonsymmetric procedure by first computing a nonsymmetric RRD as in [5] and then applying the SSVD algorithm from [11].

We present in detail one of our tests. We consider the 20×20 symmetric Vandermonde matrix A with $a = \frac{1}{2}$; see (9). The condition number of A is $\kappa_2(A) \approx 3.5 \cdot 10^{53}$. We compute its eigenvalues using the following algorithms:

- Algorithm A: The MATLAB `eig` function with 75-digit arithmetic.
- Algorithm B: Compute a symmetric RRD using the formulas in section 4 followed by the J-orthogonal algorithm [33, Algorithm 3.3.1, page 66].
- Algorithm C: Compute a nonsymmetric RRD as in [5] followed by the SSVD algorithm of [11].
- Algorithm D: The MATLAB `eig` function in double [23] precision arithmetic.

The output of Algorithms A, B, and C agreed to at least 14 digits, so we plotted only the output of Algorithms B and D in Figure 6.1. Since $\kappa_2(A) \approx 3.5 \cdot 10^{53}$, Algorithm A computed all eigenvalues with at least 16 significant decimal digits of accuracy. Algorithms B and C guarantee high relative accuracy for the computed eigenvalues. The results from those algorithms agreed with the ones from Algorithm A to at least 14 digits. In contrast, the traditional Algorithm D returned only the eigenvalues of largest absolute value accurately, with the accuracy gradually decreasing until the eigenvalues with magnitude smaller than $O(\epsilon)||A||_2$ were computed with no correct digits at all.

Appendix. Rounding error analysis for diagonally scaled Cauchy matrices. Theorem 3.1 is proved in this appendix in a more general setting. The error analysis we present remains valid when the Bunch–Parlett method is applied on any matrix for which it is possible to compute the entries of its Schur complements with relative errors bounded by $k\epsilon/(1-k\epsilon)$, where k is an integer positive number and ϵ is the machine precision. For scaled Cauchy matrices, $k = 8n$ according to (7).

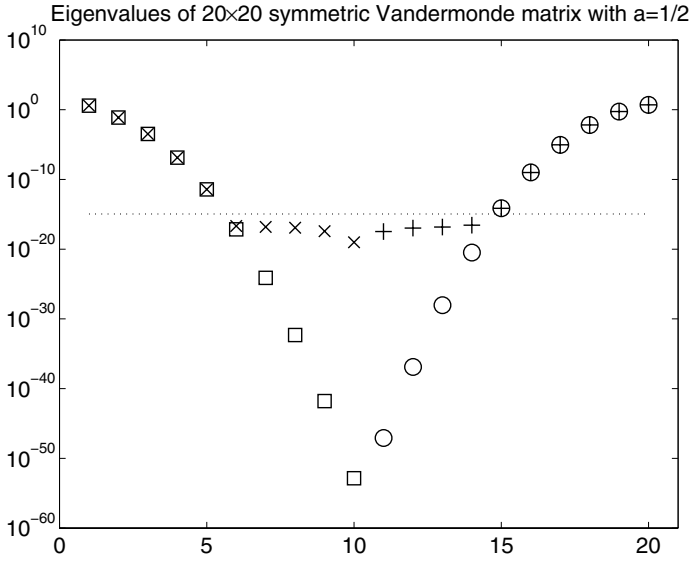


FIG. 6.1. *Plots of the absolute values of the eigenvalues of the* 20×20 *symmetric Vandermonde matrix with* $a = \frac{1}{2}$. *The "□" and "◯" symbols represent, respectively, the negative and positive eigenvalues computed with an accurate algorithm. The "×" and "+" symbols represent, respectively, the negative and positive eigenvalues computed by Algorithm D (implemented as the MATLAB function* `eig` *in double precision arithmetic). Data below the dotted line may be inaccurate for Algorithm D.*

We use the conventional error model for floating point arithmetic [22, section 2.2]:

$$\texttt{fl}(a \odot b) = (a \odot b)(1 + \delta),$$

where a and b are real floating point numbers, $\odot \in \{+, -, \times, /\}$, and $|\delta| \le \epsilon$. Moreover, we assume that neither overflow nor underflow occurs. We also use the following notation introduced in [22, Chapter 3]: θ_q is any number such that

$$|\theta_q| \le \frac{q\epsilon}{1 - q\epsilon} \equiv \gamma_q. \tag{24}$$

Moreover, the results in [22, Lemma 3.3] will be frequently used throughout this section without being explicitly referred to. We will assume that $0 < \gamma_q$ for all the symbols γ_q appearing in this section.

In what follows, α is the parameter used in the Bunch–Parlett pivoting strategy to decide whether a 1×1 or 2×2 pivot is selected (see Algorithm 1). We present the error bounds in this section depending on α, where $0 < \alpha < 1$. Thus values different from the classical one, $\alpha = (1 + \sqrt{17})/8$, are also considered.

A.1. Auxiliary results on the Jacobi method. Let us write the Jacobi procedure [21] to orthogonally diagonalize a 2×2 real symmetric matrix as a matrix factorization. The following equation holds:

$$\begin{bmatrix} a & c \\ c & b \end{bmatrix} = \begin{bmatrix} cs & sn \\ -sn & cs \end{bmatrix} \begin{bmatrix} a - c\,t & 0 \\ 0 & b + c\,t \end{bmatrix} \begin{bmatrix} cs & -sn \\ sn & cs \end{bmatrix}, \tag{25}$$

where

$$\zeta = \frac{b-a}{2c}, \qquad t = \frac{\operatorname{sign}(\zeta)}{|\zeta| + \sqrt{1+\zeta^2}}, \tag{26}$$

$$cs = \frac{1}{\sqrt{1+t^2}}, \qquad sn = cs \cdot t, \tag{27}$$

and $\operatorname{sign}(0) = 1$.

In general, disastrous cancellations may appear in the Jacobi procedure above, and the eigenvalues computed in floating point arithmetic may be inaccurate. However, it is well known that the Jacobi procedure is backward stable because only orthogonal matrices are involved. Theorem A.1 below shows this, providing precise error bounds that we will use in the detailed error analysis of the next subsections. The Jacobi method computes accurate eigenvalues for well-conditioned matrices because it is backward stable. We will see that this is the case for the 2×2 pivots selected by the Bunch–Parlett pivoting strategy.

THEOREM A.1. *Let*

$$\widetilde{A} = \begin{bmatrix} \tilde{a} & \tilde{c} \\ \tilde{c} & \tilde{b} \end{bmatrix}$$

be a matrix of real floating point numbers. Let us apply to $\widetilde{A}$ the Jacobi procedure (25) *in floating point arithmetic with machine precision ϵ. Let $\widetilde{cs}, \widetilde{sn}, \tilde{\lambda}_1 = \tilde{a} - \tilde{c}\,\tilde{t}$, and $\tilde{\lambda}_2 = \tilde{b} + \tilde{c}\,\tilde{t}$ be the exact magnitudes for $\widetilde{A}$, and let $\widehat{cs}, \widehat{sn}, \hat{\lambda}_1$, and $\hat{\lambda}_2$ be the corresponding computed counterparts. Then*

1. $\widehat{cs} = \widetilde{cs}\,(1+\theta_{113})$.
2. $\widehat{sn} = \widetilde{sn}\,(1+\theta_{141})$.
3. $\hat{\lambda}_1 = \tilde{\lambda}_1 + e_1$ *with* $|e_1| \le (|\tilde{a}| + |\tilde{c}\,\tilde{t}|)\gamma_{29}$.
4. $\hat{\lambda}_2 = \tilde{\lambda}_2 + e_2$ *with* $|e_2| \le (|\tilde{b}| + |\tilde{c}\,\tilde{t}|)\gamma_{29}$.

Moreover, the computed eigendecomposition

$$\begin{bmatrix} \widehat{cs} & \widehat{sn} \\ -\widehat{sn} & \widehat{cs} \end{bmatrix} \begin{bmatrix} \hat{\lambda}_1 & 0 \\ 0 & \hat{\lambda}_2 \end{bmatrix} \begin{bmatrix} \widehat{cs} & -\widehat{sn} \\ \widehat{sn} & \widehat{cs} \end{bmatrix}$$

is nearly the exact eigendecomposition of $\widetilde{A} + E$; more precisely,

$$\widetilde{A} + E = \begin{bmatrix} \widetilde{cs} & \widetilde{sn} \\ -\widetilde{sn} & \widetilde{cs} \end{bmatrix} \begin{bmatrix} \hat{\lambda}_1 & 0 \\ 0 & \hat{\lambda}_2 \end{bmatrix} \begin{bmatrix} \widetilde{cs} & -\widetilde{sn} \\ \widetilde{sn} & \widetilde{cs} \end{bmatrix},$$

where $\|E\|_2 \le \sqrt{2}\,\gamma_{29}\|\widetilde{A}\|_F \le 2\,\gamma_{29}\,\|\widetilde{A}\|_2$.

Proof. The bounds for $\widehat{cs}$, $\widehat{sn}$, $\hat{\lambda}_1$, and $\hat{\lambda}_2$ follow from a direct application of Lemmas 3.1 and 3.3 in [22]. For the backward error bound, notice that

$$E = \begin{bmatrix} \widetilde{cs} & \widetilde{sn} \\ -\widetilde{sn} & \widetilde{cs} \end{bmatrix} \begin{bmatrix} e_1 & 0 \\ 0 & e_2 \end{bmatrix} \begin{bmatrix} \widetilde{cs} & -\widetilde{sn} \\ \widetilde{sn} & \widetilde{cs} \end{bmatrix}.$$

Then $\|E\|_2 = \max\{|e_1|, |e_2|\} \le \gamma_{29} \max\{|\tilde{a}| + |\tilde{c}\,\tilde{t}|, |\tilde{b}| + |\tilde{c}\,\tilde{t}|\} \le \gamma_{29} \max\{|\tilde{a}| + |\tilde{c}|, |\tilde{b}| + |\tilde{c}|\} \le \sqrt{2}\,\gamma_{29}\|\widetilde{A}\|_F$. ∎

A.2. Properties of 2×2 Bunch–Parlett pivots. The 2×2 pivots selected by the Bunch–Parlett complete pivoting strategy are very well conditioned symmetric indefinite matrices. The next lemma quantifies this fact.

LEMMA A.2. *Let H be a real symmetric 2×2 matrix such that $\alpha\,|h_{21}| > \max\{|h_{11}|, |h_{22}|\}$, where $0 < \alpha < 1$. Then the spectral condition number, $\kappa_2(H)$, of H is bounded as*

$$\kappa_2(H) < \frac{1+\alpha}{1-\alpha}.$$

This bound cannot be improved. In particular, if $\alpha = 0.6404$, then $\kappa_2(H) < 4.6$.

Proof. Let us write the matrix H as

$$H = \begin{bmatrix} 0 & h_{21} \\ h_{21} & 0 \end{bmatrix} + \begin{bmatrix} h_{11} & 0 \\ 0 & h_{22} \end{bmatrix} \equiv H_0 + H_1.$$

The singular values of H_0 are both equal to $|h_{21}|$. Then using Weyl's perturbation theorem for singular values (see, for instance, [10, Corollary 5.1]), we get

$$\kappa_2(H) \leq \frac{|h_{21}| + \|H_1\|_2}{|h_{21}| - \|H_1\|_2} < \frac{|h_{21}| + \alpha|h_{21}|}{|h_{21}| - \alpha|h_{21}|} = \frac{1+\alpha}{1-\alpha}.$$

The bound cannot be improved because the matrix $H = \left[\begin{smallmatrix} \alpha & 1 \\ 1 & \alpha \end{smallmatrix}\right]$ has $\kappa_2(H) = \frac{1+\alpha}{1-\alpha}$. ∎

The entries of the eigenvectors of the 2×2 pivots selected by the Bunch–Parlett strategy are bounded below by $1/3$. This means that small normwise variations in the eigenvectors imply small variations in the components.

LEMMA A.3. *Let H be a real symmetric 2×2 matrix such that $\alpha\,|h_{21}| > \max\{|h_{11}|, |h_{22}|\}$, where $0 < \alpha < 1$. Let $\left[\begin{smallmatrix} cs & sn \\ -sn & cs \end{smallmatrix}\right]$ be the orthogonal eigenvector matrix of H; then*

$$\frac{1}{\sqrt{2}} \leq cs \leq \frac{\alpha + \sqrt{1+\alpha^2}}{\sqrt{1 + \left(\alpha + \sqrt{1+\alpha^2}\right)^2}},$$

$$\frac{1}{\sqrt{1 + \left(\alpha + \sqrt{1+\alpha^2}\right)^2}} \leq sn \leq \frac{1}{\sqrt{2}}.$$

In particular, if $\alpha = 0.6404$, then $0.47 \leq sn$ and $cs \leq 0.88$. The following simple lower bound for sn is valid for any α: $1/3 < sn$.

Proof. From (26), $|\zeta| \leq \alpha$ and $1/(\alpha + \sqrt{1+\alpha^2}) \leq |t| \leq 1$. Combining these bounds with (27), the lemma is proved. ∎

A.3. Forward errors in RRDs. The entries of the Schur complements of diagonally scaled Cauchy matrices are computed by (7) with relative errors less than γ_{8n}. In this section we assume that the entries of the Schur complements are computed with relative errors less than γ_k; thus the error analysis remains valid for other cases.

A nagging problem will arise in the analysis: the computed 2×2 pivots fulfill the conditions of Bunch and Parlett, i.e., $\alpha\,|\hat{h}_{21}| > \max\{|\hat{h}_{11}|, |\hat{h}_{22}|\}$, but the exact pivots may not. This justifies the following lemma.

LEMMA A.4. *Let*

$$\widetilde{A} = \begin{bmatrix} a(1+\beta_a) & c(1+\beta_c) \\ c(1+\beta_c) & b(1+\beta_b) \end{bmatrix} \equiv \begin{bmatrix} \tilde{a} & \tilde{c} \\ \tilde{c} & \tilde{b} \end{bmatrix}$$

be a matrix of real floating point numbers, where $\max\{|\beta_a|, |\beta_b|, |\beta_c|\} \leq \gamma_k$, *and* $\alpha\,|\tilde{c}| > \max\{|\tilde{a}|, |\tilde{b}|\}$, *with* $0 < \alpha < 1$. *Denote* $A \equiv \left[\begin{smallmatrix} a & c \\ c & b \end{smallmatrix}\right]$. *If*

$$4\sqrt{2}\,\frac{1+\alpha}{1-\alpha}\,\gamma_k \leq 1, \tag{28}$$

then

$$\kappa_2(A) \leq 2\,\frac{1+\alpha}{1-\alpha}. \tag{29}$$

Proof. Notice that

$$\widetilde{A} = A + E_1 \qquad \text{with} \qquad \|E_1\|_F \leq \gamma_k \|A\|_F \leq \sqrt{2}\,\gamma_k \|A\|_2. \tag{30}$$

Let $\sigma_1 \geq \sigma_2$ and $\tilde{\sigma}_1 \geq \tilde{\sigma}_2$ be the singular values of A and $\widetilde{A}$, respectively. Now Corollary 5.1 from [10] implies

$$\kappa_2(\widetilde{A}) = \frac{\tilde{\sigma}_1}{\tilde{\sigma}_2} \geq \frac{\sigma_1 - \sqrt{2}\,\gamma_k \|A\|_2}{\sigma_2 + \sqrt{2}\,\gamma_k \|A\|_2} = \kappa_2(A)\,\frac{1 - \sqrt{2}\,\gamma_k}{1 + \sqrt{2}\,\gamma_k\,\kappa_2(A)}.$$

From this we get

$$\kappa_2(A) \leq \frac{\kappa_2(\widetilde{A})}{1 - 2\sqrt{2}\,\gamma_k\,\kappa_2(\widetilde{A})}.$$

The result follows from (28) and Lemma A.2, which implies

$$\kappa_2(\widetilde{A}) \leq (1+\alpha)/(1-\alpha). \qquad \square$$

Obviously the rigorous factor 2 in (29) is pessimistic, and in practice $\kappa_2(A) \approx \kappa_2(\widetilde{A}) \leq (1+\alpha)/(1-\alpha)$. However, at the cost of the nonessential factor 2, Lemma A.4 allows us to get rigorous error bounds instead of first-order error bounds. In particular, we can prove the following lemma.

LEMMA A.5. *Let*

$$\widetilde{A} \equiv \begin{bmatrix} \tilde{a} & \tilde{c} \\ \tilde{c} & \tilde{b} \end{bmatrix} = \begin{bmatrix} a(1+\beta_a) & c(1+\beta_c) \\ c(1+\beta_c) & b(1+\beta_b) \end{bmatrix}$$

be a matrix of real floating point numbers, where $\max\{|\beta_a|, |\beta_b|, |\beta_c|\} \leq \gamma_k$, *and* $\alpha\,|\tilde{c}| > \max\{|\tilde{a}|, |\tilde{b}|\}$, *with* $0 < \alpha < 1$. *Denote* $A \equiv \left[\begin{smallmatrix} a & c \\ c & b \end{smallmatrix}\right]$. *Let the eigenvalues of* A *be* $\lambda_1 \geq \lambda_2$; v_1 *and* v_2 *be the corresponding orthonormal eigenvectors; and* cs *and* sn *be the components of the eigenvectors, i.e.,* $v_1 = [cs\,,\, -sn]^T$ *and* $v_2 = [sn\,,\, cs]^T$ *or vice versa. Let* $\hat{\lambda}_1, \hat{\lambda}_2, \hat{v}_1, \hat{v}_2, \widehat{cs}$, *and* $\widehat{sn}$ *be their corresponding computed counterparts by applying the Jacobi process in* (25)–(27) *to* $\widetilde{A}$ *in floating point arithmetic with machine precision* ϵ. *If*

$$4\sqrt{2}\,\frac{1+\alpha}{1-\alpha}\,\gamma_{k+29} \leq 1 \qquad \textit{and} \qquad \gamma_{141+48k} \leq 1, \tag{31}$$

then

1.

$$\frac{|\hat{\lambda}_i - \lambda_i|}{|\lambda_i|} \leq 4\,\frac{1+\alpha}{1-\alpha}\,\gamma_{k+29}, \qquad i = 1, 2; \tag{32}$$

2.

$$\|\hat{v}_i - v_i\|_2 \leq \gamma_{4k+141}, \qquad i = 1, 2; \tag{33}$$

3.

$$\widehat{cs} = cs\,(1 + \theta_{16k+113}) \qquad and \qquad \widehat{sn} = sn\,(1 + \theta_{48k+141}). \tag{34}$$

We have chosen to get error bounds for cs and sn that do not depend on α. At the cost of complicating the bounds, it is possible to get sharper bounds depending on α. Moreover, we have frequently overestimated the bounds to get simpler expressions. It is well known that the precise value of the constants appearing in roundoff error bounds are, in any case, pessimistic.

Proof of Lemma A.5. According to Theorem A.1, $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are the exact eigenvalues of

$$\widetilde{A} + E \qquad \text{with} \qquad \|E\|_2 \leq \sqrt{2}\gamma_{29}\|\widetilde{A}\|_F,$$

while $\hat{v}_1$ and $\hat{v}_2$ differ from the exact eigenvectors of $\widetilde{A} + E$ by only small relative changes in each component. Therefore, by taking into account (30), we get that $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are the exact eigenvalues of

$$A + E_2 \equiv A + E_1 + E \qquad \text{with} \qquad \|E_2\|_2 \leq \sqrt{2}\,\gamma_{k+29}\|A\|_F \leq 2\,\gamma_{k+29}\|A\|_2, \tag{35}$$

and $\hat{v}_1$, $\hat{v}_2$ are small relative componentwise perturbations of the eigenvectors of $A + E_2$. Weyl's perturbation theorem for eigenvalues implies that $|\hat{\lambda}_i - \lambda_i| \leq \|E_2\|_2 \leq 2\,\gamma_{k+29}\|A\|_2$. By using (29) we obtain (32):

$$\frac{|\hat{\lambda}_i - \lambda_i|}{|\lambda_i|} \leq 2\,\gamma_{k+29}\kappa_2(A) \leq 4\,\frac{1+\alpha}{1-\alpha}\,\gamma_{k+29}, \qquad i = 1, 2.$$

Let us focus on the eigenvectors. In the first place, we are going to relate the eigenvectors v_1 and v_2 of A to the eigenvectors $\widetilde{v}_1$ and $\widetilde{v}_2$ of $\widetilde{A}$. Notice that according to Theorem A.1, the components of $\hat{v}_1$ and $\hat{v}_2$ are small relative perturbations of the components of $\widetilde{v}_1$ and $\widetilde{v}_2$. Therefore, once $\widetilde{v}_1$ and $\widetilde{v}_2$ are related to v_1 and v_2, the difference between $\hat{v}_i$ and v_i, $i = 1, 2$, is easily obtained. Let $\theta(v_i, \widetilde{v}_i)$ be the acute angle between v_i and $\widetilde{v}_i$. Then [10, Theorem 5.4] and (30) lead to

$$\frac{1}{2}\sin 2\,\theta(v_i, \widetilde{v}_i) \leq \frac{\sqrt{2}\,\gamma_k\,\|A\|_2}{|\lambda_1 - \lambda_2|}. \tag{36}$$

Let $\tilde{\lambda}_1 \geq \tilde{\lambda}_2$ be the eigenvalues of $\widetilde{A}$. Using again Weyl's theorem, we obtain $|\tilde{\lambda}_i - \lambda_i| \leq \sqrt{2}\gamma_k\|A\|_2$, $i = 1, 2$. Therefore, $|\tilde{\lambda}_i - \lambda_i|/|\lambda_i| \leq \sqrt{2}\gamma_k\kappa_2(A)$. Lemma A.4 implies

$$\frac{|\tilde{\lambda}_i - \lambda_i|}{|\lambda_i|} \leq 2\sqrt{2}\,\frac{1+\alpha}{1-\alpha}\,\gamma_k, \tag{37}$$

and the first assumption in (31) leads to $|\tilde{\lambda}_i - \lambda_i|/|\lambda_i| \leq 1/2$. Therefore, $\tilde{\lambda}_i$ and λ_i have the same sign. The matrix $\widetilde{A}$ is indefinite, as is A, thus $|\lambda_1 - \lambda_2| > \|A\|_2$, and

$$\frac{1}{2}\sin 2\,\theta(v_i, \widetilde{v}_i) \leq \sqrt{2}\,\gamma_k. \tag{38}$$

The first assumption in (31) implies $\sin 2\,\theta(v_i, \widetilde{v}_i) < 1/2$; thus $1/\sqrt{2} \leq \cos\theta(v_i, \widetilde{v}_i)$. From this bound and (38), we obtain $\sin\theta(v_i, \widetilde{v}_i) \leq 2\,\gamma_k$, and, by using that $\|v_i - \widetilde{v}_i\|_2 \leq \sqrt{2}\sin\theta(v_i, \widetilde{v}_i)$,

$$\|v_i - \widetilde{v}_i\|_2 \;\leq\; 2\sqrt{2}\,\gamma_k < \gamma_{4k}, \qquad i = 1, 2. \tag{39}$$

Now, notice that the error bounds for $\widehat{cs}$ and $\widehat{sn}$ appearing in Theorem A.1 lead to $\|\widehat{v}_i - \widetilde{v}_i\|_2 \leq \gamma_{141}$. Finally,

$$\|\widehat{v}_i - v_i\|_2 \leq \|\widehat{v}_i - \widetilde{v}_i\|_2 + \|\widetilde{v}_i - v_i\|_2 \leq \gamma_{4k+141}, \qquad i = 1, 2,$$

which is (33).

Let us prove the third item. We prove only the error bound for sn. The bound for cs is proved in a similar way. The bound (39) implies

$$\left|\frac{sn - \widetilde{sn}}{\widetilde{sn}}\right| \leq \frac{2\sqrt{2}\,\gamma_k}{|\widetilde{sn}|} < 6\sqrt{2}\,\gamma_k,$$

where we have used that $1/3 < |\widetilde{sn}|$, according to Lemma A.3. Then

$$\left|\frac{sn - \widetilde{sn}}{sn}\right| \leq \frac{6\sqrt{2}\,\gamma_k}{1 - 6\sqrt{2}\,\gamma_k} \leq\; (2 + \sqrt{2})\,6\sqrt{2}\,\gamma_k \;<\; \gamma_{48k}, \tag{40}$$

where we have used that $6\sqrt{2}\gamma_k \leq 1/\sqrt{2}$. The previous bound can also be written as $\widetilde{sn} = sn(1 + \theta_{48k})$. Combining this expression with Theorem A.1, we get the bound in (34) for the sine. ∎

Lemma A.5 allows us to prove the main theorem of this section. In this theorem, we extend the symbols θ_x and γ_x introduced in (24) to noninteger values of $x \geq 1$. In particular, it is easy to check that Lemma 3.3 in [22] remains valid for these noninteger values.

THEOREM A.6. *Let $B = B^T$ be an $n \times n$ real matrix, and let $S^{(m)}$ be its mth Schur complement, $0 \leq m \leq n-1$. Let us assume that all the entries of the Schur complements of B can be computed with relative error bounded by γ_k in floating point arithmetic with machine precision ϵ, i.e.,*

$$\widehat{S}^{(m)}_{ij} = S^{(m)}_{ij}\,(1 + \beta^{(m)}_{ij}), \qquad |\beta^{(m)}_{ij}| \leq \gamma_k \qquad \textit{for all } i, j, m, \tag{41}$$

where $\widehat{S}^{(m)}$ are the computed Schur complements. Let us also assume that the Bunch–Parlett pivoting strategy applied to B in floating point arithmetic does not permute any rows or columns of B.

Let $\widehat{X}\widehat{D}\widehat{X}^T$ be the RRD of B computed in floating point arithmetic by applying the Bunch–Parlett method to the Schur complements $\widehat{S}^{(m)}$, $0 \leq m \leq n-1$, followed by the Jacobi spectral diagonalization of the 2×2 pivots, as in (6)*. Let us apply this algorithm to B in exact arithmetic by choosing the same dimensions for the pivots as those selected in floating point arithmetic. Let X and D be the exact factors, i.e., $B = XDX^T$. If*

$$4\sqrt{2}\,\frac{1+\alpha}{1-\alpha}\,\gamma_{k+29} \;\leq 1 \qquad \textit{and} \qquad \gamma_{141+48k} \leq 1,$$

then

1.

$$|\widehat{D}_{ii} - D_{ii}| \leq 4\,\frac{1+\alpha}{1-\alpha}\,\gamma_{k+29}\,|D_{ii}|, \qquad 1 \leq i \leq n;$$

2.

$$\|\widehat{X} - X\|_F \leq 2\sqrt{2}\,\frac{1+\alpha}{1-\alpha}\,\gamma_{h(\alpha)}\,\|X\|_F, \tag{42}$$

where

$$h(\alpha) = \left(8\,\frac{1+\alpha}{1-\alpha} + 49\right)\,k + 232\,\frac{1+\alpha}{1-\alpha} + 144; \tag{43}$$

3.

$$\|\widehat{X}(:,j) - X(:,j)\|_2 \leq \frac{4\sqrt{2n}(1+\alpha)}{(1-\alpha)^2(1-\gamma_{g(\alpha)})}\,\gamma_{h(\alpha)}\|X(:,j)\|_2, \qquad 1 \leq j \leq n; \tag{44}$$

where

$$g(\alpha) = \left(32\left(\frac{1+\alpha}{1-\alpha}\right)^2 + 196\,\frac{1+\alpha}{1-\alpha}\right)\,k, \tag{45}$$

and it is assumed that $\gamma_{g(\alpha)} < 1$.

Theorem 3.1 follows from Theorem A.6, taking $k = 8\,n$, $\alpha = 0.6404$, and increasing some of the bounds to get simpler expressions.

Proof of Theorem A.6. The first item is trivial in the case of 1×1 pivots, and it is a consequence of (32) for the 2×2 pivots, selected by the Bunch–Parlett strategy.

If $X(:,s)$ is a column of X corresponding to a 1×1 pivot, we simply combine roundoff errors to get $\widehat{X}(i,s) = X(i,s)(1+\theta_{2k+1})$, and then

$$\|\widehat{X}(:,s) - X(:,s)\|_2 \leq \gamma_{2k+1}\,\|X(:,s)\|_2. \tag{46}$$

Therefore, we need only focus on the columns corresponding to 2×2 pivots.

Let us assume for the rest of the proof that $X(:, j : j+1)$ are two columns of X corresponding to a 2×2 pivot. Let us denote the nontrivial part of X as follows: $X(j : j+1, j : j+1) \equiv X_{11}$ and $X(j+2 : n, j : j+1) \equiv X_{21}$. We will also use $S_{21} \equiv S^{(j-1)}(j+2 : n, j : j+1)$. The 2×2 pivot is $S_{11} \equiv S^{(j-1)}(j : j+1, j : j+1)$, and its orthogonal diagonalization is denoted by $S_{11} = U\Lambda U^T$. Finally, $\Lambda \equiv \mathrm{diag}(\lambda_1, \lambda_2)$. The corresponding computed magnitudes will be denoted by the same hatted letters.

According to (6),

$$\|\widehat{X}_{11} - X_{11}\|_F = \|\widehat{U} - U\|_F \leq \sqrt{2}\,\gamma_{4k+141} = \gamma_{4k+141}\|X_{11}\|_F, \tag{47}$$

where (33) has been used. To study the error in X_{21}, it is convenient to define

$$f(\alpha) \equiv 4\frac{1+\alpha}{1-\alpha}.$$

Thus, (32) implies that $\hat{\lambda}_p = \lambda_p(1+\theta_{f(\alpha)\,(k+29)})$, for $p = 1, 2$. Notice that, by (6), $X_{21} = S_{21}U\Lambda^{-1}$. Then for the computed magnitude,

$$(\widehat{X}_{21})_{pq} = \sum_{l=1}^{2} \frac{(\widehat{S}_{21})_{pl}\,(\widehat{U})_{lq}}{\hat{\lambda}_q}(1+\theta_3^{(p,l,q)}) = \sum_{l=1}^{2} \frac{(S_{21})_{pl}\,U_{lq}}{\lambda_q}(1+\theta_{h(\alpha)}^{(p,l,q)}),$$

where $h(\alpha)$ is given by (43), and (34) has been used to bound the errors in the entries of U. The previous equation leads to

$$|\widehat{X}_{21} - X_{21}| \leq \gamma_{h(\alpha)}\, |S_{21}|\, |U\Lambda^{-1}|,$$

where, for any matrix B, $|B|$ is the matrix whose entries are the absolute values of the entries of B. Now, we use that the Frobenius norm is unitarily invariant to get

$$\begin{aligned} \|\widehat{X}_{21} - X_{21}\|_F &\leq \gamma_{h(\alpha)}\, \|S_{21}U\|_F\, \|\Lambda^{-1}\|_F \\ &\leq \sqrt{2}\,\gamma_{h(\alpha)}\, \|S_{21}U\Lambda^{-1}\,\Lambda\|_F\, \|\Lambda^{-1}\|_2 \\ &\leq \sqrt{2}\,\gamma_{h(\alpha)}\, \|S_{21}U\Lambda^{-1}\|_F\, \kappa_2(\Lambda) \\ &\leq 2\sqrt{2}\,\frac{1+\alpha}{1-\alpha}\,\gamma_{h(\alpha)}\, \|X_{21}\|_F, \end{aligned} \tag{48}$$

where (29) and $\kappa_2(S_{11}) = \kappa_2(\Lambda)$ have been used. This inequality and (47) imply

$$\|\widehat{X}(:, j:j+1) - X(:, j:j+1)\|_F \leq 2\sqrt{2}\,\frac{1+\alpha}{1-\alpha}\,\gamma_{h(\alpha)}\, \|X(:, j:j+1)\|_F.$$

The normwise bound (42) is finally obtained by combining the above inequality with (46).

The proof of the columnwise error bound (44) needs more work in the case of columns of X corresponding to 2×2 pivots. It relies on two properties. The first is that the absolute values of the entries of the matrix $\widehat{S}_{21}\widehat{S}_{11}^{-1}$ are bounded by $1/(1-\alpha)$ because $\widehat{S}_{11}$ is a 2×2 pivot chosen by the Bunch–Parlett pivoting strategy [4, 22] (see also [20, page 118] for a simple proof). The second is that $X_{11} = U$, and, as a consequence, both columns of $X(:, j:j+1)$ have a norm greater than or equal to 1.

We will use some additional notation in the rest of the proof. Let $\widehat{S}_{11} = \widetilde{U}\widetilde{\Lambda}\widetilde{U}^T$ be the exact orthogonal diagonalization of $\widehat{S}_{11}$. Notice that we have previously used $S_{11} = U\Lambda U^T$, the exact orthogonal diagonalization of the *exact* block S_{11}, and $\widehat{U}\widehat{\Lambda}\widehat{U}^T$, the computed orthogonal diagonalization of $\widehat{S}_{11}$. We will also use the matrices $\widetilde{X}_{11} \equiv \widetilde{U}$ and $\widetilde{X}_{21} = \widehat{S}_{21}\widetilde{U}\widetilde{\Lambda}^{-1}$. Finally, $\widetilde{\Lambda} \equiv \mathrm{diag}(\tilde{\lambda}_1, \tilde{\lambda}_2)$.

According to [20, page 118],

$$\|\widetilde{X}_{21}\|_F = \|\widehat{S}_{21}\widehat{S}_{11}^{-1}\|_F \leq \frac{\sqrt{2(n-j-1)}}{1-\alpha} \leq \frac{\sqrt{2\,n}}{1-\alpha}. \tag{49}$$

Let us relate $\|\widetilde{X}_{21}\|_F$ to $\|X_{21}\|_F$. Notice that

$$(\widetilde{X}_{21})_{pq} = \sum_{l=1}^{2} \frac{(\widehat{S}_{21})_{pl}\,(\widetilde{U})_{lq}}{\tilde{\lambda}_q}. \tag{50}$$

The difference between the eigenvalues and eigenvectors of $\widehat{S}_{11}$ and those of S_{11} can be bounded as done in (37) and (40) for A and $\widetilde{A}$. Therefore, $\tilde{\lambda}_q = \lambda_q\,(1 + \theta_{f(\alpha)\,k})$ and $(\widetilde{U})_{lq} = U_{lq}(1 + \theta_{48k})$. Moreover, $(\widehat{S}_{21})_{pl} = (S_{21})_{pl}\,(1 + \theta_k)$, and (50) implies

$$(\widetilde{X}_{21})_{pq} = \sum_{l=1}^{2} \frac{(S_{21})_{pl}\,U_{lq}}{\lambda_q}\,(1 + \theta_{(2f(\alpha)+49)k}).$$

This implies $|\widetilde{X}_{21} - X_{21}| \leq \gamma_{(2f(\alpha)+49)k}\,|S_{21}|\,|U\Lambda^{-1}|$. An argument similar to that leading to (48) implies

$$\|\widetilde{X}_{21} - X_{21}\|_F \leq \gamma_{g(\alpha)}\,\|X_{21}\|_F,$$

where $g(\alpha)$ is given by (45). This bound and (49) yield

$$\|X_{21}\|_F \leq \|\widetilde{X}_{21}\|_F + \|X_{21} - \widetilde{X}_{21}\|_F \leq \frac{\sqrt{2\,n}}{1-\alpha} + \gamma_{g(\alpha)}\,\|X_{21}\|_F$$

and

$$\|X_{21}\|_F \leq \frac{\sqrt{2n}}{(1-\alpha)(1-\gamma_{g(\alpha)})}.$$

We substitute this bound in (48) to get

$$\|\widehat{X}_{21} - X_{21}\|_F \leq \frac{4\,\sqrt{n}\,(1+\alpha)}{(1-\alpha)^2\,(1-\gamma_{g(\alpha)})}\,\gamma_{h(\alpha)}.$$

This inequality and (47) imply

$$\|\widehat{X}(:, j:j+1) - X(:, j:j+1)\|_F \leq \frac{4\,\sqrt{2\,n}\,(1+\alpha)}{(1-\alpha)^2\,(1-\gamma_{g(\alpha)})}\,\gamma_{h(\alpha)}.$$

The bound (44) follows from (46) and the previous bound because $\max\{\|\widehat{X}(:,j) - X(:,j)\|_2, \|\widehat{X}(:,j+1) - X(:,j+1)\|_2\} \leq \|\widehat{X}(:, j:j+1) - X(:, j:j+1)\|_F$ and $1 \leq \|X(:,j)\|_2$, $1 \leq \|X(:,j+1)\|_2$. □

Acknowledgments. The authors thank the editor of this paper, Prof. Ilse Ipsen, for her careful reading of the original manuscript and for many helpful suggestions that have significantly improved the paper.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, 3rd ed., SIAM, Philadelphia, 1999.

[2] T. Ando, *Totally positive matrices*, Linear Algebra Appl., 90 (1987), pp. 165–219.

[3] D. S. Bindel and S. Givindjee, *Elastic PMLs for Resonator Anchor Loss Simulation*, Report no. UCB/SEMM-2005/01, University of California, Berkeley, CA, 2005. Available online http://www.cs.berkeley.edu/∼dbindel/papers/pml-tr.pdf.

[4] J. R. Bunch and B. N. Parlett, *Direct methods for solving symmetric indefinite systems of linear equations*, SIAM J. Numer. Anal., 8 (1971), pp. 639–655.

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

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

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

[8] J. Demmel and P. Koev, *Accurate SVDs of polynomial Vandermonde matrices involving orthonormal polynomials*, Linear Algebra Appl., 417 (2006), pp. 382–396.

[9] J. Demmel and K. Veselić, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1245.

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

[11] 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.
[12] S. C. Eisenstat and I. C. F. Ipsen, *Relative perturbation techniques for singular value problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1972–1988.
[13] K. Fernando and B. Parlett, *Accurate singular values and differential qd algorithms*, Numer. Math., 67 (1994), pp. 191–229.
[14] S. Fomin and A. Zelevinsky, *Total positivity: Tests and parametrizations*, Math. Intelligencer, 22 (2000), pp. 23–33.
[15] F. Gantmacher, *The Theory of Matrices*, Vol. 1, AMS Chelsea Publishing, Providence, RI, 1998.
[16] F. Gantmacher and M. Krein, *Oscillation Matrices and Kernels and Small Vibrations of Mechanical Systems*, revised ed., AMS Chelsea Publishing, Providence, RI, 2002.
[17] M. Gasca and C. A. Micchelli, eds., *Total Positivity and Its Applications*, Math. Appl. 359, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.
[18] M. Gasca and J. M. Peña, *Total positivity and Neville elimination*, Linear Algebra Appl., 165 (1992), pp. 25–44.
[19] M. Gasca and J. M. Peña, *On factorizations of totally positive matrices*, in Total Positivity and Its Applications, M. Gasca and C. A. Micchelli, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996, pp. 109–130.
[20] P. E. Gill, W. Murray, and M. H. Wright, *Numerical Linear Algebra and Optimization*, Vol. 1, Addison-Wesley Publishing Company, Advanced Book Program, Redwood City, CA, 1991.
[21] G. Golub and C. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[22] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., SIAM, Philadelphia, 2002.
[23] *IEEE Standard for Binary Floating Point Arithmetic*, Std 754-1985, ANSI/IEEE, New York, 1985.
[24] S. Karlin, *Total Positivity*, Vol. I, Stanford University Press, Stanford, CA, 1968.
[25] P. Koev, *Accurate computations with totally nonnegative matrices*, SIAM J. Matrix Anal. Appl., submitted.
[26] P. Koev, *Accurate eigenvalues and SVDs of totally nonnegative matrices*, SIAM J. Matrix Anal. Appl., 27 (2005), pp. 1–23.
[27] R.-C. Li, *Relative perturbation theory.* II. *Eigenspace and singular subspace variations*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 471–492.
[28] R.-C. Li, *Relative perturbation theory.* IV. $\sin 2\theta$ *theorems*, Linear Algebra Appl., 311 (2000), pp. 45–60.
[29] *MATLAB Reference Guide*, The MathWorks, Inc., Natick, MA, 1992.
[30] M. J. Peláez and J. Moro, *High accuracy eigenvalue algorithms for symmetric DSTU and TSC matrices*, SIAM J. Matrix Anal. Appl., submitted.
[31] J. M. Peña, *LDU decompositions with L and U well conditioned*, Electron. Trans. Numer. Anal., 18 (2004), pp. 198–208 (electronic).
[32] I. Slapničar, *Componentwise analysis of direct factorization of real symmetric and Hermitian matrices*, Linear Algebra Appl., 272 (1998), pp. 227–275.
[33] I. Slapničar, *Accurate Symmetric Eigenreduction by a Jacobi Method*, Ph.D. thesis, Fernuniversität—Hagen, Hagen, Germany, 1992.
[34] G. W. Stewart, *Matrix Algorithms*, Vol. I, SIAM, Philadelphia, 1998.
[35] K. Veselić, *A Jacobi eigenreduction algorithm for definite matrix pairs*, Numer. Math., 64 (1993), pp. 241–269.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 4, pp. 1157–1172

NUMERICAL METHODS FOR THE TRIDIAGONAL HYPERBOLIC QUADRATIC EIGENVALUE PROBLEM*

BOR PLESTENJAK†

Abstract. We consider numerical methods for the computation of the eigenvalues of the tridiagonal hyperbolic quadratic eigenvalue problem. The eigenvalues are computed as zeros of the characteristic polynomial using the bisection, Laguerre's method, and the Ehrlich–Aberth method. Initial approximations are provided by a divide-and-conquer approach using rank two modifications, and we show that these initial approximations interlace with the exact eigenvalues. The above methods need a stable and efficient evaluation of the quadratic eigenvalue problem's characteristic polynomial and its derivatives. We discuss how to obtain these values using three-term recurrences, the QR factorization, and the LU factorization. Numerical results show that the presented methods are more efficient than solving a linearized generalized eigenvalue problem.

Key words. quadratic eigenvalue problem, inertia, Laguerre's method, Ehrlich–Aberth method, bisection, LU factorization, QR factorization, divide-and-conquer

AMS subject classifications. 65F15, 15A18, 15A69

DOI. 10.1137/050624856

1. Introduction.

We consider a Hermitian quadratic eigenvalue problem (QEP)

$$Q(\lambda)x = (\lambda^2 M + \lambda C + K)x = 0, \tag{1.1}$$

where M, C, and K are $n \times n$ Hermitian matrices. If (1.1) is satisfied for a nonzero $x \in \mathbb{C}^n$ and $\lambda \in \mathbb{C}$, then λ is an eigenvalue and x is the corresponding (right) eigenvector. The characteristic polynomial $f(\lambda) = \det(Q(\lambda))$ is of degree less than or equal to $2n$. A QEP is regular when f is not identically zero. A regular QEP has $2n$ eigenvalues, finite or infinite. The finite eigenvalues are the zeros of f while the infinite eigenvalues correspond to the zero eigenvalues of the reversed QEP $\lambda^2 Q(1/\lambda) = \lambda^2 K + \lambda C + M$. If M is nonsingular, then we have $2n$ finite eigenvalues with up to $2n$ eigenvectors, which are not necessarily linearly independent. QEPs appear in various applications; for a recent survey of the QEP see [21].

We say that a QEP is *hyperbolic* [13] if M is positive definite and

$$(x^*Cx)^2 > 4(x^*Mx)(x^*Kx)$$

for all $x \neq 0$. For a hyperbolic QEP the eigenvalues are all real. In this paper we focus on the tridiagonal hyperbolic QEP, where matrices M, C, and K are Hermitian and tridiagonal. An example of a tridiagonal quadratic eigenvalue problem is a damped mass-spring system (see, e.g., [21]). Our goal is to compute all or some of the eigenvalues. For the computation of the eigenvalues we apply polynomial solvers to the characteristic polynomial.

New theoretical results are presented in two theorems. In the first theorem we generalize the inertia. The second result is that the initial approximations, obtained

*Received by the editors February 20, 2005; accepted for publication (in revised form) by I. C. F. Ipsen June 2, 2006; published electronically December 18, 2006. The research was supported in part by Ministry of Higher Education, Science and Technology of Slovenia research project Z1-3136.
http://www.siam.org/journals/simax/28-4/62485.html
†Department of Mathematics, University of Ljubljana, Jadranska 19, SI-1000 Ljubljana, Slovenia (bor.plestenjak@fmf.uni-lj.si).

by a divide-and-conquer approach, interlace with the exact eigenvalues. These two results enable us to show that all the eigenvalues can be computed using Laguerre's method. The eigenvectors can be later obtained by the inverse iteration; for a stable algorithm see [7].

We show that some of the presented methods can be applied to more general problems, e.g., to the banded polynomial eigenvalue problems, though there is as yet no theory to support this approach.

The paper is organized as follows. In section 2 we recall some results on hyperbolic QEPs. The inertia of a hyperbolic QEP is discussed in section 3. In sections 4, 5, and 6 three different approaches, based respectively on the three-term recurrences, QR factorization, and LU factorization, for the computation of the derivatives of the characteristic polynomial are presented. The divide-and-conquer approach for the initial approximations is presented in section 7. In sections 8 and 9 Laguerre's method and the Ehrlich–Aberth method are applied to the computation of the zeros of the characteristic polynomial, respectively.

Some numerical examples are given in section 10, followed by conclusions.

2. Auxiliary results. The following properties of the hyperbolic QEPs are gathered from [8, 13, 18]. A hyperbolic QEP has $2n$ real eigenvalues and eigenvectors. All eigenvalues are semisimple and there is a gap between the largest n (*primary*) and the smallest n (*secondary*) eigenvalues. There are n linearly independent eigenvectors associated with the primary and the secondary eigenvalues, respectively.

For each $x \neq 0$ the equation

$$\mu^2 x^* M x + \mu x^* C x + x^* K x = 0$$

has two distinct real solutions $\mu_1(x) < \mu_2(x)$. If x is an eigenvector, then at least one of $\mu_1(x)$ and $\mu_2(x)$ is the corresponding eigenvalue. Values $\mu_1(x)$ and $\mu_2(x)$ are generalizations of the Rayleigh quotient. As for the single symmetric matrix case, there exists a minimax theorem for hyperbolic QEPs as well.

THEOREM 2.1 (Duffin [8]). *If $\lambda_{2n} \leq \cdots \leq \lambda_1$ are the eigenvalues of a hyperbolic QEP, then*

$$\lambda_{n+i} = \max_{\substack{S \subset \mathbb{C}^n \\ \dim(S)=i}} \min_{0 \neq x \in S} \mu_1(x) \qquad \text{and} \qquad \lambda_i = \max_{\substack{S \subset \mathbb{C}^n \\ \dim(S)=i}} \min_{0 \neq x \in S} \mu_2(x)$$

for $i = 1, \ldots, n$.

THEOREM 2.2 (Markus [18]). *A Hermitian QEP where M is positive definite is hyperbolic if and only if there exists $\gamma \in \mathbb{R}$ such that the matrix $Q(\gamma)$ is negative definite.*

Remark 2.3. The scalar γ in Theorem 2.2, such that $Q(\gamma)$ is negative definite, lies in the gap between the primary and the secondary eigenvalues, i.e., $\lambda_{n+1} < \gamma < \lambda_n$.

3. Inertia of a hyperbolic QEP. The inertia of a Hermitian matrix A is a triplet of nonnegative integers (ν, ζ, π), where ν, ζ, and π are, respectively, the number of negative, zero, and positive eigenvalues of A. The following theorem shows that the inertia of a Hermitian matrix $Q(\sigma)$ is related to the number of eigenvalues of the QEP Q that are larger or smaller than σ, respectively.

THEOREM 3.1. *Let M, C, and K be Hermitian $n \times n$ matrices such that $Q(\lambda) = \lambda^2 M + \lambda C + K$ is a hyperbolic QEP, and let $\lambda_{2n} \leq \cdots \leq \lambda_{n+1} < \lambda_n \leq \cdots \leq \lambda_1$ be the eigenvalues of the QEP Q. If (ν, ζ, π) is the inertia of the matrix $Q(\sigma)$, then ζ is the algebraic multiplicity of σ as an eigenvalue of the QEP Q and*

(a) *if $\sigma > \lambda_n$, then ν is the number of eigenvalues of Q larger than σ and $\pi + n$ is the number of eigenvalues of Q smaller than σ;*
(b) *if $\sigma < \lambda_{n+1}$, then ν is the number of eigenvalues of Q smaller than σ and $\pi + n$ is the number of eigenvalues of Q larger than σ.*

Proof. For each $\lambda \in \mathbb{R}$, $Q(\lambda)$ is a Hermitian $n \times n$ matrix with n real ordered eigenvalues

$$\mu_n(\lambda) \le \cdots \le \mu_1(\lambda), \tag{3.1}$$

where $\mu_1, \ldots, \mu_n$ are continuous functions of λ. It is easy to see that σ is an eigenvalue of the QEP Q of algebraic multiplicity k exactly when there exists $1 \le i \le n$ such that

$$\mu_i(\sigma) = \mu_{i+1}(\sigma) = \cdots = \mu_{i+k-1}(\sigma) = 0.$$

Since M is a Hermitian positive definite matrix,

$$\lim_{\lambda \to \pm\infty} \mu_i(\lambda) = \infty$$

for all i. By Theorem 2.2 there exists $\sigma_0 \in \mathbb{R}$ such that $\mu_i(\sigma_0) < 0$ for all i. Because each μ_i is a continuous function, it has at least two zeros, one on the right and one on the left side of σ_0. As each zero of μ_i is also an eigenvalue of the QEP Q which has $2n$ eigenvalues, it follows that each μ_i has exactly two zeros.

As $\mu_1, \ldots, \mu_n$ are continuous and ordered as in (3.1), it is not hard to deduce that if $\sigma > \sigma_0$ and σ is not an eigenvalue of Q, then the number of negative eigenvalues of $Q(\sigma)$ equals the number of eigenvalues of Q that are larger than σ. This proves (a), and similarly we can prove (b). ∎

Remark 3.2. Theorem 3.1 is a generalization of a similar theorem in [20], where M is a positive definite matrix and K is a negative definite matrix. In this case $Q(0)$ is negative definite and a proof similar to the one above can be done without applying Theorem 2.2.

Based on the inertia we could apply the bisection to obtain the kth eigenvalue. The algorithm is similar to the algorithm for the symmetric eigenvalue problem. To derive more efficient methods, we use some faster methods that were successfully applied to tridiagonal eigenvalue problems: Laguerre's method [16, 17] and the Ehrlich–Aberth method [4].

The above methods need stable and efficient computation of $\nu(Q(\lambda))$, $f(\lambda)$, $f'(\lambda)/f(\lambda)$, and $f''(\lambda)/f(\lambda)$, where $f(\lambda) = \det(Q(\lambda))$. We discuss how to obtain these values using the three-term recurrences, the QR factorization, and the LU factorization in the next three sections.

4. Three-term recurrences. Let $Q(\lambda) = (\lambda^2 M + \lambda C + K)$, where M, C, and K are $n \times n$ tridiagonal matrices. We can write

$$Q(\lambda) = \begin{bmatrix} a_1(\lambda) & b_1(\lambda) & & & & 0 \\ b_1(\lambda) & a_2(\lambda) & b_2(\lambda) & & & \\ & \ddots & \ddots & \ddots & & \\ & & b_{n-2}(\lambda) & a_{n-1}(\lambda) & b_{n-1}(\lambda) & \\ 0 & & & b_{n-1}(\lambda) & a_n(\lambda) & \end{bmatrix},$$

where $a_i(\lambda) = \lambda^2 M_{ii} + \lambda C_{ii} + K_{ii}$ and $b_i(\lambda) = \lambda^2 M_{i+1,i} + \lambda C_{i+1,i} + K_{i+1,i}$ are quadratic polynomials. The determinant of a tridiagonal matrix can be computed using a three-term recurrence; see, e.g., [11]. If $f_k(\lambda) = \det(Q_k(\lambda))$, where $Q_k(\lambda)$ is the leading

$k \times k$ submatrix of $Q(\lambda)$, then

$$\begin{array}{rclcrcl} f_0 & = & 1, & & f_1 & = & a_1, \\ f_0' & = & 0, & & f_1' & = & a_1', \\ f_0'' & = & 0, & & f_1'' & = & a_1'', \end{array}$$

and

$$\begin{aligned} f_{r+1} &= a_{r+1}f_r - b_r^2 f_{r-1}, \\ f_{r+1}' &= a_{r+1}'f_r + a_{r+1}f_r' - 2b_r b_r' f_{r-1} - b_r^2 f_{r-1}', \\ f_{r+1}'' &= a_{r+1}''f_r + 2a_{r+1}'f_r' + a_{r+1}f_r'' - 2b_r'^2 f_{r-1} - 2b_r b_r'' f_{r-1} - 4b_r b_r' f_{r-1}' - b_r^2 f_{r-1}'' \end{aligned}$$

for $r = 1, \ldots, n-1$. For the sake of brevity the argument λ is omitted in the above equations.

As the above recurrences may suffer from overflow and underflow problems [15], we define

$$d_i = \frac{f_i}{f_{i-1}}, \quad g_i = \frac{f_i'}{f_i}, \quad h_i = \frac{f_i''}{f_i}.$$

Then $f_n = d_1 \cdots d_n$,

$$\begin{array}{rclcrcl} d_1 & = & a_1, & & & & \\ g_0 & = & 0, & & g_1 & = & \dfrac{a_1'}{a_1}, \\ h_0 & = & 0, & & h_1 & = & \dfrac{a_1''}{a_1}, \end{array}$$

and

$$\begin{aligned} d_{r+1} &= a_{r+1} - \frac{b_r^2}{d_r}, \\ g_{r+1} &= \frac{1}{d_{r+1}}\left(a_{r+1}' + a_{r+1}g_r - \frac{1}{d_r}(2b_r b_r' + b_r^2 g_{r-1})\right), \\ h_{r+1} &= \frac{1}{d_{r+1}}\left(a_{r+1}'' + 2a_{r+1}'g_r + a_{r+1}h_r - \frac{1}{d_r}(2b_r'^2 + 2b_r b_r'' + 4b_r b_r' g_{r-1} + b_r^2 h_{r-1})\right) \end{aligned}$$

for $r = 1, \ldots, n-1$.

Remark 4.1. One can see that $d_1, \ldots, d_n$ are the diagonal elements from the LDL^* factorization of the matrix $Q(\lambda)$.

Remark 4.2. The algorithm may break down if $d_r = 0$ for some $r = 1, \ldots, n-1$. In such a case we introduce small perturbations and set

$$d_r = \frac{\varepsilon}{d_{r-1}}\left(|\lambda|^2|M_{r-1,r-1}| + |\lambda||C_{r-1,r-1}| + |K_{r-1,r-1}| + \varepsilon\right),$$

where ε is the machine precision. This corresponds to a small relative perturbation of the matrices M, C, and K. A similar approach is used in [16].

5. A QR factorization approach. If $f(\lambda) \neq 0$, then it follows from Jacobi's formula for the derivative of the determinant that

$$f'(\lambda)/f(\lambda) = \operatorname{tr}(Q(\lambda)^{-1}Q'(\lambda)). \tag{5.1}$$

If we denote $A = Q(\lambda)$ and $B = Q'(\lambda)$, then we need to compute $\mathrm{tr}(A^{-1}B)$, where in our case A and B are tridiagonal matrices. In [4] one can find a stable $\mathcal{O}(n)$ computation of $\mathrm{tr}(A^{-1})$ via QR factorization. In this section we generalize this algorithm to compute $\mathrm{tr}(A^{-1}B)$. We start with a sketch of the algorithm for $\mathrm{tr}(A^{-1})$; for details and the theory, see [4].

Let A be a tridiagonal matrix and let $A = UR$, where

$$R = \begin{bmatrix} r_1 & s_1 & t_1 & & \\ & \ddots & \ddots & \ddots & \\ & & r_{n-2} & s_{n-2} & t_{n-2} \\ & & & r_{n-1} & s_{n-1} \\ & & & & r_n \end{bmatrix}$$

is an upper triangular tridiagonal matrix and U is the product of $n-1$ Givens rotations, $U^* = G_{n-1} \cdots G_2 G_1$, where

$$G_i([i, i+1], [i, i+1]) = \begin{bmatrix} \psi_i & \theta_i \\ -\overline{\theta_i} & \overline{\psi_i} \end{bmatrix} \quad \text{and} \quad |\psi_i|^2 + |\theta_i|^2 = 1.$$

Then

$$U^* = \begin{bmatrix} v_1u_1 & \psi_1 & & & 0 \\ v_2u_1 & v_2u_2 & \psi_2 & & \\ \vdots & \vdots & \ddots & \ddots & \\ \vdots & \vdots & & v_{n-1}u_{n-1} & \psi_{n-1} \\ v_nu_1 & v_nu_2 & \cdots & v_nu_{n-1} & v_nu_n \end{bmatrix},$$

where

$$\begin{aligned} D &= \mathrm{diag}(1, -\overline{\psi_1}, \overline{\psi_1\psi_2}, \ldots, (-1)^{n-1}\overline{\psi_1\psi_2\cdots\psi_{n-1}}), \\ u &= D^{-1}[1, \overline{\theta_1}, \ldots, \overline{\theta_{n-1}}]^T, \\ v &= D[\theta_1, \ldots, \theta_{n-1}, 1]^T. \end{aligned}$$

If we solve $Rw = v$, then

$$\mathrm{tr}(A^{-1}) = \sum_{i=1}^{n} u_i w_i. \tag{5.2}$$

Kressner [12] generalized the above approach into an $\mathcal{O}(n)$ algorithm for the computation of $\mathrm{tr}(A^{-1}B)$, where both matrices A and B are tridiagonal. Suppose that

$$B = \begin{bmatrix} x_1 & z_1 & & & 0 \\ y_1 & x_2 & z_2 & & \\ & \ddots & \ddots & \ddots & \\ & & y_{n-2} & x_{n-1} & z_{n-1} \\ 0 & & & y_{n-1} & x_n \end{bmatrix}.$$

To compute $\mathrm{tr}(A^{-1}B)$ we need the diagonal elements of $A^{-1}B$. From

$$(A^{-1}B)_{ii} = e_i^T R^{-1} U^* B e_i$$

$$= z_{i-1}e_i^T R^{-1}U^* e_{i-1} + x_i e_i^T R^{-1}U^* e_i + y_i e_i^T R^{-1}U^* e_{i+1}$$

and

$$\begin{aligned} e_i^T R^{-1}U^* e_{i-1} &= u_{i-1}w_i, \\ e_i^T R^{-1}U^* e_i &= u_i w_i, \\ e_i^T R^{-1}U^* e_{i+1} &= u_{i+1}w_i + \frac{1}{r_i}(\psi_i - v_i u_{i+1}) \end{aligned}$$

it follows that

(5.3)
$$\operatorname{tr}(A^{-1}B) = \sum_{i=2}^{n} z_{i-1}u_{i-1}w_i + \sum_{i=1}^{n} x_i u_i w_i + \sum_{i=1}^{n-1} y_i \left(u_{i+1}w_i + \frac{1}{r_i}(\psi_i - v_i u_{i+1}) \right).$$

As reported in [4], formula (5.2) is not stable. To make it stable, we have to avoid the explicit multiplication by the matrix D or D^{-1}. If we define $\hat{R} = D^{-1}RD$, $\hat{v} = D^{-1}v$, $\hat{u} = Du$, and solve $\hat{R}\hat{w} = \hat{v}$ for $\hat{w}$, then

$$\operatorname{tr}(A^{-1}) = \sum_{i=1}^{n} \hat{u}_i \hat{w}_i. \tag{5.4}$$

Notice that

$$\hat{R} = \begin{bmatrix} \hat{r}_1 & \hat{s}_1 & \hat{t}_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \hat{r}_{n-2} & \hat{s}_{n-2} & \hat{t}_{n-2} \\ & & & \hat{r}_{n-1} & \hat{s}_{n-1} \\ & & & & \hat{r}_n \end{bmatrix},$$

where $\hat{r}_i = r_i$, $\hat{s}_i = -\overline{\psi_i} s_i$, and $\hat{t}_i = -\overline{\psi_i \psi_{i+1}} t_i$.

Using the same notation it follows from

$$\begin{aligned} u_i w_i &= \hat{u}_i \hat{w}_i, \\ u_{i-1}w_{i-1} &= -\overline{\psi_{i-1}}\hat{u}_{i-1}\hat{w}_i, \\ u_{i+1}w_{i-1} &= -\hat{u}_{i+1}\hat{w}_i(\overline{\psi_i})^{-1}, \\ v_i u_{i+1} &= -\hat{v}_i \hat{u}_{i+1}(\overline{\psi_i})^{-1} \end{aligned}$$

that we may rewrite formula (5.3) in a stable form:

$$\operatorname{tr}(A^{-1}B) = \sum_{i=2}^{n} x_i \hat{u}_i \hat{w}_i - \sum_{i=1}^{n} z_{i-1}\overline{\psi_{i-1}}\hat{u}_{i-1}\hat{w}_i - \sum_{i=1}^{n-1} \frac{y_i}{\overline{\psi_i}} \left(\hat{u}_{i+1}\hat{w}_i + \frac{1}{r_i}(|\psi_i|^2 + \hat{v}_i \hat{u}_{i+1}) \right).$$

6. An LU factorization approach. In [5] one can find an algorithm for the computation of the derivative of the determinant using the LU factorization. Suppose that $\det(Q(\lambda)) \neq 0$ and that $PQ(\lambda) = LU$ is the result of Gaussian elimination with partial pivoting for $Q(\lambda)$, where L is a lower triangular matrix with ones on the diagonal and U is an upper triangular matrix. Then

$$f(\lambda) = \det(Q(\lambda)) = \det(P) \cdot u_{11}u_{22}\cdots u_{nn}.$$

If we fix the permutation matrix P, then for each μ in a small neighborhood of λ there exist analytic matrices $L(\mu)$ and $U(\mu)$ such that

$$L(\mu)U(\mu) = PQ(\mu) \tag{6.1}$$

is the LU factorization of $PQ(\mu)$. By differentiating (6.1) at $\mu = \lambda$ we get

$$PQ' = L'U + LU' = MU + LV,$$

where $M = L'$ is a lower triangular matrix with zeros on the diagonal and $V = U'$ is an upper triangular matrix. Matrices M and V of the proper form and such that $PQ' = MU + LV$ can be computed from Q', P, L, and U (see Algorithm 6.1). It follows that

$$f'(\lambda) = \det(P) \sum_{i=1}^{n} v_{ii} \prod_{\substack{j=1 \\ j \neq i}}^{n} u_{jj}$$

and

$$\frac{f'(\lambda)}{f(\lambda)} = \sum_{i=1}^{n} \frac{v_{ii}}{u_{ii}}.$$

ALGORITHM 6.1 (Bohte [5]). *The algorithm solves the equation $B = MU + LV$ for M and V, where L is a lower triangular matrix with ones on the main diagonal, U is a nonsingular upper triangular matrix, B is a square $n \times n$ matrix, M is a lower triangular matrix with zeros on the main diagonal, and V is an upper triangular matrix.*

for $r = 1$ to n
  for $k = r$ to n
    $v_{rk} = b_{rk} - \sum_{j=1}^{r-1}(m_{rj}u_{jk} + l_{rj}v_{jk})$
  for $i = r + 1$ to n
    $m_{ir} = \frac{1}{u_{rr}}\left(b_{ir} - \sum_{j=1}^{r-1}(m_{ij}u_{jr} + l_{ij}v_{jr}) - l_{ir}v_{rr}\right)$

For the second derivative we have

$$PQ'' = L''U + 2L'U' + LU'' = NU + 2MV + LW, \tag{6.2}$$

where $N = L''$ is a lower triangular matrix with zeros on the diagonal and $W = U''$ is an upper triangular matrix. It follows that

$$\frac{f''(\lambda)}{f(\lambda)} = \sum_{i=1}^{n} \frac{w_{ii}}{u_{ii}} + \left(\sum_{i=1}^{n} \frac{v_{ii}}{u_{ii}}\right)^2 - \sum_{i=1}^{n} \frac{v_{ii}^2}{u_{ii}^2}.$$

From the relation (6.2) we get $PQ'' - 2MV = NU + LW$, which means that we can apply Algorithm 6.1 for the computation of N and W as well.

An implementation of Algorithm 6.1 for banded matrices computes f'/f and f''/f in a linear time. The algorithm is more expensive than the three-term recurrences in section 4, but its advantage is that it can be applied to nontridiagonal matrices as well. Let us also mention that in [5] one can find a slightly modified algorithm that is able to compute $f'(\lambda)$ even if $f(\lambda) = 0$.

7. Divide-and-conquer. We choose $m \approx n/2$ and write

$$Q(\lambda) = Q_0(\lambda) + b_m(\lambda)(e_{m-1}e_{m+1}^T + e_{m+1}e_{m-1}^T),$$

where

$$Q_0(\lambda) = \begin{bmatrix} Q_1(\lambda) & 0 \\ 0 & Q_2(\lambda) \end{bmatrix}.$$

$Q_0(\lambda)$ is a rank two modification of $Q(\lambda)$. If we apply Theorem 2.2, then it is not hard to see that Q_1 and Q_2 are hyperbolic QEPs. The eigenvalues $\widetilde{\lambda}_{2n} \le \cdots \le \widetilde{\lambda}_1$ of Q_0, a union of the eigenvalues of Q_1 and Q_2, are approximations to the eigenvalues $\lambda_{2n} \le \cdots \le \lambda_1$ of Q.

We can show that the eigenvalues of Q_0 and Q interlace. To show this useful property we introduce a convex combination of Q_0 and Q. Let Q_t be a QEP defined by

$$Q_t(\lambda) = (1-t)Q_0(\lambda) + tQ(\lambda).$$

LEMMA 7.1. *The QEP Q_t is hyperbolic for $t \in [0,1]$.*

Proof. From Theorem 2.2 it follows that there exists γ such that $Q(\gamma)$ is negative definite. Being principal submatrices of $Q(\gamma)$, matrices $Q_1(\gamma)$ and $Q_2(\gamma)$ are negative definite as well. Since it is a block diagonal matrix with negative definite blocks $Q_1(\gamma)$ and $Q_2(\gamma)$, the matrix $Q_0(\gamma)$ is negative definite, too. For $t \in [0,1]$ it is now easy to see that $Q_t(\gamma) = (1-t)Q_0(\gamma) + tQ(\gamma)$ is negative definite and Theorem 2.2 yields that Q_t is a hyperbolic QEP. ∎

The following theory is a generalization of Theorem 5.2 in [17] for the generalized symmetric tridiagonal eigenvalue problem.

LEMMA 7.2. *Let $\lambda_{2n}(t) \le \cdots \le \lambda_1(t)$ be the ordered eigenvalues of the QEP Q_t for $t \in [0,1]$. Each eigencurve $\lambda_i(t)$ is then either constant or strictly monotone for $t \in [0,1]$ and $i = 1, \ldots, 2n$. If we define $\widetilde{\lambda}_0 = \infty$ and $\widetilde{\lambda}_{2n+1} = -\infty$, then each eigencurve $\lambda_i(t)$ lies on the interval $[\widetilde{\lambda}_{i+1}, \widetilde{\lambda}_{i-1}]$.*

Proof. From the construction of Q_t (see, for example, the three-term recurrences in section 4) it follows that the determinant of $Q_t(\lambda)$ can be expressed as

$$p(t,\lambda) := \det Q_t(\lambda) = p_1(\lambda) + t^2 p_2(\lambda),$$

where p_1 and p_2 are polynomials of degree $2n$.

If for a chosen λ_0 we have $p_2(\lambda_0) \neq 0$, then the equation $p(t,\lambda_0) = 0$ has at most one solution on $(0,1]$. If $p_2(\lambda_0) = 0$ and $p_1(\lambda_0) \neq 0$, then none of the eigencurves passes the line $\lambda = \lambda_0$. If $p_2(\lambda_0) = 0$ and $p_1(\lambda_0) = 0$, then λ_0 is an eigenvalue of Q_t for $t \in [0,1]$ and at least one eigencurve $\lambda_i(t)$ is constant and equal to λ_0.

It follows from the above discussion that the eigencurves $\lambda_i(t)$ for $i = 1, \ldots, 2n$ are either constant or strictly monotone for $t \in [0,1]$ (see Figure 7.1). For each $\widetilde{\lambda}_i$ either the only solution of $p(t, \widetilde{\lambda}_i) = 0$ is at $t = 0$ or the eigencurve $\lambda_i(t)$ is constant and equal to $\widetilde{\lambda}_i$. Therefore, $\lambda_i(t)$ is bounded below and above by $\widetilde{\lambda}_{i+1}$ and $\widetilde{\lambda}_{i-1}$, respectively. ∎

THEOREM 7.3. *Let $\widetilde{\lambda}_{2n} \le \cdots \le \widetilde{\lambda}_1$ be the eigenvalues of $Q_0(\lambda)$ and $\lambda_{2n} \le \cdots \le \lambda_1$ the eigenvalues of $Q(\lambda)$. Then*

(a) *$\widetilde{\lambda}_1 \le \lambda_1$ and $\lambda_{2n} \le \widetilde{\lambda}_{2n}$,*

(b) *$\widetilde{\lambda}_{i+1} \le \lambda_i \le \widetilde{\lambda}_{i-1}$ for $i = 2, \ldots, n-1$ and $i = n+2, \ldots, 2n-1$,*

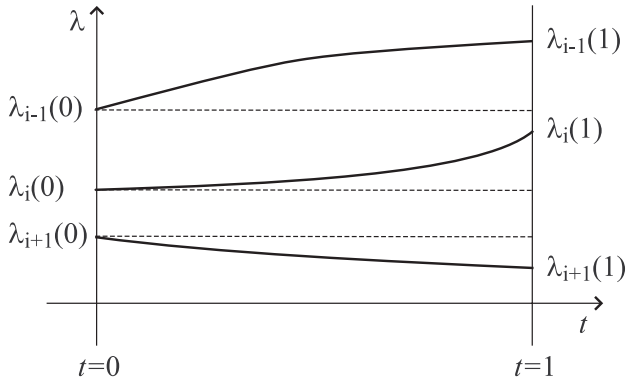


FIG. 7.1. *Eigenvalues of Q_0 and Q interlace.*

(c) $\widetilde{\lambda}_{n+1} \le \lambda_{n+1} < \lambda_n \le \widetilde{\lambda}_n$.

Proof. As matrices $Q_1(\lambda)$ and $Q_2(\lambda)$ are submatrices of $Q(\lambda)$, it follows from Theorem 2.1 that the primary eigenvalues of Q_1 and Q_2 lie in the interval $[\lambda_n, \lambda_1]$. Because of that, the primary eigenvalues of Q_0 lie in the interval $[\lambda_n, \lambda_1]$. Similarly we can show that the secondary eigenvalues of Q_0 lie in the interval $[\lambda_{2n}, \lambda_{n+1}]$. Thus we prove points (a) and (c).

Point (b) follows from Lemma 7.2. We know that λ_i and $\widetilde{\lambda}_i$ are connected by a monotone eigencurve $\lambda_i(t)$, which is bounded below and above by $\widetilde{\lambda}_{i+1}$ and $\widetilde{\lambda}_{i-1}$, respectively. ∎

Remark 7.4. Unlike the divide-and-conquer method for the symmetric tridiagonal matrices, here $\widetilde{\lambda}_i = \widetilde{\lambda}_{i+1}$ does not imply that one of the eigenvalues of Q is $\widetilde{\lambda}_i$. Only if $\widetilde{\lambda}_{i-1} = \widetilde{\lambda}_i = \widetilde{\lambda}_{i+1}$ can one deduce that $\widetilde{\lambda}_i$ is an eigenvalue of Q.

In the conquer phase we use a numerical method that computes the eigenvalues $\lambda_1, \ldots, \lambda_{2n}$ of the QEP Q from the initial approximations $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_{2n}$. Two numerical methods that may be applied for this task are presented in the next two sections. We are not claiming that these are the optimal methods. Other polynomial solvers applied to the classical or to the generalized eigenvalue problem with tridiagonal matrices (see, e.g., [14, 19]) could be applied to the QEP as well.

8. Laguerre's method. To the characteristic polynomial $f(\lambda) = \det(Q(\lambda))$ we can apply Laguerre's method, a well-known globally convergent method for finding polynomial zeros. One step of Laguerre's iteration is

$$L_{\pm}(x) = x + \frac{2n}{\left(\dfrac{-f'(x)}{f(x)} \pm \sqrt{(2n-1)\left((2n-1)\left(\dfrac{-f'(x)}{f(x)}\right)^2 - 2n\dfrac{f''(x)}{f(x)}\right)}\right)}. \tag{8.1}$$

For more details on the method and its properties see, e.g., [16, 22].

For a real polynomial having all real roots the method is globally convergent with a cubic convergence in a neighborhood of a simple eigenvalue. If we add $\lambda_{2n+1} = -\infty$ and $\lambda_0 = \infty$, then for $x \in (\lambda_{i+1}, \lambda_i)$ we have

$$\lambda_{i+1} < L_-(x) < x < L_+(x) < \lambda_i.$$

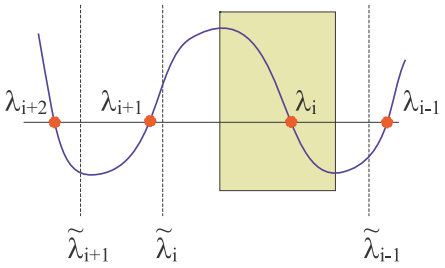


FIG. 8.1. *The cubic convergence region around a simple eigenvalue* λ_i.

In the divide-and-conquer algorithm we use Laguerre's method to compute the eigenvalues $\lambda_{2n} \le \cdots \le \lambda_1$ of Q from the initial approximations $\widetilde{\lambda}_{2n} \le \cdots \le \widetilde{\lambda}_1$ that are the eigenvalues of Q_0. We know from Theorem 7.3 that $\widetilde{\lambda}_{i+1} \le \lambda_i \le \widetilde{\lambda}_{i-1}$ and that we can use $\widetilde{\lambda}_i$ as an initial approximation for λ_i. From $\nu(Q(\widetilde{\lambda}_i))$ we see if $\lambda_i > \widetilde{\lambda}_i$ or $\widetilde{\lambda}_i < \lambda_i$ and then use the appropriate L_+ or L_- sequence. The global convergence of Laguerre's method guarantees that we get all the eigenvalues by computing them independently one by one.

Although the convergence close to a simple eigenvalue should be cubic, we can expect very slow convergence at the beginning if $\widetilde{\lambda}_i$ is closer to λ_{i-1} or λ_{i+1} than to λ_i (see Figure 8.1).

The necessary condition [22] for the cubic convergence near a simple eigenvalue λ is that the sign of $-f'(x)/f(x)$ agrees with the sign of $\lambda - x$ (see Figure 8.1). To improve the convergence we first use the bisection on interval $[\widetilde{\lambda}_{i+1}, \widetilde{\lambda}_i]$ (or $[\widetilde{\lambda}_i, \widetilde{\lambda}_{i-1}]$) until the condition for the cubic convergence is achieved.

Due to rounding errors, the condition $-f'(x)/f(x)(\lambda - x) > 0$ might also be achieved near λ_{i-1} or λ_{i+1}. An additional criterion that we use is that near λ_i the sign of $f'(x)$ has to agree with $(-1)^{i+1}$.

9. Ehrlich–Aberth method. This method simultaneously approximates all zeros of a polynomial $f(\lambda) = \det(Q(\lambda))$. From an initial approximation $x^{(0)} \in \mathbb{C}^{2n}$ the method generates a sequence $x^{(j)} \in \mathbb{C}^{2n}$ that locally converges to the eigenvalues of the QEP Q. The Ehrlich–Aberth iteration is given by

$$x_j^{(k+1)} = x_j^{(k)} - \frac{\dfrac{f(x_j^{(k)})}{f'(x_j^{(k)})}}{1 - \dfrac{f(x_j^{(k)})}{f'(x_j^{(k)})} \displaystyle\sum_{\substack{l=1 \\ l \ne j}}^{2n} \frac{1}{x_j^{(k)} - x_l^{(k)}}} \tag{9.1}$$

for $j = 1, \ldots, 2n$. For details on the method and its properties see, e.g., [3, 4].

If the method is implemented in the Gauss–Seidel style then the convergence for simple roots is cubical and linear for multiple roots. We iterate only those eigenvalues that have not converged yet.

As in the previous section, we use the Ehrlich–Aberth method to compute the eigenvalues of Q using the eigenvalues of Q_0 as initial approximations. It may happen that Q_0 has multiple eigenvalues. In such a case we have a division by zero in (9.1).

In IEEE arithmetic this leads to ∞ in the denominator and consequently to $x_j^{(k+1)} = x_j^{(k)}$. To prevent this, we always slightly perturb the eigenvalues of Q_0 before we use them as initial approximations.

10. Numerical examples. We implemented Laguerre's method and the Ehrlich–Aberth method for the computation of the eigenvalues of the tridiagonal QEPs in Fortran 95. The code can be downloaded from author's web site.[1] Using Compaq Visual Fortran 6.6 on PC Pentium 4 2.6 GHz 1 GB RAM we tested both methods on a limited set of tridiagonal QEPs. In the numerical examples we compare the average number of iterations, the computational time and the accuracy of the computed eigenvalues. As a measure of the accuracy we use the maximum relative error

$$\max_{i=1,\dots,2n} \frac{|\widetilde{\lambda}_i - \lambda_i|}{|\lambda_i|},$$

where λ_i is the exact eigenvalue computed either analytically or using variable precision in Mathematica 5. For all tridiagonal QEPs in this section we tested all three algorithms for the evaluation of the derivative of the determinant. As the choice has almost no effect on the accuracy and the number of needed steps, we include only the results for the fastest method, the three-term recurrences.

For comparison we also applied the Lapack [1] routine ZGGEV to the linearized generalized eigenvalue problem

$$\begin{bmatrix} 0 & K \\ K & C \end{bmatrix} z = \lambda \begin{bmatrix} K & 0 \\ 0 & -M \end{bmatrix} z. \tag{10.1}$$

The main stopping criterion is the relative size of a correction. We take $\varepsilon = 10^{-15}$ and stop the iteration for λ_j when

$$|\lambda_j^{(k+1)} - \lambda_j^{(k)}| \le |\lambda_j^{(k+1)}|\varepsilon.$$

Another stopping criterion for Laguerre's method are different inertias of $Q(\lambda_j^{(k+1)})$ and $Q(\lambda_j^{(k)})$. In the Ehrlich–Aberth method we use a heuristic that stops the iteration once the large majority of the eigenvalues has converged and the corrections for the remaining eigenvalues stop becoming smaller.

In both methods one step (an iteration for one eigenvalue approximation) has linear time complexity. If we compare the number of operations needed for (8.1) and (9.1) and for the three-term recurrences in section 4, then we can observe that one step of Laguerre's method is more expensive and is roughly equivalent to 1.8 Ehrlich–Aberth steps.

Example 10.1. In the first numerical example we use random tridiagonal matrices, where the elements are uniformly distributed in such intervals that the obtained QEP is hyperbolic. For the matrices M and K, the diagonal and codiagonal elements are uniformly distributed in $[0.5, 1]$ and $[0, 0.1]$, respectively. The diagonal and codiagonal elements of the matrix C are uniformly distributed in $[4, 5]$ and $[0, 0.5]$, respectively.

The numerical results are presented in Table 10.1. In the first two columns are the results for the Ehrlich–Aberth method; in the first column we use real arithmetic while in the second column we use complex perturbations and complex arithmetic.

[1] http://www-lp.fmf.uni-lj.si/plestenjak/papers.htm

TABLE 10.1

The average number of iterations in the last divide-and-conquer step, the computational time, and the maximum relative error of the computed eigenvalues in Example 10.1.

n	Ehrlich-Aberth $\mathbb{R}$	Ehrlich-Abert $\mathbb{C}$	Laguerre-bisection	ZGGEV
Average number of iterations in the last D&C				
100	1.9	1.9	1.9	
200	1.8	1.8	2.1	
400	1.6	1.6	1.2	
800	1.6	1.5	1.3	
Time in seconds				
100	0.02	0.03	0.02	0.60
200	0.05	0.13	0.06	5.02
400	0.13	0.39	0.23	52.95
800	0.48	1.48	0.83	684.63
Maximum relative error				
100	5e-15	4e-16	5e-15	5e-14
200	5e-15	4e-16	5e-15	9e-14
400	5e-15	4e-16	5e-15	1e-13
800	5e-15	4e-16	5e-15	1e-13

Complex perturbations increase the computational time for one iteration but in some cases (see, e.g., Example 10.2), where we have multiple or close eigenvalues, we might have faster convergence. In the third column are the results for Laguerre's method, and in the last column are the results for the Lapack routine ZGGEV applied to the linearized generalized eigenvalue problem (10.1) of size $2n$. The cost of ZGGEV, which is not optimized for block tridiagonal matrices, is $\mathcal{O}(n^3)$, compared to $\mathcal{O}(n^2)$ for the methods presented in this paper. Because of that ZGGEV is slower than the presented methods even for a moderate size of matrices.

We tested the methods on matrix dimensions from 100 to 800. The results in Table 10.1 are organized in three parts. In the upper part is the average number of iterations in the last divide-and-conquer step. For Laguerre's method we count bisection steps as well. As the dimension of the matrices increases, better the eigenvalues of $Q_0(\lambda)$ approximate the eigenvalues of $Q(\lambda)$ and fewer iterations are needed in the final phase. The middle part in Table 10.1 contains the computational times in seconds. One can see that although Laguerre's method needs fewer iterations, it runs slower than the Ehrlich–Aberth method which does not compute the second derivatives. In the lower part of the table are the maximum relative errors of the computed eigenvalues. In this example all methods perform well and give small relative errors. The maximum condition number of the eigenvalues (we use the condition number defined in [6] and implemented in MATLAB 7.0 routine `polyeig`) in Example 10.1 is of order 1.

Example 10.2. In this example we use matrices with constant diagonals and codiagonals, such that the QEP is hyperbolic. We take $M = \text{tridiag}(0.1, 1, 0.1)$, $C = \text{tridiag}(0.5, 5, 0.5)$, and $K = \text{tridiag}(0.2, 1, 0.2)$. For such problem the eigenvalues can be computed analytically. All eigenvalues are simple, but we can expect problems in the divide-and-conquer approach because the eigenvalues of Q_0 appear in pairs. The eigenvalues are not sensitive, the maximum condition number for the eigenvalues in this example is of order 1.

Numerical results, organized in the same way as in Example 10.1, are presented in Table 10.2. We can see that the number of iterations is larger than in Example 10.1.

TABLE 10.2
The average number of iterations in the last divide-and-conquer step, the computational time, and the maximum relative error of the computed eigenvalues in Example 10.2.

n	Ehrlich-Aberth $\mathbb{R}$	Ehrlich-Abert $\mathbb{C}$	Laguerre-bisection	ZGGEV
Average number of iterations in the last D&C				
100	19.9	18.5	5.8	
200	19.6	17.6	5.7	
400	19.0	17.1	5.7	
800	18.5	16.5	5.6	
Time in seconds				
100	0.08	0.22	0.03	0.64
200	0.33	0.92	0.13	5.33
400	1.30	3.70	0.53	58.42
800	5.17	14.59	2.09	865.13
Maximum relative error				
100	3e-15	5e-16	3e-15	6e-15
200	3e-15	5e-16	3e-15	5e-15
400	3e-15	6e-16	3e-15	2e-14
800	3e-15	6e-16	3e-15	3e-15

TABLE 10.3
The average number of iterations in the last divide-and-conquer step, the computational time, and the maximum relative error of the computed eigenvalues in Example 10.3.

	QEP 1					QEP 2				
	ZGGEV		Ehrlich–Aberth $\mathbb{C}$			ZGGEV		Ehrlich–Aberth $\mathbb{C}$		
n	Time	Error	Time	Avg. iter	Error	Time	Error	Time	Avg. iter	Error
100	0.75	4e-13	0.03	1.9	1e-13	0.59	7e-15	0.23	20.1	2e-15
200	6.16	3e-12	0.11	1.7	9e-15	5.23	4e-14	1.02	19.5	2e-15
400	67.09	4e-12	0.39	1.6	4e-14	46.64	6e-14	4.09	18.9	4e-15

The Ehrlich–Aberth method has problems with close initial approximations. In this case Laguerre's method gives the best performance.

Example 10.3. The Ehrlich–Aberth method in complex arithmetic can also be applied to the QEPs that are nonhyperbolic and where the eigenvalues might be complex. The interlacing property of the eigenvalues of Q_0 and Q is no longer true, but we can still expect that the eigenvalues of Q_0 are good initial approximations to the eigenvalues of Q. When the solutions are complex, Laguerre's method is not globally convergent anymore, and without the inertia and the interlacing property we have no guarantee that the method returns all the eigenvalues.

For the first nonhyperbolic QEP we use random symmetric tridiagonal matrices. The diagonal elements of matrices M, C, and K are uniformly distributed in $[0, 1]$. The codiagonal elements of matrices M, C, and K are uniformly distributed in $[0, 0.1]$, $[0, 0.5]$, and $[0, 0.2]$, respectively. The maximum condition number of the eigenvalues is of order 10^2 and this reflects in slightly larger errors than in the previous examples.

For the second nonhyperbolic QEP we use the example from [21], where $M = \text{tridiag}(0.1, 1, 0.1)$, $C = \text{tridiag}(-3, 9, -3)$, and $K = \text{tridiag}(-5, 15, -5)$. All eigenvalues are simple, but the eigenvalues of Q_0 are double. The maximum condition number of the eigenvalues is of order 10^2.

Numerical results in Table 10.3 show that the Ehrlich–Aberth method can be applied to such QEPs.

TABLE 10.4
The average number of iterations in the last divide-and-conquer step for the banded quadratic eigenvalue problem with matrices of dimension n and bandwidth p from Example 10.5.

p	$n = 50$	$n = 100$	$n = 200$
1	3.9	2.9	2.3
2	5.8	4.2	3.3
3	6.2	5.3	4.4
4	6.4	5.9	5.3
5	9.3	6.4	6.4

Example 10.4. We consider the second-order model of vibration of a rotating axel in a power plant from [2]. We have a second-order differential equation

$$M\ddot{z} + C\dot{z} + Kz = Du,$$

where M, C, and K are tridiagonal symmetric matrices of dimension $n = 211$. After the Fan, Lin, and Van Dooren scaling [9], the norms of the matrices are $\|M\| = 2$, $\|C\| = 8 \cdot 10^{-6}$, and $\|K\| = 2$. One eigenvalue of the corresponding QEP is 0, which makes the resulting system neither observable nor detectable. The largest real part of the remaining nonzero eigenvalues is $\rho = -0.01626759$.

If we apply the Ehrlich–Aberth method, then the relative error of the computed ρ is of order 10^{-12}. The average number of iterations in the last divide-and-conquer step is 6.8. If we use the linearization (10.1) and ZGGEV, then the relative error of the computed ρ is of order 10^{-9}. If we reduce the linearized 422×422 problem into a 421×421 problem for the nonzero eigenvalues as in [2], then the relative error of the computed ρ falls to 10^{-10}. This example shows that we can get more accurate results without a linearization. The eigenvalues in this example have condition numbers of orders from 1 up to 10^6.

Example 10.5. The above ideas can be extended to QEPs with banded matrices as well. We can apply the Ehrlich–Aberth method as long as we have an efficient method for the computation of the characteristic polynomial and its derivative. For banded matrices these values can be computed in a linear time using the algorithm based on the LU factorization from section 6.

As in the previous examples, the initial approximations are obtained by the divide-and-conquer scheme. The matrices M, C, and K are represented as 2×2 block matrices and then the approximations are obtained by a recursive application of the method to the diagonal block subproblems.

The following example was done in MATLAB 7.0. We take three matrices of dimension n with normally distributed elements: `M=randn(n)`, `C=randn(n)`, and `K=randn(n)`, set $m_{ij} = c_{ij} = k_{ij} = 0$ for $|i - j| > p$, where p is the bandwidth, and apply a MATLAB implementation of the Ehrlich–Aberth method.

As expected, the results in Table 10.4 show that the average number of iterations in the last divide-and-conquer step does increase with the bandwidth. However, for a small bandwidth, one step is performed in linear time and the results in Table 10.4 show that the Ehrlich–Aberth method can be considered as an alternative for the banded quadratic eigenvalue problems. For all combinations of p and n in Table 10.4 the maximum relative error of the computed eigenvalues is below 10^{-14} and smaller than the error obtained by the MATLAB function `polyeig` that applies QZ to the linearized problem.

11. Conclusions. We have presented two numerical methods for the tridiagonal hyperbolic QEP that use the divide-and-conquer approach. Both methods can be easily parallelized. Laguerre's method and the bisection require hyperbolicity, while the Ehrlich–Aberth method might be applied to more general problems—for instance, non-Hermitian tridiagonal quadratic eigenvalue problems, tridiagonal polynomial eigenvalue problems, banded polynomial eigenvalue problems, and others. In these applications, the algorithm based on the LU factorization might be used for an efficient computation of the derivative of the determinant.

Let us mention that at the moment there are no methods for transforming a general QEP to a tridiagonal form. In future, this might change with structure preserving transformations (SPT) [10].

Acknowledgments. The author would like to thank Daniel Kressner for providing the generalization of the QR factorization approach in section 5. The author would also like to thank Dario Bini and Françoise Tisseur for useful suggestions and comments. The author is also grateful to the referees for careful reading of the paper and several helpful comments.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK User's Guide*, 3rd ed., SIAM, Philadelphia, 1999.

[2] J. Abels and P. Benner, *CAREX—A Collection of Benchmark Examples for Continuous-Time Algebraic Riccati Equations (version 2.0)*, SLICOT Working Note 1999-14, 1999.

[3] D. A. Bini, *Numerical computation of polynomial zeros by means of Aberth's method*, Numer. Algorithms, 13 (1996), pp. 179–200.

[4] D. A. Bini, L. Gemignani, and F. Tisseur, *The Ehrlich–Aberth method for the nonsymmetric tridiagonal eigenvalue problem*, SIAM J. Matrix Anal. Appl., 27 (2005), pp. 153–175.

[5] Z. Bohte, *Calculation of the derivative of the determinant*, Obzornik Mat. Fiz., 28 (1981), pp. 33–50.

[6] J.-P. Dedieu and F. Tisseur, *Perturbation theory for homogeneous polynomial eigenvalue problems*, Linear Algebra Appl., 358 (2003), pp. 71–94.

[7] I. S. Dhillon, *Reliable computation of the condition number of a tridiagonal matrix in $O(n)$ time*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 776–796.

[8] R. J. Duffin, *A minimax theory for overdamped networks*, J. Rational Mech. Anal., 4 (1955), pp. 221–233.

[9] H.-Y. Fan, W.-W. Lin, and P. Van Dooren, *Normwise scaling of second order polynomial matrices*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 252–256.

[10] S. D. Garvey, *Structure Preserving Transformations for Linear Dynamic Systems*, available at http://www.nottingham.ac.uk/∼eazsg/SPT/index.htm.

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

[12] D. Kressner, *personal communication*, 2003.

[13] P. Lancaster, *Quadratic eigenvalue problems*, Linear Algebra Appl., 150 (1991), pp. 499–506.

[14] K. Li, *Durand–Kerner root-finding method for the generalized tridiagonal eigenproblem*, Missouri J. Math. Sci., (1999), pp. 33–43.

[15] K. Li and T. Y. Li, *An algorithm for symmetric tridiagonal eigenproblems: Divide and conquer with homotopy continuation*, SIAM J. Sci. Comput., 14 (1993), pp. 735–751.

[16] T. Y. Li and Z. Zeng, *The Laguerre iteration in solving the symmetric tridiagonal eigenproblem, revisited*, SIAM J. Sci. Comput., 15 (1994), pp. 1145–1173.

[17] K. Li, T. Y. Li, and Z. Zeng, *An algorithm for generalized symmetric tridiagonal eigenvalue problems*, Numer. Algorithms, 8 (1994), pp. 269–291.

[18] A. S. Markus, *Introduction to the Spectral Theory of Polynomial Operator Pencils*, AMS, Providence, RI, 1988.

[19] V. Pereyra and G. Scherer, *Eigenvalues of symmetric tridiagonal matrices: A fast, accurate and reliable algorithm*, J. Inst. Math. Appl., 12 (1973), pp. 209–222.

[20] D. Sapagovene, *The Sturm sequence for a nonlinear algebraic eigenvalue problem*, Differencial'nye Uravnenija i Primenen., 16 (1976), pp. 87–95.
[21] F. Tisseur and K. Meerbergen, *The quadratic eigenvalue problem*, SIAM Rev., 43 (2001), pp. 235–286.
[22] J. Wilkinson, *Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 4, pp. 1173–1198

ACCURATE FACTORIZATION AND EIGENVALUE ALGORITHMS FOR SYMMETRIC DSTU AND TSC MATRICES*

MARÍA JOSÉ PELÁEZ† AND JULIO MORO†

Abstract. Two algorithms are presented which compute, with small componentwise relative error, a block LDL^T factorization of symmetric matrices belonging to two classes of matrices: diagonally scaled totally unimodular (DSTU) and total signed compound (TSC) matrices. This accuracy is achieved by taking advantage of some special properties of such structures in order to avoid subtractions throughout the factorization process. Once an accurate block LDL^T decomposition is available, it is proved that one can easily obtain an accurate symmetric rank-revealing decomposition, which is the starting point for algorithms computing with high relative accuracy the eigenvalues and eigenvectors of arbitrary, possibly indefinite, symmetric matrices. This proves that eigenvalues and eigenvectors of symmetric DSTU and TSC matrices can be computed with high relative accuracy.

Key words. symmetric eigenproblem, high relative accuracy, rank-revealing decomposition, structured matrices

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

DOI. 10.1137/050631537

1. Introduction.

It is well known that, given an arbitrary real square matrix, conventional, general-purpose eigenvalue algorithms like QR or divide-and-conquer can only guarantee full accuracy in the computation of the eigenvalues with largest absolute value. If one is interested in obtaining *all* eigenvalues with correct sign and correct leading digits, then more specifically devised algorithms are needed. Given a matrix $A \in \mathbb{R}^{n\times n}$, the best one can expect is that all computed eigenvalues $\widehat{\lambda}_i$ and corresponding eigenvectors $\widehat{q}_i$, $i = 1, \ldots, n$, satisfy

$$\begin{aligned} \frac{|\widehat{\lambda}_i - \lambda_i|}{|\lambda_i|} &= O(\kappa\epsilon), \\ \Theta(\widehat{q}_i, q_i) &= \frac{O(\kappa\epsilon)}{\operatorname{relgap}(\lambda_i)}, \end{aligned} \tag{1}$$

where λ_i, q_i and $i = 1, \ldots, n$, are, respectively, the exact eigenvalues and eigenvectors of A; ϵ is the machine precision; $O(\cdot)$ is the customary big-oh Landau notation; $\Theta(\cdot, \cdot)$ stands for the angle between two n-vectors; $\operatorname{relgap}(\lambda_i)$ is the usual relative gap $\operatorname{relgap}(\lambda_i) = \min\{\min_{j\neq i} \frac{|\lambda_j - \lambda_i|}{|\lambda_i|}, 1\}$, $i = 1, \ldots, n$; and κ is some constant of moderate size, eventually depending on the dimension n but independent of the condition number of A. The algorithms achieving these, or some slightly relaxed, error bounds are the so-called *high relative accuracy algorithms* for the eigenvalue problem. A similar definition can be given for high relative accuracy algorithms for the singular value decomposition (SVD), replacing eigenvalues by singular values.

Several high relative accuracy algorithms are known so far, both for the SVD and for the standard eigenvalue problem. Their scope, however, is limited, in the

*Received by the editors May 17, 2005; accepted for publication (in revised form) by J. Barlow October 9, 2006; published electronically December 18, 2006. The research conducted in this paper was supported by Spanish Ministerio de Ciencia y Tecnología grant BFM-2003-00223.
http://www.siam.org/journals/simax/28-4/63153.html
†Departamento de Matemáticas, Universidad Carlos III de Madrid, 28911 Leganés, Spain (mpelaez@math.uc3m.es, jmoro@math.uc3m.es).

sense that each algorithm is specifically designed for a particular class of matrices, taking advantage of the structure properties specific to that class in order to preserve the accuracy of the computation. Some such classes for the SVD are, for instance, bidiagonal matrices [8, 17], matrices of the form BD with D diagonal and B well-conditioned [10, 16, 23], acyclic, Cauchy and Vandermonde matrices [11], unit-displacement-rank matrices [7], or weakly diagonally dominant M-matrices [9, 24]. Some of these classes can be grouped into a larger class of matrices for which it is possible to compute accurately a rank-revealing decomposition [11]. For the eigenvalue problem, some classes allowing accurate eigendecomposition are symmetric scaled diagonally dominant [4]; symmetric positive definite [10]; symmetric tridiagonal [21, 12, 13]; symmetric Cauchy and Vandermonde [15]; matrices of the form $H = BD$, where D is diagonal and B is well conditioned [30]; and symmetric indefinite matrices allowing for an accurate initial factorization [25, 27, 14]. So far, the only class of nonsymmetric matrices whose eigenvalues can be computed to high relative accuracy is the class of totally nonnegative matrices [22].

Among all these different methods we will concentrate on a family of algorithms [11, 14, 25] which proceed in two stages: In the first stage, the algorithm computes an initial factorization of the matrix. Then an appropriately chosen Jacobi-type algorithm is applied to the factors. To achieve the accuracy bounds (1), both stages must be performed accurately enough. The accuracy of the second stage is usually ensured once and for all through a detailed error analysis, valid for any factorized matrix. Once this is done, the accuracy of the overall algorithm depends entirely on the accuracy of the preprocessing factorization in the first stage. In other words, *the classes of matrices such that these algorithms compute all its eigenvalues and eigenvectors to high relative accuracy become those classes for which it is known how to compute a sufficiently accurate initial factorization.*

In the case of the SVD, for instance, several classes of matrices were identified in [11] such that special versions of Gaussian elimination with complete pivoting (GECP), conveniently adapted to each class, lead to accurate *nonsymmetric* factorizations of the form $A = X\Delta Y^T$. Two of these classes are the *diagonally scaled totally unimodular* (DSTU) matrices and *total signed compound* (TSC) matrices (see sections 3.1 and 3.2 below for definitions). Our main contribution in this paper is to prove that, for both DSTU and TSC structures, the subclass of *symmetric* matrices allows for the computation of *symmetric* factorizations of the form $A = X\Delta X^T$ in a sufficiently accurate way to compute eigenvalues and eigenvectors with the accuracy (1). Therefore, for any symmetric matrix which is either DSTU or TSC, all its eigenvalues and eigenvectors can be computed to high relative accuracy. The leading idea of the factorization algorithms we propose is, as in [11], to take advantage of the special properties of each structure in order to completely avoid subtraction throughout the factorization process.

Two high relative accuracy algorithms are available to compute eigenvalues and eigenvectors of general, possibly indefinite, symmetric matrices: the *J-orthogonal algorithm* [29, 25] and the *signed SVD algorithm* [14]. None of them, strictly speaking, has error bounds of the form (1). For the J-orthogonal method, the constant κ in (1) is the maximum of the condition numbers of some intermediate matrices produced by the algorithm, and these condition numbers could be, in principle, arbitrarily large. The signed SVD method, on the other hand, has an error bound for eigenvectors of the form (1), but with a potentially smaller quantity, relgap*, in the denominator instead of the usual relative gap (see (9) below). Nevertheless, both algorithms are

able in practice to compute eigenvalues and eigenvectors to high relative accuracy.

Both algorithms begin by initially factorizing the matrix, although in a slightly different way. To be more precise, we begin by defining *rank-revealing decompositions.*

DEFINITION 1. *Given $A \in \mathbb{R}^{m\times n}$ with $m \geq n$, a rank-revealing decomposition (RRD) of A is any factorization $A = X\Delta Y^T$ such that $X \in \mathbb{R}^{m\times r}$, $Y \in \mathbb{R}^{n\times r}$, $\Delta \in \mathbb{R}^{r\times r}$ for $r \leq \min\{m, n\}$, where Δ is diagonal and nonsingular and both matrices X, Y are well-conditioned.*

For instance, the SVD is an RRD factorization, but others can be obtained, for instance, via Gaussian elimination with complete pivoting (GECP), QR with complete pivoting, or, as we will see, via the diagonal pivoting method.

The signed SVD method begins by computing an RRD, either symmetric with $X = Y$, or nonsymmetric with $X \neq Y$.[1] Since in our case the matrix A is symmetric, preservation of structure advises keeping the symmetry in the factorization. Therefore, we restrict ourselves in this paper to the analysis of *symmetric* RRDs of the form

$$A = X\Delta X^T. \tag{2}$$

The J-orthogonal algorithm, on the other hand, begins by computing a so-called *symmetric indefinite factorization*

$$PAP^T = GJG^T, \tag{3}$$

where P is a permutation matrix, J is square diagonal with diagonal elements ± 1 and G has full column rank. Although this is not exactly an RRD, its computation is equivalent to computing the symmetric RRD above, since it suffices to scale Δ on both sides with $|\Delta|^{-1/2} = \mathrm{diag}(|\Delta_{ii}|^{-1/2})$ to obtain $A = GJG^T$, where $G = X|\Delta|^{1/2}$ and $J = |\Delta|^{-1/2}\Delta|\Delta|^{-1/2}$. Therefore, we concentrate in what follows in obtaining a symmetric RRD (2).

The way we obtain the symmetric RRD is via the so-called *block LDL^T factorization*

$$PAP^T = LDL^T, \tag{4}$$

where P is a permutation matrix, L is unit lower triangular, and D is block-diagonal with 1×1 and 2×2 diagonal blocks. Of course, this is not an RRD, since D is not diagonal. However, it suffices to orthogonally diagonalize the 2×2 blocks in D via Givens rotations to obtain an RRD, a procedure which was introduced in [26, 27] to obtain symmetric indefinite decompositions: let $Q \in \mathbb{R}^{n\times n}$ be an orthogonal, block-diagonal matrix conformal to D, each 2×2 diagonal block of Q being the Givens rotation used to diagonalize the corresponding diagonal block of D. Then $A = X\Delta X^T$ with

$$X = P^T LQ \qquad \text{and} \qquad \Delta = Q^T DQ. \tag{5}$$

Notice that L and X have the same condition number in any unitarily invariant norm.

[1] There might be advantages in computing a nonsymmetric RRD of a symmetric matrix, due to the additional freedom in pivoting: a class of structured symmetric matrices might allow for accurate *nonsymmetric* RRDs but not for accurate symmetric ones. Whether such a class exists, however, is still an open question.

One can easily show,[2] adapting the proof of [11, Theorem 2.1] from the SVD context to the eigenvalue problem, that the symmetric RRD (2) determines the eigendecomposition to high relative accuracy, i.e., that, as stated in [11] for the SVD, having any symmetric RRD is as good as having an eigendecomposition, because any small change (in the sense given by (7) below) in the factors of the RRD produces small changes in the eigenvalues and eigenvectors.

Once we have this, the error analyses of both the J-orthogonal and the signed SVD method guarantee the accuracy of the computed eigenvalues and eigenvectors *only if* the initial factorization is computed accurately enough. Proving this for DSTU ad TSC matrices is our goal in the present paper, but since the error analyses of both methods are very different, the accuracy requirements on the factorizations are also diverse. We will deal in what follows only with the signed SVD method, whose error analysis is more amenable to our approach, but we stress that similar results hold for the J-orthogonal method.

To analyze the accuracy of the computed RRD we proceed in two stages: First, we see how to compute block LDL^T factorizations of symmetric DSTU and TSC matrices with *componentwise* small relative error, i.e., if $\widehat{L}$ and $\widehat{D}$ are the factors computed in floating point arithmetic and L and D are the exact factors, it will be shown that

$$|\widehat{l}_{ij} - l_{ij}| = O(\epsilon)|l_{ij}|, \qquad |\widehat{d}_{ij} - d_{ij}| = O(\epsilon)|d_{ij}| \tag{6}$$

for every $i, j \in \{1, \ldots, n\}$. To prove this it is enough to show that *no subtraction is ever performed throughout the factorization process.* Products, quotients, square roots, and sums of quantities of like sign are harmless operations from the point of view of producing large forward errors. The only possible source of forward instability is cancellation, and we will rule it out by avoiding subtraction (even if the subtracted quantities are not close to each other).

In a second stage, we will show that the RRD obtained from the block LDL^T factorization via Givens diagonalization, as in (5), satisfies the requirements which ensure the accuracy (1). According to the error analysis in [14], these requirements are that the factor Δ in (2) be computed with small *componentwise* relative errors, and the factor X be computed with small *normwise* relative error in any norm, i.e.,

$$\|\widehat{X} - X\| = O(\epsilon)\|X\|, \qquad |\widehat{\Delta}_{ii} - \Delta_{ii}| = O(\epsilon)|\Delta_{ii}|, \quad i = 1, \ldots, n, \tag{7}$$

if $\widehat{X}, \widehat{\Delta}$ are the factors computed in floating point arithmetic and X, Δ are the exact ones (actually, we will prove a sharper bound for X in Theorem 6). All this leads to the conclusion that *all eigenvalues and eigenvectors of symmetric DSTU and TSC matrices can be computed with high relative accuracy via the signed SVD method, provided the initial RRD is computed with these special factorization algorithms.* To be more precise, the eigenvalues are computed with an error of the form (1), where κ is given by

$$\kappa = \kappa(R')\kappa(X), \tag{8}$$

[2]This has been done in [15]: if $A = XDX^T$ and $\widetilde{A} = \widetilde{X}\widetilde{D}\widetilde{X}^T$ are RRDs of the symmetric matrices A and $\widetilde{A}$, and both the normwise relative error $\|\widetilde{X} - X\|/\|X\|$ in X and the componentwise relative error $|\widetilde{D}_{ii} - D_{ii}|/|D_{ii}|$ in D are bounded by a quantity β smaller than 1, then, setting $\eta = \beta(2 + \beta)\kappa(X)$, the relative error in the eigenvalues is bounded by $O(\eta)$ and the sine of the canonical angles between the eigenvectors of A and $\widetilde{A}$ is bounded by $O(\eta)$ divided by the relative gap (see [15, section 2] for more details).

$\kappa(\cdot)$ denotes the condition number in the two-norm, X is the nondiagonal factor in (2), and R' is the best conditioned row diagonal scaling of the triangular factor R of a QR factorization with column pivoting of the product $X\Delta$. Since it was proved in [11, Theorem 3.2] that $\kappa(R')$ is at most of order $O(n^{3/2}\kappa(X))$, we see that, up to a moderate constant, the factor κ is of the order of the condition number of the factor X in the RRD. Therefore, it is important to guarantee that $\kappa(X)$ is moderate. We will do this for symmetric DSTU matrices in Theorem 3, proving that $\kappa(X) = O(n^2)$. No such bound is available so far for TSC matrices.

As for the eigenvectors, they are computed with an error of the form (1) but replacing the usual relative gap with

$$\text{relgap}^*(|\lambda_i|) = \min\left\{ \min_{\substack{j\in\mathcal{S}\\ j\neq i}} \left| \frac{|\lambda_j| - |\lambda_i|}{\lambda_i} \right|, 1 \right\}, \tag{9}$$

where the index set $\mathcal{S}$ is equal to $\{1,\ldots,n\}$ unless the eigenvalue, say λ_{j_0}, whose absolute value is closest to $|\lambda_i|$ has opposite sign to λ_i. In that case, $\mathcal{S}$ is obtained from $\{1,\ldots,n\}$ by removing j_0 and the index k of any other eigenvalue within a relative distance of order $O(\kappa\epsilon)$ of λ_{j_0}.

The paper is organized as follows. We begin in section 2 with a brief review of some basic properties of the block LDL^T factorization, which will be needed later on. Then, we show in section 3 how to achieve accurate block LDL^T decompositions of symmetric DSTU and TSC matrices. The factorization algorithm for $n \times n$ DSTU matrices has a computational cost of order $O(n^3)$, while the one for TSC matrices has a worst-case cost of $O(n^4)$. Section 4 is devoted to showing that, given any block LDL^T factorization satisfying (6), all further manipulations required to derive (2) from (4) do not spoil the accuracy. More precisely, we show that the componentwise accuracy is preserved in Δ, and it is tranformed at worse in columnwise accuracy for X. Some numerical experiments are presented in section 5 which confirm the high accuracy of the proposed algorithms. Finally, we collect in Appendix A the proofs of the results in section 2.

We end this introduction with a brief comment on singular matrices: we will assume that all matrices A under examination are nonsingular. If A is singular, the number of zero eigenvalues is determined from any RRD satisfying (7), and the signed SVD method can be enhanced to compute the null vectors using a complete QR factorization with complete pivoting. However, this is out of the scope of our error analysis in this paper.

2. Block LDL^T factorizations of symmetric matrices. One of the possible symmetric analogues of the LU decomposition is the block LDL^T decomposition (4). Any symmetric matrix admits such a factorization [18, Chapter 11], and the most common procedure to compute it is the *diagonal pivoting method*: it begins by choosing a permutation matrix P, an integer $s = 1$ or $s = 2$, and an $s \times s$ nonsingular pivot matrix E such that

$$PAP^T = \begin{pmatrix} E & C^T \\ C & B \end{pmatrix},$$

so

$$PAP^T = \begin{pmatrix} I_s & 0 \\ CE^{-1} & I_{n-s} \end{pmatrix} \begin{pmatrix} E & 0 \\ 0 & B - CE^{-1}C^T \end{pmatrix} \begin{pmatrix} I_s & E^{-1}C^T \\ 0 & I_{n-s} \end{pmatrix}. \tag{10}$$

The block LDL^T factorization of A follows from simply repeating this same process on the succesive $(n-s)\times(n-s)$ Schur complements $B-CE^{-1}C^T$. The whole process costs $n^3/3$ arithmetic operations plus the cost of determining the permutations.

Several symmetric strategies are available for choosing the pivot matrix E, analogous to either partial, complete, or rook pivoting in LU. Since our final goal is an RRD, we will mostly use the Bunch–Parlett pivoting strategy [3], a symmetric analogue of complete pivoting which usually produces well-conditioned factors L (see [2] for a detailed analysis of element growth of the L factor). This pivoting strategy can be summarized as follows:

$$
\begin{array}{ll}
 & \alpha = (1+\sqrt{17})/8 \ 0.64 \\
 & \mu_0 = \max_{i,j}|a_{ij}| =: |a_{pq}| \\
 & \mu_1 = \max_i |a_{ii}| =: |a_{rr}| \\
(11) & \textbf{If } \mu_1 \geq \alpha\mu_0 \textbf{ then} \\
 & \qquad \text{choose } E = [a_{rr}] \text{ as } 1\times 1 \text{ pivot} \\
 & \textbf{else} \\
 & \qquad \text{choose } E = \begin{pmatrix} a_{pp} & a_{pq} \\ a_{pq} & a_{qq} \end{pmatrix} \text{ as } 2\times 2 \text{ pivot}
\end{array}
$$

Any 2×2 pivot E chosen with this strategy is a symmetric indefinite, well-conditioned matrix whose condition number is bounded by $(1+\alpha)/(1-\alpha)\approx 4.6$ in the 2-norm. The value $(1+\sqrt{17})/8$ of the constant α is chosen to ensure that the growth factor corresponding to two successive 1×1 pivots equals the growth factor corresponding to a 2×2 pivot.

It is well known that any final or intermediate value computed by Gaussian elimination with any pivoting strategy is either a minor or a quotient of minors of the original matrix (see Lemma 5.1 in [11]). This is a consequence of the properties of Schur complements: given any matrix A (eventually nonsymmetric), partitioned as

$$
A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},
$$

with A_{11} square and nonsingular, the *Schur complement*[3] of A_{11} in A is

$$
\mathcal{C} = A_{22} - A_{21}A_{11}^{-1}A_{12}. \tag{12}
$$

One of the simplest properties of $\mathcal{C}$ is that $\det A = \det A_{11}\det\mathcal{C}$ or, equivalently, that $\det\mathcal{C} = \det A/\det A_{11}$. Hence, $\det\mathcal{C}$ is a quotient of minors of A. To prove that every intermediate quantity in the block LDL^T factorization (4) is also a quotient of minors of the original matrix, some properties of Schur complements are needed. Due to their technical character, all proofs are postponed to Appendix A. We will use MATLAB notation to state the results; i.e., $A([r_1,\dots,r_p],[c_1,\dots,c_p])$ denotes the $p\times p$ submatrix of A containing the elements in rows $r_1,\dots,r_p$ and columns $c_1,\dots,c_p$. We also abbreviate as $1:k$ the list of all integers from 1 to k.

LEMMA 1. *Let $A\in\mathbb{R}^{n\times n}$, let $k<n$ and let A_k be the upper left $k\times k$ principal submatrix of A. Then*

[3]In order not to complicate the notation, the definition (12) and Lemma 1 are given in terms of Schur complements of the upper left leading principal submatrix, which is the only one we will employ below. Of course, all results hold true if A_{11} is replaced by any square nonsingular submatrix of A.

(a) *the* (i,j) *element of the* $(n-k)\times(n-k)$ *Schur complement* $\mathcal{C}_k$ *of* A_k *in* A *is given by*

$$\mathcal{C}_k(i,j) = \frac{\det A([1:k,k+i],[1:k,k+j])}{\det A([1:k],[1:k])} \tag{13}$$

for each $i,j \in \{1,\dots,n-k\}$*;*

(b) *for any* $s \le n-k$*, the minor of* $\mathcal{C}_k$ *containing rows* $i_1 < \cdots < i_s$ *and columns* $j_1 < \cdots < j_s$ *is given by*

$$\det \mathcal{C}_k([i_1,\dots,i_s],[j_1,\dots,j_s]). \tag{14}$$
$$= \frac{\det A([1:k,k+i_1,\dots,k+i_s],[1:k,k+j_1,\dots,k+j_s])}{\det A[1:k],[1:k]}.$$

In particular, any minor of $\mathcal{C}_k$ *is a quotient of minors of* A*.*

As a consequence of Lemma 1, we prove our previous claim.

THEOREM 1. *Let* A *be a real symmetric matrix and let* $PAP^T = LDL^T$ *be a block* LDL^T *factorization of* A *as described in* (4)*, obtained using any pivoting strategy. Then every entry of* L *or* D *is either zero or a quotient of minors (or just a minor) of* A*.*

It is interesting to observe at this point that the block LDL^T factorization does not enjoy all the good properties of the LDU decomposition obtained from Gaussian elimination. For instance, it is not true, as it is for Gaussian elimination, that every minor of L is a quotient of minors of A: if we consider the symmetric matrix

$$A = \begin{pmatrix} 13 & 39 & 65 \\ 39 & 128 & 274 \\ 65 & 274 & 903 \end{pmatrix}$$

which can be factorized as $A = LDL^T$ with

$$L = \begin{pmatrix} 1 & & \\ 3 & 1 & \\ 5 & 7 & 1 \end{pmatrix}, \qquad D = \left(\begin{array}{c|cc} 13 & & \\ \hline & 11 & 2 \\ & 2 & 11 \end{array}\right),$$

one can check that

$$\det L([2,3],[1,2]) = \begin{vmatrix} 3 & 1 \\ 5 & 7 \end{vmatrix} = 16$$

is not a quotient of minors of A.

We finish this section with explicit formulas for the elements of L and D as quotients of minors of A. These formulas, which will be needed later to recompute the entries of L, are, to our knowledge, new in the literature. They are an extension of classical formulas for Gaussian elimination (see, e.g., [20, section 1.4]). Such classical formulas no longer hold in our case, since, as shown by the example above, not every minor of L is a quotient of minors of A, and this is an essential ingredient of the proofs for Gaussian elimination. For the sake of simplicity, the formulas are written with no reference to the pivoting permutations, but the same result holds for the factorization (4), appropriately renaming rows and columns according to the permutation matrix P.

LEMMA 2. *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix factorized as $A = LDL^T$ with $L \in \mathbb{R}^{n\times n}$ unit lower triangular and $D \in \mathbb{R}^{n\times n}$ block-diagonal with 1×1 and 2×2 diagonal blocks. Let $D = \operatorname{diag}(D_1, \ldots, D_r)$ with $D_k \in \mathbb{R}^{s_k\times s_k}$, $s_k = 1$ or 2, $k = 1, \ldots, r$, and partition L conformally as $L = [L_1 \,|\, \cdots \,|\, L_r]$ with $L_k \in \mathbb{R}^{n\times s_k}$. For each $k \in \{1, \ldots, r\}$, set $n_k = s_1 + \ldots + s_k$. Then, for each $k \in \{1, \ldots, r\}$, the elements (i, j) of L with $j \in \{n_k - s_k + 1, \ldots, n_k\}$ are given by*

$$L(i,j) = \frac{\det A([1:j-1, i, j+1:n_k], [1:n_k])}{\det A([1:n_k], [1:n_k])}, \qquad i = n_k + 1, \ldots, n, \tag{15}$$

and if $n_0 = 0$, the elements (i, j) of D with $i, j \in \{n_k - s_k + 1, \ldots, n_k\}$ are given by

$$D(i,j) = \frac{\det A([1:n_{k-1}, i], [1:n_{k-1}, j])}{\det A([1:n_{k-1}], [1:n_{k-1}])}. \tag{16}$$

3. Accurate factorization of symmetric DSTU and TSC matrices. In this section we show that, for both DSTU and TSC matrices, the block LDL^T factorization (4) can be computed with *componentwise* small relative error, as stated in (6). To prove it, we will modify the diagonal pivoting method in such a way that no subtraction is ever performed throughout the factorization process. Since cancellation is the only possible source of forward instability, the overall process will produce a small relative forward error in each single component of L and D.

3.1. Diagonally scaled totally unimodular matrices.

DEFINITION 2. *A matrix Z with integer entries is called* totally unimodular (TU) *if all its minors are -1, 0, or 1. In particular, the entries of Z must be either -1, 0, or 1. A matrix A is* diagonally scaled totally unimodular (DSTU) *if $A = \mathcal{D}_L Z \mathcal{D}_R$, where Z is totally unimodular, and $\mathcal{D}_L$ and $\mathcal{D}_R$ are diagonal matrices.*

The class of TU matrices contains some well-known classes as particular cases (see, e.g., [1, section 2.3]). Some of them are acyclic matrices, finite element matrices from linear mass-spring systems, or reduced node-arc incidence (RNAI) matrices [28]. In our case we are interested only in *symmetric* DSTU matrices, i.e., symmetric matrices A which can be written as

$$A = \mathcal{D} Z \mathcal{D}, \tag{17}$$

where Z is symmetric TU and $\mathcal{D} = \operatorname{diag}(d_1, \ldots, d_n)$ is diagonal. The matrix Z is supposed to be known exactly, but the elements of $\mathcal{D}$ are known only to high relative accuracy.

A first important property of DSTU matrices is that, as can be easily checked, *any Schur complement of a DSTU is still DSTU*. Hence, we may use the special properties of DSTU matrices at any stage of the factorization. The second important property, which is the key to avoid subtraction, is the following.

LEMMA 3. *Any minor of a symmetric DSTU matrix* (17) *is a monomial with coefficients $0, 1$ or -1 in the diagonal elements d_i of $\mathcal{D}$.*

As a consequence of this, any minor of a DSTU matrix is determined to high relative accuracy, and so are the elements of any of its Schur complements since, according to Theorem 1, each of them is just a quotient of minors of the original matrix.

Another interesting property of DSTU matrices is that the 2×2 pivots chosen by the Bunch–Parlett strategy have a very special structure.

LEMMA 4. *Any* 2×2 *pivot choosen by the Bunch–Parlett pivoting strategy on a symmetric DSTU matrix has, at least, one zero entry on the diagonal.*

Proof. A 2×2 pivot is chosen whenever

$$\mu_1 = \max_i |a_{ii}| := |a_{rr}| < \alpha\, |a_{pq}| =: \alpha \max_{i,j} |a_{ij}| = \alpha\, \mu_0.$$

If we suppose that both a_{pp} and a_{qq} are nonzero, then

$$|a_{pp}| \le \mu_1 < \alpha\, |a_{pq}|, \qquad |a_{qq}| \le \mu_1 < \alpha\, |a_{pq}|, \tag{18}$$

so $a_{pq} = d_p d_q z_{pq}$ is nonzero and, consequently, also d_p and d_q are different from zero. This, together with (18), implies that

$$|d_p| < \alpha^2 |d_p|, \tag{19}$$

in contradiction with the fact that $\alpha < 1$. Hence, either $a_{pp} = 0$ or $a_{qq} = 0$. ∎

Notice that inequality (19) is also in contradiction with $\alpha = 1$. Therefore, *Lemma 4 holds even if* $\alpha = 1$, a fact we will need in section 3.1.2 once we slightly modify the pivoting strategy.

3.1.1. Accurate block LDL^T factorization of symmetric DSTU matrices. We distinguish two cases, depending on the size of the pivot chosen at each stage.

- *Case 1: The chosen pivot* $E = [a_{rr}]$ *is* 1×1.
  Notice first that the elements of L computed at this stage are just a quotient
  $$l_{ir} = (CE^{-1})_{ir} = \frac{a_{ir}}{a_{rr}}$$
  of elements of the Schur complement computed in the previous stage. Therefore, l_{ir} is computed with small forward error, provided a_{ir} and a_{rr} are computed with small forward error as well. Computing the elements of D, however, may involve subtraction, since they are given by the formulas
  $$(B - CE^{-1}C^T)_{ij} = a_{ij} - l_{ir} a_{rj}. \tag{20}$$
  According to Theorem 1 and Lemma 3, each of the operands above is either zero or, up to a sign, a product of powers of the diagonal elements of $\mathcal{D}$. Actually, a simple computation shows that each operand is a monomial with coefficients ± 1 or 0 in the two variables d_i and d_j. Hence, (20) can be rewritten as
  $$m_1 = m_2 + m_3$$
  where each operand m_i for $i = 1, 2, 3$ is a monomial with coefficient ± 1 or 0 in the two variables d_i and d_j. There are four possibilities for this arithmetic operation, depending on whether m_2 and m_3 are zero or not. Three of the possibilities give rise to no operation at all. The fourth possibility, in which both m_2 and m_3 are nonzero, can have only $m_1 = 0$ as the result of the arithmetic operation, since the only way to obtain a coefficient $1, -1$, or 0 in the monomial m_1 is that the nonzero coefficients of m_2 and m_3 are ± 1 and cancel each other. In other words, *whenever the arithmetic operation* (20) *has two nonzero operands we assign the result to zero without performing the arithmetic operation.* Thus we avoid the possible cancellation which this operation might have produced.

- *Case* 2*: The chosen pivot* E *is* 2×2 *with largest off-diagonal element* a_{pq}*,* $p < q$*.*
  Taking into account Lemma 4, the entries in the two columns of L are computed as

$$l_{ip} = (CE^{-1})_{ip} = \frac{-a_{ip}a_{qq}}{a_{pq}^2} + \frac{a_{iq}a_{pq}}{a_{pq}^2}, \tag{21}$$

$$l_{iq} = (CE^{-1})_{iq} = \frac{a_{ip}a_{pq}}{a_{pq}^2} - \frac{a_{iq}a_{pp}}{a_{pq}^2}. \tag{22}$$

  Again, both expressions can be written in the form $m_1 = m_2 + m_3$, where each m_i is a monomial with coefficients $0, 1$, or -1 in three variables, namely d_i and the reciprocals of d_p and d_q. The same argument employed above applies here, i.e., we can avoid any potential subtraction by setting $m_1 = 0$ whenever both operands m_2 and m_3 are nonzero.
  Something similar happens with the elements of D: the elements of the Schur complement of an arbitary matrix are of the form

$$(B - CE^{-1}C^T)_{ij}$$

$$= a_{ij} - \frac{a_{ip}a_{qq}a_{pj} - a_{iq}a_{pq}a_{pj} - a_{ip}a_{pq}a_{qj} + a_{iq}a_{pp}a_{qj}}{a_{pp}a_{qq} - a_{pq}^2}, \tag{23}$$

  but in our case, due to Lemma 4, we have $a_{pp}a_{qq} = 0$. Replacing this fact in (23), we obtain that the entries of the Schur complement are a sum

$$m_1 = m_2 + m_3 + m_4 + m_5 \tag{24}$$

  of at most four operands, each of which is a monomial in the variable $d_i d_j$ with coefficients ± 1 or 0, as is the result m_1 of (24). If all four operands are zero, or if there is a single nonzero operand, no operation is performed. If we have exactly two or four nonzero operands in (24), the same argument employed above implies that m_1 must be zero, since the only possible sum in $\{1, -1, 0\}$ of two or four numbers in $\{1, -1\}$ is zero. Finally, in the case when exactly three operands are nonzero, two of them must necessarily cancel each other, and the result m_1 is equal to the third operand. Therefore, if the three nonzero operands are m_r, m_s, m_t, then we can assign

$$m_1 = -|m_t|\text{sign}(m_r m_s m_t),$$

  where m_t is any of the three operands, since all three have the same absolute value $|d_i d_j|$.

Thus we have proved the following result.

THEOREM 2. *Algorithm* 1 *computes all entries of the factors* L *and* D *of the block* LDL^T *factorization of a symmetric DSTU matrix to high relative accuracy, i.e.,*

$$|\widehat{l}_{ij} - l_{ij}| = O(\epsilon)|l_{ij}|, \quad |\widehat{d}_{ij} - d_{ij}| = O(\epsilon)|d_{ij}|,$$

where $\widehat{L}$ *and* $\widehat{D}$ *are the factors computed in floating point arithmetic by Algorithm* 1 *and* L*,* D *are the the exact factors which the diagonal pivoting method would compute*

in exact arithmetic choosing the pivots with the same dimensions and positions as those chosen in floating point arithmetic to compute $\widehat{L}$ *and* $\widehat{D}$.

It should be noted that any attempt to theoretically estimate the constants inside the $O(\epsilon)$ in Theorem 2 is bound to be pessimistic. Take, for instance, the number p_k of floating point operations required to compute the elements of D at stage k of the LDL^T factorization. This quantity is given by the recursive formula $p_{k+1} = 2p_k + 1$, so $p_k = 2^{k+1} - 1$. Therefore, the constant inside the $O(\epsilon)$ given by a straightforward error analysis would be exponential. However, this is never observed in practice.

Finally, the previous analysis suggests the following $O(n^3)$ algorithm.

ALGORITHM 1. BLOCK LDL^T FACTORIZATION OF A SYMMETRIC DSTU MATRIX A.

Input: symmetric $n \times n$ DSTU matrix A

Output: unit lower triangular matrix L, block diagonal matrix D with 1×1 and 2×2 diagonal blocks, permutation matrix P such that $PAP^T = LDL^T$.

for $i = 1$ **to** n
choose pivot according to Bunch–Parlett pivoting strategy
  if 1×1 pivot a_{ii}
    $D_{ii} = a_{ii}$
    for $j = i + 1$ **to** n
      $l_{ji} = a_{ji}/a_{ii}$
    endfor
    for $j = i + 1$ **to** n
      for $k = i + 1$ **to** n
        $a_{jk} = a_{jk} - \dfrac{a_{ji}a_{ik}}{D_{ii}}$
        (*)If the last subtraction has two nonzero operands, set $a_{jk} = 0$
      endfor
    endfor
  elseif 2×2 pivot, $\begin{pmatrix} a_{ii} & a_{i,i+1} \\ a_{i+1,i} & a_{i+1,i+1} \end{pmatrix}$
    $D_{ii} = a_{ii}, D_{i,i+1} = D_{i+1,i} = a_{i,i+1}, D_{i+1,i+1} = a_{i+1,i+1}$
    for $j = i + 1$ **to** n
      $l_{ji} = \dfrac{a_{j,i+1}a_{i,i+1}}{a_{i,i+1}^2} - \dfrac{a_{ji}a_{i+1,i+1}}{a_{i,i+1}^2}$
      (*) If the last subtraction has two nonzero operands, set $l_{ji} = 0$
    endfor
    for $j = i + 2$ **to** n
      $l_{j,i+1} = \dfrac{a_{j,i}a_{i,i+1}}{a_{i,i+1}^2} - \dfrac{a_{j,i+1}a_{ii}}{a_{i,i+1}^2}$
      (*) If the last subtraction has two nonzero operands, set $l_{j,i+1} = 0$
    endfor
    for $j = i + 1$ **to** n
      for $k = i + 1$ **to** n
        $m_2 = a_{jk}, m_3 = -\dfrac{a_{jq}a_{pq}a_{pk}}{a_{pq}^2}, m_4 = -\dfrac{a_{jp}a_{pq}a_{qk}}{a_{pq}^2}$
        if $a_{pp} = 0$
          $m_5 = \dfrac{a_{jp}a_{qq}a_{pk}}{a_{pq}^2}$

```
29.                    else m_5 = a_jq a_pp a_qk / a_pq^2
30.                endif
31.                a_jk = m_2 + m_3 + m_4 + m_5
32.                (*) If the last addition has two or four nonzero operands,
                       set a_jk = 0
33.                (*) If the last subtraction has three nonzero operands m_r, m_s, m_t,
34.                   set a_jk = -|m_r| sign(m_r m_s m_t)
35.             endfor
36.          endfor
37.       endif
38. endfor
```

3.1.2. A new pivoting strategy. In addition to its accuracy, another feature of the GECP decomposition $PAP^T = LDU$ of nonsymmetric $n \times n$ DSTU matrices is that the condition numbers of L and U grow at most quadratically with the dimension n [11, Theorem 10.2]. To prove the same for the triangular factor L in the block LDL^T decomposition we will slightly change the pivoting strategy. Consider the following one:

$$\begin{array}{l} \textbf{if} \quad \mu_0 = \max_{i,j}|a_{ij}| = \max_i |a_{ii}| = \mu_1 \\ \qquad \text{choose } 1 \times 1 \text{ pivot} \\ \textbf{else} \\ \qquad \text{choose } 2 \times 2 \text{ pivot} \end{array} \tag{25}$$

With this strategy, the entries of L are trivially bounded by 1 in absolute value, while the best one can say for Bunch–Parlett is that $|l_{ij}| \leq 1/\alpha \approx 1.6$ (for the elements generated by 1×1 pivots). Also, the condition number in the 2-norm of the pivots is bounded by 4.6 for Bunch–Parlett and by 2.6 for this new strategy. Notice that the change in the value of α does not affect the validity of the results in section 3.1.1, since Lemma 4 remains true for all $\alpha \leq 1$.

We are now in the position to prove that, with this modified pivoting strategy, the condition number of the factor L of the block LDL^T factorization (4) grows at most quadratically with the dimension of the factorized matrix.

THEOREM 3. *Let A be a symmetric DSTU matrix. There is a DSTU matrix B whose unit lower triangular factor computed by Gaussian elimination with complete pivoting coincides with the triangular factor of the block LDL^T factorization of A obtained using the pivoting strategy* (25)*. Therefore, the condition number of the latter triangular factor grows at most quadratically with the dimension of A.*

Proof. Without loss of generality, we may restrict ourselves to comparing the first two steps of GECP with the corresponding steps of the diagonal pivoting method. If the first pivot in the diagonal pivoting method is 1×1, then it applies to A the same permutations as GECP, and the entries of the first column of both triangular factors trivially coincide. Moreover, since the pivot is chosen from the diagonal, the matrix is symmetrically permuted by GECP and the Schur complements also coincide for both methods.

Now, suppose that the first pivot in the diagonal pivoting method is 2×2, say

$$\begin{pmatrix} a_{pp} & a_{pq} \\ a_{qp} & 0 \end{pmatrix}.$$

The proof for the case $a_{pp} = 0$ is completely analogous. Then, if we denote by P_{ij} the permutation which interchanges rows i and j, the diagonal pivoting method permutes the matrix A into $P_{2q}P_{1p}\ A\ P_{1p}P_{2q}$ in order to place the 2×2 pivot matrix in the upper left corner. GECP, on the other hand, would permute A to $P_{1p}AP_{1q}$ to place a_{pq} in the upper left corner of the matrix. However, it will be convenient for our purpose to apply some additional permutations: applying P_{2q} on the rows places the zero entry a_{qq} in the $(2,1)$ position. Applying P_{2q} and P_{2p} on the columns places the entry $a_{qp} = a_{pq}$ in the $(2,2)$ position and reorders the entries in such a way that, if we rename

$$M_1 = P_{2q}P_{1p}\ A\ P_{1p}P_{2q}, \qquad M_2 = P_{2q}P_{1p}AP_{1q}P_{2q}P_{2p},$$

both matrices M_1 and M_2 are identical, except for the first two columns, which are switched. Now, denote by m_{ij} the (i,j) element of the symmetric matrix M_1 (recall that $m_{22} = 0$, and m_{12} is the entry of M_1 with largest absolute value). Then the diagonal pivoting method computes the entries of the first two columns of L as

$$\begin{aligned} l_{i1} &= \frac{m_{i2}}{m_{12}}, \quad i = 3, \ldots, n, \\ l_{i2} &= \frac{m_{i1}m_{12} - m_{i2}m_{11}}{m_{12}^2}, \quad i = 3, \ldots, n, \end{aligned}$$

and the entries of the $(n-2) \times (n-2)$ Schur complement are

(26)
$$(B - CE^{-1}C^T)_{ij} = m_{ij} + \frac{-m_{i2}m_{12}m_{1j} - m_{i1}m_{12}m_{2j} + m_{i2}m_{11}m_{2j}}{m_{12}^2}, \quad i,j = 3, \ldots, n.$$

We will prove that the first two steps of Gaussian elimination on M_2 produce exactly the same two rows of L and the same $(n-2) \times (n-2)$ Schur complement. The first step of Gaussian elimination on M_2 trivially produces a first column with a zero in the first position, and $l_{i1} = m_{i2}/m_{12},\ i = 3, \ldots, n$ as above. The $(n-1) \times (n-1)$ resulting Schur complement has the form

(27)
$$\left(\begin{array}{c|ccc} m_{12} & \cdots & m_{2j} & \cdots \\ \hline \vdots & \vdots & \vdots & \vdots \\ m_{i1} - \dfrac{m_{i2}m_{11}}{m_{12}} & \vdots & m_{ij} - \dfrac{m_{i2}m_{1j}}{m_{12}} & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{array}\right).$$

In the second step, GECP looks for the entry with largest absolute value in this matrix. We claim this entry is again m_{12}; if not, there must be some new entry, which was not in M_2, larger than m_{12} in absolute value. Any entry of the Schur complement (27) vanishes whenever it is the result of a subtraction with nonzero operands, so the only possibly new elements are quotients $-(m_{i2}m_{1j})/(m_{12})$. These quotients are still monomials with coefficient ± 1 in the two diagonal entries of $\mathcal{D}$ corresponding to the indices i and j before A was permuted to M_2. In any case, since both i, j are different from either 1 or 2, both $\widetilde{d}_i$ and $\widetilde{d}_j$ are different from d_p and d_q. Consequently, the absolute value of any new element in (27) is strictly smaller than the maximum $|m_{12}| = |d_p d_q|$.

Hence, no permutation is needed in the second step of GECP on M_2. This step produces a second column for L with entries

$$l_{i2} = \frac{m_{i1} - \dfrac{m_{11}m_{i2}}{m_{12}}}{m_{12}} = \frac{m_{i1}m_{12} - m_{11}m_{i2}}{m_{12}^2}$$

which coincide with the entries l_{i2} above, replacing i by j (recall that $m_{ij} = m_{ji}$). Finally, the $(n-2) \times (n-2)$ Schur complement computed by GE in this second step has entries

$$\left(m_{ij} - \frac{m_{i2}m_{1j}}{m_{12}}\right) - \frac{\left(m_{1i} - \dfrac{m_{11}m_{i2}}{m_{12}}\right) m_{2j}}{m_{12}} \ .$$

A straightforward computation shows the equality of this formula with (26). This proves that, as claimed, two steps of Gaussian elimination on M_2 produce the same two columns for L as one 2×2 step of the diagonal pivoting method on M_1. Furthermore, the remaining $(n-2) \times (n-2)$ Schur complement is also the same.

Therefore, repeating the argument for the subsequent steps of both decompositions, we obtain that the factor L of the symmetric decomposition (4) of A is equal to the lower triangular factor of the LDU decomposition of a nonsymmetric matrix $B = AQ$ for an appropriate permutation matrix Q. Notice that B is DSTU if A is DSTU, since

$$B = D_L \widetilde{Z} D_R,$$

with $D_L = D$, $D_R = Q^T D Q$, $\widetilde{Z} = ZQ$, and the latter is trivially TU. The bound on the condition number of L follows trivially from Theorem 10.2 in [11]. ∎

The proof of Theorem 3 relies somewhat indirectly on [11, Theorem 10.2]. Trying a more direct path, with a proof analogous to the one of [11, Theorem 10.2] is unfeasible, since we lack one of the essential ingredients, namely the property of Gaussian elimination that the elements of any minor of L are a quotient of minors of the original matrix A and, therefore, the elements of L^{-1} are quotients of minors of A. This no longer holds for the LDL^T factorization, as shown by the 3×3 example in section 2.

3.2. Total signed compound matrices.

DEFINITION 3. *Let $\mathcal{S}$ be a set of matrices with given sparsity and sign pattern, i.e., all matrices in $\mathcal{S}$ have their nonzero entries in the same position and with the same sign. The set $\mathcal{S}$ is* total signed compound (TSC) *if, for every $A \in \mathcal{S}$ and for every square submatrix M of A, the Laplace expansion*

$$\det M = \sum_{\pi} [\mathrm{sign}(p) m_{1,\pi_1} m_{2,\pi_2} \dots m_{s,\pi_s}] \tag{28}$$

of the determinant of M is either a sum of monomials of like sign, with at least one nonzero monomial, or identically zero (i.e., no nonzero monomial appears in the expression).

There are well-known classes of pattern matrices among the TSC, provided their particular sign distribution conforms to the TSC condition. Two such examples are,

for instance, the tridiagonal pattern

$$\begin{pmatrix} + & + & & & \\ & + & - & + & \\ & + & + & + & \\ & & + & - & + \\ & & & + & + \end{pmatrix}$$

and the arrowhead pattern

$$\begin{pmatrix} + & + & + & + & + \\ + & - & & & \\ + & & - & & \\ + & & & - & \\ + & & & & - \end{pmatrix}.$$

TSC matrices are rather sparse (there are at most $3n-2$ nonzero entries in an $n \times n$ TSC matrix), and there are $O(n)$ algorithms for computing the determinant of an $n \times n$ TSC matrices. Moreover, such algorithms compute the determinant to high relative accuracy, since, due to the TSC property, no cancellation occurs in the calculation (recall that the determinant is determined to high relative accuracy, since it is subtraction-free). The $O(n)$ cost is achieved by making use of an alternative, constructive definition of TSC matrices: every TSC matrix can be constructed starting from a 1×1 nonzero matrix and repeatedly applying four construction rules (see [1, 11] for more details). If we restrict ourselves to *symmetric* TSC matrices, one can prove that only three construction rules are needed.

THEOREM 4. *Every TSC symmetric matrix can be obtained by starting with a* 1×1 *nonzero matrix and applying the following three construction rules repeatedly in some order:*

1. *If A is symmetric and TSC, then permuting two rows and the corresponding columns, or multiplying by -1 one row and the corresponding column, leaves A symmetric TSC.*
2. *If A_1 and A_2 are symmetric TSC matrices, then so is the direct sum*
$$\begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}.$$
3. *If the $n \times n$ $\widetilde{A}$ is symmetric and TSC, with $\widetilde{a}_{ii} \neq 0$, then so is the $(n+1) \times (n+1)$ matrix A obtained as follows:*
$$A = \left(\begin{array}{ccccccc|c} & & & & & & & 0 \\ & & & & & & & \vdots \\ & & & & & & & 0 \\ & & & \widetilde{A} & & & & a_{i,n+1} \\ & & & & & & & 0 \\ & & & & & & & \vdots \\ & & & & & & & 0 \\ \hline 0 & \dots & 0 & a_{i,n+1} & 0 & \dots & 0 & a_{n+1,n+1} \end{array}\right),$$
where we can also set $\widetilde{a}_{ii}$ to zero. The new possibly nonzero entries $a_{i,n+1}$ and $a_{n+1,n+1}$ must be chosen so that the two monomials in the minor $a_{n+1,n+1}\widetilde{a}_{i,i} - a_{i,n+1}a_{n+1,i}$ have the same sign (or are zero).

These rules will allow us to inexpensively generate TSC matrices in section 5.

3.2.1. Accurate block LDL^T factorization of symmetric TSC matrices. Our interest in TSC matrices stems from the fact that all their minors can be computed without subtraction and therefore with no cancellation. According to Theorem 1, any intermediate quantity in the process of the block LDL^T factorization is a quotient of minors (or just a minor) of the original matrix. Although the standard formulas for the diagonal pivoting method may require subtraction and therefore lead to cancellation, that subtraction can be avoided if the corresponding element of L or of the Schur complement is recomputed as a quotient of minors of the original TSC matrix. This can be done using the formulas in Lemma 1 for the Schur complements and the formulas in Lemma 2 for the elements of L. Moreover, the lack of cancellation implies that these minors are computed to high relative accuracy.

The argument above amounts to proving the following theorem.

THEOREM 5. *Algorithm* 2 *computes all entries of the L and D factors of the block LDL^T factorization of a symmetric TSC matrix to high relative accuracy, i.e.,*

$$\frac{|\widehat{l}_{ij} - l_{ij}|}{|l_{ij}|} = O(\epsilon), \quad \frac{|\widehat{d}_{ij} - d_{ij}|}{|d_{ij}|} = O(\epsilon),$$

where $\widehat{L}$ and $\widehat{D}$ are the factors computed in floating point arithmetic by Algorithm 2 *and L and D are the the exact factors which the diagonal pivoting method would compute in exact arithmetic choosing the pivots with the same dimensions and positions as those chosen in floating point arithmetic to compute $\widehat{L}$ and $\widehat{D}$.*

We now write the pseudocode for the corresponding algorithm. Of course, recomputing represents an overhead cost, since every $s \times s$ minor costs $O(s)$ operations instead of the $O(1)$ operations for the standard formula (any submatrix of a TSC matrix is trivially TSC). Therefore, the following modification of the diagonal pivoting method can cost in the worst case as much as $O(n^4)$ arithmetic operations, the same asymptotic order of the algorithm computing nonsymmetric RRDs of TSC matrices in [11] (see [11, Theorem 7.2, p. 60]). For more on this question, see the experiment below at the end of section 5.2.

ALGORITHM 2. BLOCK LDL^T FACTORIZATION OF A SYMMETRIC TSC MATRIX A.

Input: symmetric $n \times n$ TSC matrix A

Output: unit lower triangular matrix L, block diagonal matrix D with 1×1 and 2×2 diagonal blocks, permutation matrix P such that $PAP^T = LDL^T$.

1. **for** $i = 1$ **to** n
2. choose pivot according to Bunch–Parlett pivoting strategy
3. **if** 1×1 pivot, a_{ii}
4. $D_{ii} = a_{ii}$
5. **for** $j = i + 1$ **to** n
6. $l_{ji} = a_{ji}/a_{ii}$
7. **endfor**
8. **for** $j = i + 1$ **to** n
9. **for** $k = i + 1$ **to** n
10. $a_{jk} = a_{jk} - \dfrac{a_{ji}a_{ik}}{D_{ii}}$
11. (*) If the last subtraction has two nonzero operands with the same sign, recompute a_{jk} as the quotient of two minors of A according to formula (13) in Lemma 1

endfor
endfor
elseif 2×2 pivot, $\begin{pmatrix} a_{ii} & a_{i,i+1} \\ a_{i+1,i} & a_{i+1,i+1} \end{pmatrix}$
$D_{ii} = a_{ii}, D_{i,i+1} = D_{i+1,i} = a_{i,i+1}, D_{i+1,i+1} = a_{i+1,i+1}$
for $j = i + 1$ **to** n
$dpiv = a_{ii}a_{i+1,i+1} - a_{i,i+1}^2$
(*) If this subtraction has two nonzero operands with the same sign, recompute $dpiv$ as the quotient of two minors of A according to formula (14) in Lemma 1
$l_{ji} = \frac{a_{ji}a_{i+1,i+1}}{dpiv} - \frac{a_{j,i+1}a_{i,i+1}}{dpiv}$
(*) If the last subtraction has two nonzero operands with the same sign, recompute l_{ji} as the quotient of two minors of A according to formula (15) in Lemma 2
endfor
for $j = i + 2$ **to** n
$l_{j,i+1} = -\frac{a_{j,i+1}a_{i,i+1}}{dpiv} + \frac{a_{j,i+1}a_{ii}}{dpiv}$
(*) If the last subtraction has two nonzero operands with the same sign, recompute $l_{j,i+1}$ as the quotient of two minors of A according to formula (15) in Lemma 2
endfor
for $j = i + 1$ **to** n
for $k = i + 1$ **to** n
$a_{jk} = a_{jk} - \frac{a_{jp}a_{qq}a_{pk}}{dpiv} - \frac{a_{jq}a_{pq}a_{pk}}{dpiv} - \frac{a_{jp}a_{pq}a_{qk}}{dpiv} + \frac{a_{jq}a_{pp}a_{qk}}{dpiv}$
(*) If the last subtraction has two nonzero operands with the same sign, recompute a_{jk} as the quotient of two minors of A according to formula (13) in Lemma 1
endfor
endfor
endif
endfor

4. From LDL^T to RRD. Once we have a block LDL^T factorization, we have seen in (5) how to obtain a symmetric RRD by Givens diagonalization. It is not hard to show that, since L is computed with small elementwise error and $X = P^T LQ$ is, up to permutations, the result of a floating point Givens transformation (see, e.g., [6, Lemma 3.1]), the computed X satisfy (7). However, we will prove a tighter bound in Theorem 6 below, namely that X is computed with *columnwise* small relative errors. Note also that X and L have the same condition number, so if L is well-conditioned, then so is X (this is guaranteed, for instance, for DSTU matrices, according to Theorem 3).

4.1. Error analysis. We present here the error analysis showing that the block LDL^T factorization, followed by Givens diagonalization, leads to RRDs satisfying the requirement (7) for accurately computing eigenvalues and eigenvectors via the signed SVD method. We make no distinction between DSTU and TSC matrices, since the error analysis is valid for any matrix such that its block LDL^T decomposition can be

computed with small componentwise error as in (6). The analysis is related to the one in [15] and uses some results appearing in [15]. To be more precise we need some notation: we assume the conventional model for floating point arithmetic,

$$\texttt{fl}(a \odot b) = (a \odot b)(1+\delta), \tag{29}$$

where a and b are real floating point numbers, $\odot \in \{+,-,\times,/\}$, and $|\delta| \leq \epsilon$, where ϵ is the machine precision. Moreover, we assume that neither overflow nor underflow occur. Also, for each $k > 0$ we set

$$\gamma_k = \frac{k\epsilon}{1-k\epsilon}, \tag{30}$$

and as in [18, section 3.4], we denote by θ_k any positive quantity bounded by γ_k. Finally, given a real symmetric 2×2 matrix, we write the Jacobi orthogonal diagonalization procedure as

$$\begin{pmatrix} a & c \\ c & b \end{pmatrix} = \begin{pmatrix} cs & sn \\ -sn & cs \end{pmatrix} \begin{pmatrix} \lambda_1 & \\ & \lambda_2 \end{pmatrix} \begin{pmatrix} cs & -sn \\ sn & cs \end{pmatrix}, \tag{31}$$

with $\lambda_1 = a - ct,\ \lambda_2 = b + ct$, where

$$t = \frac{\operatorname{sign}(\zeta)}{|\zeta| + \sqrt{1+\zeta^2}} \qquad \text{for} \qquad \zeta = \frac{b-a}{2c} \tag{32}$$

and

$$cs = \frac{1}{\sqrt{1+t^2}}, \qquad sn = cs \cdot t. \tag{33}$$

The main result in this section, written with this notation, is the following.

THEOREM 6. *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix and let $\widehat{L}, \widehat{D}$ be the computed factors of a block LDL^T factorization of A obtained through the diagonal pivoting method using the Bunch–Parlett pivoting strategy* (11) *(i.e., $\alpha = (1+\sqrt{17})/8$). Suppose that $\widehat{L}, \widehat{D}$ have been computed with small componentwise relative error*

$$\begin{aligned} \widehat{l}_{ij} &= l_{ij}(1+\theta_{K_L}^{(ij)}), \qquad i,j = 1,\dots,n, \\ \widehat{d}_{ij} &= l_{ij}(1+\theta_{K_D}^{(ij)}), \qquad i,j = 1,\dots,n, \end{aligned} \tag{34}$$

for appropriate constants $K_L, K_D > 0$, and let $\widehat{X}, \widehat{\Delta}$ (resp., X, Δ) be the computed (resp., exact) factors of a symmetric RRD obtained by Givens diagonalization (5) *in floating point (resp., in exact arithmetic) using formulas* (31)–(33). *Then*

$$\frac{|\widehat{\Delta}_{jj} - \Delta_{jj}|}{|\Delta_{jj}|} \leq 4\,\frac{1+\alpha}{1-\alpha}\,\gamma_{K_D+29}, \qquad j = 1,\dots,n, \tag{35}$$

and

$$\frac{\|\widehat{X}(:,j) - X(:,j)\|_2}{\|X(:,j)\|_2} \leq \sqrt{2nC^2+1}\ \gamma_M, \qquad j = 1,\dots,n, \tag{36}$$

where C is

$$C = \frac{1}{1-\alpha} + O(\epsilon).$$

and

$$M = \max\{48K_{_L} + 141, K_{_L} + 48K_{_D} + 143\}. \tag{37}$$

To prove it we will use the fact that the computed diagonalizing transformation is close entrywise to the exact one. The following result is taken from [15, Appendix A.3].

LEMMA 5 (see [15]). *Let*

$$\widetilde{A} = \begin{pmatrix} \widetilde{a} & \widetilde{c} \\ \widetilde{c} & \widetilde{b} \end{pmatrix} = \begin{pmatrix} a(1+\delta_a) & c(1+\delta_c) \\ c(1+\delta_c) & b(1+\delta_b) \end{pmatrix}$$

be a matrix of real floating point numbers, with $\max\{|\delta_a|, |\delta_b|, |\delta_c|\} \le \gamma_k$ *and* $\alpha|\widetilde{c}| \ge \max\{|\widetilde{a}|, |\widetilde{b}|\}$. *Let*

$$A = \begin{pmatrix} a & c \\ c & b \end{pmatrix},$$

with eigenvalues $\lambda_1 \ge \lambda_2$, *and orthonormal eigenvectors* $v_1 = [cs, -sn]^T$ *and* $v_2 = [sn, cs]$. *Let* $\widehat{\lambda}_1, \widehat{\lambda}_2, \widehat{cs}$ *and* $\widehat{sn}$ *be the versions of* λ_1, λ_2, cs *and* sn *computed in floating point arithmetic for* $\widetilde{A}$ *according to formulas* (31)–(33). *If*

$$4\sqrt{2}\,\frac{1+\alpha}{1-\alpha}\gamma_{k+29} \le 1 \qquad \textit{and} \qquad \gamma_{141+48k} \le 1,$$

then

$$\frac{|\widehat{\lambda}_i - \lambda_i|}{|\lambda_i|} \le 4\,\frac{1+\alpha}{1-\alpha}\gamma_{k+29}, \quad i = 1, 2, \tag{38}$$

and

$$\widehat{cs} = cs(1 + \theta_{16k+113}), \qquad \widehat{sn} = cs(1 + \theta_{48k+141}). \tag{39}$$

Proof of Theorem 6. First, we may assume that $P = I$, since no error is introduced by the pivoting permutations. Let $\widehat{Q}$ be the computed orthogonal matrix diagonalizing $\widehat{D}$; i.e., if $\widehat{D} = \operatorname{diag}(\widehat{D}_1, \ldots, \widehat{D}_r)$ with $D_k \in \mathbb{R}^{s_k \times s_k}$, $s_k = 1$ or 2, $k = 1, \ldots, r$, then $\widehat{Q} = \operatorname{diag}(\widehat{Q}_1, \ldots, \widehat{Q}_r)$ with $Q_k \in \mathbb{R}^{s_k \times s_k}$, $k = 1, \ldots, r$. The 1×1 blocks of $\widehat{Q}_k$ are equal to 1, and each 2×2 block

$$\widehat{Q}_k = \begin{pmatrix} \widehat{cs} & -\widehat{sn} \\ \widehat{sn} & \widehat{cs} \end{pmatrix} \tag{40}$$

is the version computed in floating point arithmetic of the Jacobi rotation which would diagonalize the 2×2 block $\widehat{D}_k$ in exact arithmetic. Analogously, $Q = \operatorname{diag}(Q_1, \ldots, Q_r)$, where

$$Q_k = \begin{pmatrix} cs & -sn \\ sn & cs \end{pmatrix}$$

is the exact Jacobi rotation diagonalizing the diagonal block D_k of D. For those columns j corresponding to a pivot with $s_k = 1$, we have $\widehat{\Delta}_{jj} = \widehat{d}_{jj}$, $\Delta_{jj} = d_{jj}$ and

$\widehat{X}(:,j) = \widehat{L}(:,j)$, $X(:,j) = L(:,j)$, so (35) and (36) are trivially satisfied. Therefore, only the columns corresponding to 2×2 pivots must be considered. Let the jth and $(j+1)$st be two such columns. First, inequality (35) follows directly from applying Lemma 5 in our setting, i.e., taking $\widetilde{A}$, A, λ_1, λ_2 , and k to be equal, respectively, to $\widehat{D}$, D, Δ_{jj}, $\Delta_{j+1,j+1}$, and $K_{_D}$. With this choice, inequality (38) reduces to (35). To prove (36), note that

$$X = LQ, \qquad \widehat{X} = \mathtt{fl}(\widehat{L}\widehat{Q}),$$

where $\mathtt{fl}(expr)$ denotes the computed result in finite precision of expression $expr$. Reading these identities entrywise for columns j and $j+1$, we get

$$\widehat{X}(i,j) = \begin{cases} 0 & \text{if} \quad i < j\ , \\ \widehat{cs} & \text{if} \quad i = j\ , \\ \widehat{sn} & \text{if} \quad i = j+1\ , \\ \mathtt{fl}(\widehat{l}_{ij}\,\widehat{cs} + \widehat{l}_{i,j+1}\,\widehat{sn}) & \text{if} \quad i > j+1\ , \end{cases} \tag{41}$$

$$\widehat{X}(i,j+1) = \begin{cases} 0 & \text{if} \quad i < j\ , \\ -\widehat{sn} & \text{if} \quad i = j\ , \\ \widehat{cs} & \text{if} \quad i = j+1\ , \\ \mathtt{fl}(-\widehat{l}_{ij}\widehat{sn} + \widehat{l}_{i,j+1}\widehat{cs}) & \text{if} \quad i > j+1 \end{cases}$$

and the same equalities without hats for the entries of X. Therefore,

$$||\widehat{X}(:,j)-X(:,j)||_2^2 = (\widehat{cs}-cs)^2+(\widehat{sn}-sn)^2+\sum_{i=j+2}^{n}[\mathtt{fl}(\widehat{l}_{ij}\widehat{cs}+\widehat{l}_{i,j+1}\widehat{sn})-(l_{ij}cs+l_{i,j+1}sn)]^2.$$

Using (34), (29), and (39), we may write

$$\begin{aligned} \mathtt{fl}(\widehat{l}_{ij}\widehat{cs} + \widehat{l}_{i,j+1}\widehat{sn}) &= \Big[l_{ij}\,cs\,(1+\theta_{K_{_L}}^{(i,j)})\,(1+\theta_{16K_{_D}+113})\,(1+\delta_1) \\ &\qquad + l_{i,j+1}\,sn\,(1+\theta_{K_{_L}}^{(i,j+1)})\,(1+\theta_{48K_{_D}+141})\,(1+\delta_2)\Big]\,(1+\delta_3) \\ &= l_{ij}\,cs\,(1+\theta_{K_{_L}+16K_{_D}+115}) + l_{i,j+1}\,sn\,(1+\theta_{K_{_L}+48K_{_D}+143}). \end{aligned}$$

Hence, again using (39),

$$\begin{aligned} \|\widehat{X}(:,j) - X(:,j)\|_2^2 &= (cs\,\theta_{16K_{_D}+113})^2 + (sn\,\theta_{48K_{_D}+141})^2 + \\ &\qquad + \sum_{i=j+2}^{n}\Big(l_{ij}\,cs\,\theta_{K_{_L}+16K_{_D}+115} + l_{i,j+1}\,sn\,\theta_{K_{_L}+48K_{_D}+143}\Big)^2 \\ &\le (\gamma_{48K_{_L}+141})^2 + 2n\,(\gamma_{K_{_L}+48K_{_D}+143})^2 \max\{|l_{ij}|^2, |l_{i,j+1}|^2\}, \end{aligned}$$

where we have used the monotonicity of γ_k in k. At this point, we must observe that, although the the Bunch–Parlett strategy ensures that the entries of the *computed* $\widehat{L}$ satisfy $|\widehat{l}_{ik}| \le 1/(1-\alpha)$ for all i,k, this may not be true for the entries l_{ik} of the *exact* L. The entrywise bound (34), however, implies (after some calculations) that

$|l_{ik}| \leq C$ for a constant C which is equal to $1/(1-\alpha)$ up to first order terms[4] in ϵ. Therefore,

$$\|\widehat{X}(:,j) - X(:,j)\|_2 \leq \sqrt{1+2nC^2}\,\gamma_M$$

with M given by (37), which leads trivially to (36), since $\|X(:,j)\|_2^2 \geq cs^2 + sn^2 \geq 1$. □

5. Numerical experiments. We have performed extensive numerical tests which confirm the correctness of our algorithms. All of them were done in MATLAB 5.3 using an AMD Athlon (tm) XP 2000+ processor with IEEE arithmetic.

We have used as a reference the eigenvalues and eigenvectors computed using Maple's variable-precision arithmetic available from the Symbolic Math Toolbox of MATLAB through the command `vpa`. For each matrix A, the "exact" eigendecomposition is obtained using MATLAB's usual command `eig` (i.e., with the QR algorithm) but setting the variable `digits`, which specifies the number of significant decimal digits used by Maple to $18+d$ if the condition number of A is $O(10^d)$. We denote by λ_i and q_i the eigenvalues and eigenvectors computed in this way and by $\widehat{\lambda}_i$ and $\widehat{q}_i$ those computed via the signed SVD method implemented in MATLAB. Therefore, $\widehat{\lambda}_i, \widehat{q}_i$ are computed in double precision arithmetic, i.e., $\epsilon \approx 2.2 \cdot 10^{-16}$. The initial RRD factorization is implemented in MATLAB, using Algorithm 1 for DSTU matrices and Algorithm 2 for TSC matrices.

We analyzed the following quantities:

1. the maximum relative error in eigenvalues:

$$e_\lambda = \max_i \left| \frac{\lambda_i - \widehat{\lambda}_i}{\lambda_i} \right|; \tag{42}$$

2. a control quantity for eigenvalues:

$$\vartheta_\lambda = \frac{e_\lambda}{\kappa\epsilon}, \tag{43}$$

   where ϵ is the machine precision, and $\kappa = \kappa(R')\,\kappa(X)$ as in (8)—this quantity is expected to be roughly $O(1)$ in the experiments;
3. the maximum relative error in the eigenvectors:

$$e_q = \max_i ||\widehat{q}_i - q_i||_2; \tag{44}$$

4. a control quantity for eigenvectors:

$$\xi_q = \max_i \frac{||\widehat{q}_i - q_i||_2\,\mathrm{relgap}^*(\widehat{\lambda}_i)}{\kappa\epsilon} \tag{45}$$

   with κ as above and relgap* defined as in (9). Again, ξ_q should be $O(1)$ for the experiments to confirm our analysis.

[4] See [15] for a more detailed analysis, which proves that

$$C = \frac{1}{(1-\alpha)(1-\gamma_{g(\alpha)})}, \qquad g(\alpha) = \left(32\left(\frac{1+\alpha}{1-\alpha}\right)^2 + 196\,\frac{1+\alpha}{1-\alpha}\right) K_D.$$

5.1. Diagonally scaled totally unimodular matrices. We generated TU nonsingular matrices of sizes $6, 8, 10$, and 12. We were unable to generate matrices of larger dimension due to the high cost of the generating routine: it generates TU matrices recursively, starting from a TU matrix of size 1, i.e., either $-1, 1$, or 0. Given a generated matrix of size s, the algorithm constructs a $s+1 \times s+1$ TU matrix by adjoining a new row and column, with entries randomly chosen among $-1, 1, 0$, and checking whether all new minors containing entries from that row and column are equal to $-1, 1$, or 0. The computational cost of checking the minors is what makes the algorithm so costly.

Once we have a TU matrix, we scale it on both sides with diagonal matrices with powers of 10 on the diagonal, their condition numbers ranging from 10^5 to 10^{20}. Therefore, the corresponding DSTU matrices will have condition numbers ranging roughly from 10^{10} to 10^{40}. For each size we divide the experiments in three groups according to their condition number: condition numbers ranging from 10^{10} to 10^{20}, from 10^{20} to 10^{30}, and from 10^{30} to 10^{40}. We generate 100 matrices for each range, so the following tables reflect the results on 1200 matrices, 300 for each dimension. Table 1 shows the control quantities ϑ_λ for eigenvalues, while Table 2 shows the control quantities ξ_q for eigenvectors. For each dimension there are two columns, the left one displaying the average over the 100 tests made in that range of condition numbers and the right one displaying the largest value for the control quantity among the 100 experiments.

TABLE 1
Statistical data for accuracy in eigenvalues of DSTU matrices: ϑ_λ.

	$n=6$		$n=8$		$n=10$		$n=12$	
$\kappa(A) = O(10^d)$	Mean	Max	Mean	Max	Mean	Max	Mean	Max
$10 \le d \le 20$	1.412	6.689	1.746	32.34	1.879	19.14	1.425	9.310
$20 \le d \le 30$	1.460	16.34	1.652	38.14	1.432	13.49	1.696	45.45
$30 \le d \le 40$	1.699	26.65	1.338	11.34	1.157	3.949	1.719	33.02

TABLE 2
Statistical data for accuracy in eigenvectors of DSTU matrices: ξ_q.

	$n=6$		$n=8$		$n=10$		$n=12$	
$\kappa(A) = O(10^d)$	Mean	Max	Mean	Max	Mean	Max	Mean	Max
$10 \le d \le 20$	0.508	2.653	0.508	1.886	0.579	1.989	0.605	2.364
$20 \le d \le 30$	0.502	1.914	0.518	1.716	0.623	2.214	0.603	1.928
$30 \le d \le 40$	0.447	1.884	0.582	2.795	0.571	2.840	0.621	2.697

5.2. Total signed compound matrices. We generated TSC matrices of sizes $10, 20, 40$, and 60 by starting from a nonzero 1×1 matrix and repeatedly applying rules 2 and 3 in Theorem 4. Rule 2 was applied with a probability of 5%, choosing as A_2 one of the blocks of the TSC matrix A_1 computed in the previous stage. Otherwise, rule 3 was applied, generating the new quantities $a_{i,n+1}$, $a_{n+1,n+1}$ with MATLAB's `rand` command. Whenever rule 3 was employed, the diagonal entry $\widetilde{a}_{ii}$ was set to zero with a probability of 20%. Finally, large condition numbers were induced by scaling the resulting matrices on both sides with ill-conditioned diagonal matrices, exactly as in the experiment for DSTU matrices. Notice that, since the scaling matrices were positive, the sign pattern of the matrix does not change under scaling. Again, 1200 matrices were generated, 100 for each dimension and each range of condition numbers.

The results are summarized in Tables 3 and 4.

TABLE 3
Statistical data for accuracy in eigenvalues of TSC matrices: ϑ_λ.

	$n = 10$		$n = 20$		$n = 40$		$n = 60$	
$\kappa(A) = O(10^d)$	Mean	Max	Mean	Max	Mean	Max	Mean	Max
$10 \le d \le 20$	1.446	10.024	1.449	4.196	1.940	8.802	2.280	9.639
$20 \le d \le 30$	1.332	6.579	2.170	38.68	2.033	5.172	2.278	9.528
$30 \le d \le 40$	1.362	5.973	1.591	7.411	2.841	44.70	2.502	9.583

TABLE 4
Statistical data for accuracy in eigenvectors of TSC matrices: ξ_q.

	$n = 10$		$n = 20$		$n = 40$		$n = 60$	
$\kappa(A) = O(10^d)$	Mean	Max	Mean	Max	Mean	Max	Mean	Max
$10 \le d \le 20$	0.682	3.044	0.843	2.987	1.292	3.342	1.418	3.641
$20 \le d \le 30$	0.717	7.438	0.889	4.215	1.294	3.034	1.405	3.665
$30 \le d \le 40$	0.800	3.768	0.893	3.386	1.265	2.802	1.471	3.672

As can be seen from the tables, the results confirm our theoretical predictions. In parallel, we also computed eigenvalues and eigenvectors of the test matrices with MATLAB's `eig` command. As expected, the relative errors were huge, providing no correct digit in the smaller eigenvalues.

We conclude with an experiment to estimate the actual computational cost of the LDL^T factorization for symmetric TSC matrices: we randomly generate one hundred symmetric TSC matrices for each size from 10 to 100 in steps of ten, i.e., we generate one hundred 10×10 matrices, one hundred 20×20 matrices, one hundred 30×30 matrices and so on, i.e., one thousand test matrices in all. For each matrix we compute an LDL^T factorization using Algorithm 2, and we record the number of flops employed by the factorization procedure. Each star in Figure 1 corresponds to a given size and represents the arithmetic mean of the one hundred data obtained for that size, plotted in a log-log scale, with the logarithm of the size n of the matrix on the horizontal axis. The solid line corresponds to `flops`$= n^4$, and the dashed line to `flops`$= n^3$. As can be seen in the figure, the cost seems to be somewhere in between. However, since we have no estimation of the constants involved in the big-oh, it is hard to draw any specific conclusion, other than that the cost seems not to be too high.

Appendix A. Proofs of results in section 2.

Underlying the results in section 2 is one of the most useful properties of Schur complements, usually known as the *quotient property* (see, e.g., [5, section 2]).

LEMMA 6. *Let M be any square matrix, partitioned as*

$$M = \begin{pmatrix} B & * \\ * & * \end{pmatrix}, \qquad \textit{with} \qquad B = \begin{pmatrix} B_1 & * \\ * & * \end{pmatrix},$$

where B and B_1 are square nonsingular. Let $\mathcal{C}_1^B$ (resp., $\mathcal{C}_1^M$) be the Schur complement of B_1 in B (resp., in M). Then the Schur complement of B in M is equal to the Schur complement of $\mathcal{C}_1^B$ in $\mathcal{C}_1^M$.

With this result one can easily prove Lemma 1.

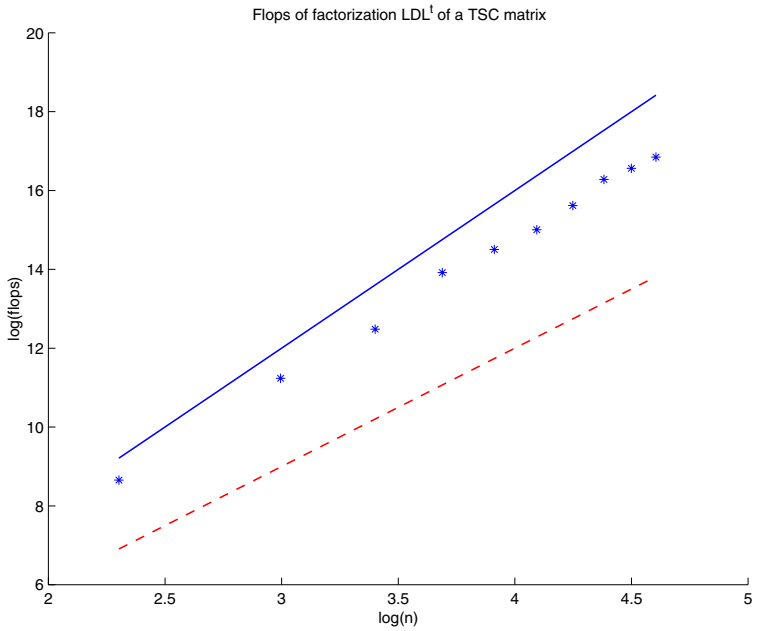


FIG. 1. *Computational cost of LDL^T factorization for symmetric TSC matrices of sizes ranging from* 10 *to* 100.

Proof of Lemma 1. We will prove (a) by induction on k. If $k = 1$, the elements of $\mathcal{C}_1$ are

$$\mathcal{C}_1(i,j) = a_{i+1,j+1} - \frac{a_{i+1,1}\, a_{1,j+1}}{a_{11}} = \frac{a_{i+1,j+1}\, a_{11} - a_{i+1,1}\, a_{1,j+1}}{a_{11}}$$

$$= \frac{\det A([1, i+1], [1, j+1])}{\det A([1], [1])},$$

which is just (13) with $k = 1$. Now, suppose that (13) is true for some $k \in \{1, 2 \ldots n\}$. We will show that then it is also true for $k + 1$. According to Lemma 6, the Schur complement $\mathcal{C}_{k+1}$ of A_{k+1} in A is the result of taking two successive Schur complements: first the Schur complement $\mathcal{C}_k$ of A_k in A and then the Schur complement of the (1, 1) entry in $\mathcal{C}_k$. We know from the induction hypothesis that

$$\mathcal{C}_k(i,j) = \frac{\det A([1:k, k+i], [1:k, k+j])}{\det A([1:k], [1:k])}.$$

Substituting this into the formula

$$\mathcal{C}_{k+1}(i,j) = \mathcal{C}_k(i+1, j+1) - \frac{\mathcal{C}_k(i+1,1)\mathcal{C}_k(1,j+1)}{\mathcal{C}_k(1,1)}$$

for the elements of $\mathcal{C}_{k+1}$ leads to

$$\mathcal{C}_{k+1}(i,j) = \frac{\begin{vmatrix} \det A([1:k, k+i+1], [1:k, k+j+1]) & \det A([1:k, k+1], [1:k, k+j+1]) \\ \det A([1:k, k+i+1], [1:k, k+1]) & \det A([1:k, k+1], [1:k, k+1]) \end{vmatrix}}{\det A([1:k], [1:k])\, \det A([1:k, k+1], [1:k, k+1])}.$$

It suffices to apply Sylvester's identity [19, p. 22] to the numerator to obtain (13) with $k+1$ instead of k.

Once (a) has been proved, all elements of $\mathcal{C}_k([i_1,\dots,i_s],[j_1,\dots,j_s])$ can be written as quotients with the same denominator $d_k = \det A([1:k],[1:k]$. Hence, the submatrix can be written as $(1/d_k)\,M$, where, for each $l,m \in \{1,\dots,s\}$, the element (l,m) of M is $\det A([1:k,k+i_l],[1:k,k+j_m])$. Applying again Sylvester's formula to M proves part (b). ∎

Proof of Theorem 1. The proof is similar to the one of Lemma 5.1 in [11, p. 52]. The entries of D are either entries of A or entries of a Schur complement of A. Hence, by Lemma 1, any entry of D is either an entry of A or a quotient of minors of A. Now consider an entry l_{ij} of L generated by a 1×1 pivot. Then l_{ij} is a quotient of two elements of the corresponding Schur complement of A, and since both elements have been created at the same stage of the factorization algorithm, by part (a) of Lemma 1 they are quotients of the form (13) with the same denominator. Hence, both denominators cancel out in the quotient and l_{ij} is a quotient of minors of A. The argument is similar for the entries of L generated by 2×2 pivots, using part (b) of Lemma 1 instead of part (a). ∎

Proof of Lemma 2.

We distinguish the cases $s_k = 1$ and $s_k = 2$. If $s_k = 1$, then $n_k = n_{k-1}+1$ and

$$L(i,n_k) = \frac{\mathcal{C}_{k-1}(i,1)}{\mathcal{C}_{k-1}(1,1)}, \qquad i \in \{n_k+1,\dots,n\},$$

which, according to part (a) of Lemma 1, is equal, after simplifying, to

$$L(i,n_k) = \frac{\det A([1:n_{k-1},n_{k-1}+i],[1:n_{k-1},n_{k-1}+1])}{\det A([1:n_{k-1},n_{k-1}+1],[1:n_{k-1},n_{k-1}+1])} = \frac{\det A([1:n_{k-1},i],[1:n_k])}{\det A([1:n_k],[1:n_k])}.$$

If $s_k = 2$, then $n_k = n_{k-1}+2$ and, for each $i \in \{n_k+1,\dots,n\}$, we have

$$L(i,n_k-1) = \frac{\begin{vmatrix} \mathcal{C}_{k-1}(i,1) & \mathcal{C}_{k-1}(i,2) \\ \mathcal{C}_{k-1}(1,2) & \mathcal{C}_{k-1}(2,2) \end{vmatrix}}{\begin{vmatrix} \mathcal{C}_{k-1}(1,1) & \mathcal{C}_{k-1}(1,2) \\ \mathcal{C}_{k-1}(2,1) & \mathcal{C}_{k-1}(2,2) \end{vmatrix}}$$

and

$$L(i,n_k) = \frac{\begin{vmatrix} \mathcal{C}_{k-1}(1,1) & \mathcal{C}_{k-1}(2,1) \\ \mathcal{C}_{k-1}(i,1) & \mathcal{C}_{k-1}(i,2) \end{vmatrix}}{\begin{vmatrix} \mathcal{C}_{k-1}(1,1) & \mathcal{C}_{k-1}(1,2) \\ \mathcal{C}_{k-1}(2,1) & \mathcal{C}_{k-1}(2,2) \end{vmatrix}}.$$

In both cases, Sylvester's identity, combined with Lemma 1, lead to the formula in the statement. Finally, the formulas for the elements of D are trivially obtained if we use the fact that the elements of D are either elements of the original matrix or elements of some Schur complement. ∎

Acknowledgments. The authors thank Prof. Froilán M. Dopico for many helpful discussions and for suggesting the pivoting strategy in section 3.1.2. They also thank both authors of reference [15] for making it available to them, since it very much helped to simplify the presentation of section 4.1.

REFERENCES

[1] R. Brualdi and H. Ryser, *Combinatorial Matrix Theory*, Cambridge University Press, Cambridge, UK, 1991.

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

[3] J. R. Bunch and B. Parlett, *Direct methods for solving symmetric indefinite systems of linear equations*, SIAM J. Numer. Anal., 8 (1971), pp. 639–655.

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

[5] R. W. Cottle, *Manifestations of the Schur complement*, Linear Algebra Appl., 8 (1974), pp. 189–211.

[6] J. W. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 2002.

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

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

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

[10] J. W. Demmel and K. Veselić, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1245.

[11] J. W. 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] I. S. Dhillon, *A New $O(n^2)$ Algorithm for the Symmetric Tridiagonal Eigenvalue/Eigenvector Problem*, Doctoral thesis, University of California at Berkeley, Berkeley, CA, 1997.

[13] I. S. Dhillon and B. N. Parlett, *Orthogonal eigenvectors and relative gaps*, SIAM J. Matrix Anal. Appl., 25 (2004), pp. 858–899.

[14] 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.

[15] 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.

[16] Z. Drmač, *Accurate computation of the product-induced singular value decomposition with applications*, SIAM J. Numer. Anal., 35 (1998), pp. 1969–1994.

[17] K. V. Fernando and B. Parlett, *Accurate singular values and differential qd algorithms*, Numer. Math., 67 (1994), pp. 191–229.

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

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

[20] A. Householder, *The Theory of Matrices in Numerical Analysis*, reprint, Dover, New York, 1975.

[21] W. Kahan, *Accurate Eigenvalues of a Symmetric Tridiagonal Matrix*, Technical Report CS-41, Department of Computer Science, Stanford University, July 1966 (revised edition in June 1968).

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

[23] R. Mathias, *Accurate eigensystem computations by Jacobi methods*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 977–1003.

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

[25] I. Slapničar, *Accurate Symmetric Eigenreduction by a Jacobi Method*, Doctoral thesis, Fernuniversität Hagen, Hagen, Germany, 1992.

[26] I. Slapničar, *Componentwise analysis of direct factorization of real symmetric and Hermitian matrices*, Linear Algebra Appl., 272 (1998), pp. 227–275.

[27] I. Slapničar, *Highly accurate symmetric eigenvalues decomposition and hyperbolic SVD*, Linear Algebra Appl., 358 (2003), pp. 387–424.

[28] S. Vavasis, *Stable numerical algorithms for equilibrium systems*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1108–1131.

[29] K. Veselić, *A Jacobi eigenreduction algorithm for definite matrix pairs*, Numer. Math., 64 (1993), pp. 241–269.

[30] K. Veselić and I. Slapničar, *Floating-point perturbations of Hermitian matrices*, Linear Algebra Appl., 195 (1993), pp. 81–116.