Computing Stable Eigendecompositions of Matrices

James Weldon Demmel
Courant Institute of Mathematical Sciences
251 Mercer Street
New York, New York 10012

Submitted by James H. Wilkinson

ABSTRACT

If a matrix T is known only to within a tolerance ϵ (because of measurement or roundoff errors), then it may be difficult to compute an eigendecomposition of T, since its invariant subspaces are discontinuous functions of its entries. In this paper we show how to compute a stable decomposition of an uncertain matrix T which varies continuously and boundedly as T varies in a ball of radius ϵ.

1. INTRODUCTION

If we are given a complex n by n matrix T_0 which we only know to within a tolerance $\epsilon > 0$, what does it mean to compute an eigendecomposition of T_0? By only knowing T_0 to a tolerance ϵ we mean that T_0 is indistinguishable from any matrix in the set

$$\mathrm{T}(\epsilon) \equiv \{T : \|T_0 - T\|_E \leqslant \epsilon\}$$

($\|T\|_E$ denotes the Frobenius norm of the matrix T, although other norms could be used as well). An eigendecomposition of T will be written

$$T = S\theta S^{-1} \tag{1.1}$$

where θ is block diagonal, $\theta = \mathrm{diag}(\theta_1, \dots, \theta_b)$. The spectrum of T, $\sigma(T)$, will be the union of the spectra of the θ_i's: $\sigma(T) = \bigcup_{i=1}^{b} \sigma(\theta_i)$. We would like to produce an eigendecomposition which is valid in some way for all matrices in $\mathrm{T}(\epsilon)$, and gives as much information about all matrices in $\mathrm{T}(\epsilon)$ as possible.

To illustrate and motivate our approach, we indicate how we would decompose $\mathrm{T}(\epsilon)$ for various values of ϵ, where T_0 is the following matrix,

52 Vanderbilt Ave., New York, NY 10017 0024-3795/86/$3.50

essentially in Jordan canonical form:

$$T_0 = \begin{bmatrix} 1 & & & & & \\ & 1.001 & & & & \\ & & 0 & 100 & & \\ & & & 0 & & \\ & & & & 0 & \\ & & & & & -1 \end{bmatrix}$$

(blanks and 0 both denote zero entries). This decomposition tells us several things: that T_0 has four distinct eigenvalues at 1.001, 1, 0, and -1, that each nonzero eigenvalue has a one dimensional invariant subspace (i.e. an eigenvector) associated with it, and that associated with 0 are one two dimensional and one one dimensional invariant subspace.

Let us ask if this information remains valid for all matrices in $T(\epsilon)$ as ϵ increases from 0. As soon as ϵ becomes nonzero, it is no longer true that all matrices in $T(\epsilon)$ have a triple eigenvalue at 0, nor two invariant subspaces associated with eigenvalues near 0. For example,

$$T_1 = \begin{bmatrix} 1 & & & & & \\ & 1.001 & & & & \\ & & 0 & 100 & & \\ & & \epsilon & 0 & & \\ & & & & 0 & \\ & & & & & -1 \end{bmatrix}$$

has three simple eigenvalues at 0, $10\sqrt{\epsilon}$, and $-10\sqrt{\epsilon}$, each with its own eigenvector, and

$$T_2 = \begin{bmatrix} 1 & & & & & \\ & 1.001 & & & & \\ & & 0 & 100 & & \\ & & & 0 & \epsilon & \\ & & & & 0 & \\ & & & & & -1 \end{bmatrix}$$

has one triple eigenvalue at 0 with just one three dimensional invariant subspace associated with it. Thus, all that we can say that is true of all matrices in $T(\epsilon)$ (for ϵ small enough) is that there are three eigenvalues, all of which could simultaneously equal 0, which together have a three dimensional invariant subspace associated with them.

In fact, we can draw four nonintersecting simple closed curves in the complex plane—one around 1.001, one around 1, one around 0, and one around -1—such that any $T \in \mathrm{T}(\epsilon)$ (for ϵ small enough) will have one eigenvalue strictly inside each of the curves round 1.001, 1, and -1, and three inside the curve around 0. Furthermore, it is impossible to draw any larger number of such curves such that each will contain a constant number of eigenvalues in its interior for all $T \in \mathrm{T}(\epsilon)$.

As ϵ increases to 0.00071, we find matrices in $\mathrm{T}(\epsilon)$ which no longer have two simple eigenvalues around 1 and 1.001:

$$T_3 = \begin{bmatrix} 1.0005 & \eta & & & & \\ & 1.0005 & & & & \\ & & 0 & 100 & & \\ & & & 0 & & \\ & & & & 0 & \\ & & & & & -1 \end{bmatrix}.$$

All we can say about T_3 is that it has two eigenvalues at 1.0005 associated with a two dimensional invariant subspace, since for $\eta \neq 0$ but arbitrarily small this subspace cannot be split into two one dimensional subspaces. Thus, when ϵ is a little larger than 0.00071, all we can say about $\mathrm{T}(\epsilon)$ is that there is one three dimensional invariant subspace with three eigenvalues indistinguishable from 0, one two dimensional subspace with two eigenvalues indistinguishable from 1.0005, and one simple eigenvalue at -1.

In particular, one may draw three simple disjoint simple closed curves in the complex plane—one around 0, one around 1.0005, and one around -1—such that any $T \in \mathrm{T}(0.00071)$ will have three eigenvalues inside the first curve, two inside the second, and one inside the third. As before, it is impossible to draw any larger number of such curves such that each one will strictly contain a fixed number of eigenvalues of each $T \in \mathrm{T}(0.00071)$.

For values of ϵ exceeding 0.01, the clustering of eigenvalues changes again. The matrix

$$T_4 = \begin{bmatrix} 1 & & & & & \\ & 1.001 & & & & \\ & & 0 & 100 & & \\ & & 0.01 & 0 & & \\ & & & & 0 & \\ & & & & & -1 \end{bmatrix}$$

has simple eigenvalues at 1.001 and 0, and double eigenvalues at 1 and -1.

Looking at the eigenvalues as functions of the entry containing 0.01 (the 4,3 entry), we see that T_4 has a pair of eigenvalues at $\pm 10\sqrt{T_{44,3}} = \pm 1$ when $T_{44,3} = 0.01$. Thus, if one tries to draw a family of simple disjoint curves in the complex plane separating the eigenvalues of all $T \in \mathrm{T}(\epsilon)$, one finds that all can draw is one curve enclosing all eigenvalues. The eigenvalues "near 0" can no longer be separated from the eigenvalues near -1 nor 1, and neither can the eigenvalue at 1.001 be separated from 1.

When we were previously unable to draw a curve surrounding exactly one eigenvalue (around 1.0005 and around 0), we could find a matrix $T \in \mathrm{T}(\epsilon)$ which indeed contained exactly one multiple eigenvalue within each circle (a double at 1.0005 and a triple at 0). Is it possible to find a matrix with a sextuple eigenvalue in T(0.01)? The answer is no, although we will not prove this here.

This example motivates the following definition of *stable eigendecomposition:* the entries of S and θ_i in (1.1) must be continuous functions of the entries of T as long as $T \in \mathrm{T}(\epsilon)$. In particular, we insist that $\dim(\theta_i) \equiv n_i$ remain constant for $T \in \mathrm{T}(\epsilon)$. This will in general only be possible if the $\sigma(\theta_i)$ are disjoint and remain so for all T in $\mathrm{T}(\epsilon)$. Thus, the approach we take is to estimate, given a possible partitioning $\sigma = \bigcup_{i=1}^{b} \sigma_i$ of T's spectrum into disjoint subsets, the norm of the smallest perturbation of T that makes some $\lambda_i \in \sigma_i$ coalesce with some $\lambda_j \in \sigma_j$ $(j \neq i)$ (we make this more precise in the next section). It will turn out that it suffices to consider partitioning $\sigma(T)$ into only two disjoint sets $\sigma(T) = \sigma_1 \cup \sigma_2$.

In this case we call the norm of the smallest perturbation that makes some $\lambda_1 \in \sigma_1$ coalesce with some $\lambda_2 \in \sigma_2$ the *dissociation* of σ_1 and σ_2, denoted $\mathrm{diss}(\sigma_1, \sigma_2)$. We present new inclusion theorems, i.e. upper and lower bounds on $\mathrm{diss}(\sigma_1, \sigma_2)$, or the norm of the smallest δT that causes an eigenvalue of σ_1 to coalesce with one from σ_2.

Thus, an algorithm for deciding if $\sigma = \bigcup_{i=1}^{b} \sigma_i$ is a stable decomposition will be to test whether the lower bound on $\mathrm{diss}(\sigma_i, \sigma - \sigma_i)$ exceeds ϵ for all i (see Theorem 7.10).

In addition to requiring stability of a decomposition as defined above, we are also interested in demanding another property of our decomposition, namely that the invariant subspaces (spaces spanned by the columns of S corresponding to each θ_i) not change very much as T varies inside $\mathrm{T}(\epsilon)$. Equivalently, we may ask that the condition number of S in (1.1) not exceed some threshold for any $T \in \mathrm{T}(\epsilon)$. Numerically, this means limiting the precision lost in computing S or θ to a given number of bits (specified by the user). Thus the user may ask given a matrix T_0, an uncertainty ϵ, and a tolerance TOL to compute a stable decomposition (1.1) of $\mathrm{T}(\epsilon)$ such that $\kappa(S) < \mathrm{TOL}$ for all $T \in \mathrm{T}(\epsilon)$. (Here TOL equals 2^B, where B is the number of bits of precision the user is willing to lose.) Theorem 7.10 provides a criterion for deciding if a given partitioning $\sigma(T) = \bigcup_{i=1}^{b} \sigma_i$ satisfies these constraints.

In Section 2 we define basic notation. Section 3 defines stable decomposition precisely. Section 4 defines projections, sep and sep_λ, which are basic quantities on which our bounds on $\text{diss}(\sigma_1, \sigma_2)$ depend. Section 5 presents a lower bound on $\text{diss}(\sigma_1, \sigma_2)$ and compares it with previous ones in the literature. Section 6 presents an upper bound on $\text{diss}(\sigma_1, \sigma_2)$ and compares it with previous results also. In Section 7 we present bounds on the condition number of S in our decomposition (1.1) of $T \in \mathrm{T}(\epsilon)$.

In a later paper we intend to extend these results to matrix pencils, including new inclusion theorems on generalized eigenvalues and deflating subspaces analogous to those results presented here. The results in this paper are largely part of the author's Ph.D. dissertation [3]. For existing software for computing the Jordan canonical form see [9].

2. NOTATION

$\|x\|$, $\|x\|_1$, and $\|x\|_\infty$ will denote the Euclidean length of the n-vector x, the one-norm of x, and the infinity-norm of x:

$$\|x\| \equiv \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2},$$

$$\|x\|_1 \equiv \sum_{i=1}^{n} |x_i|,$$

$$\|x\|_\infty \equiv \max_{1 \leqslant i \leqslant n} |x_i|.$$

$\|T\|$, $\|T\|_1$, $\|T\|_\infty$, and $\|T\|_E$ will denote the matrix norms:

$$\|T\| \equiv \sup_{x \neq 0} \frac{\|Tx\|}{\|x\|},$$

$$\|T\|_1 \equiv \max_j \sum_i |T_{ij}|,$$

$$\|T\|_\infty \equiv \max_i \sum_j |T_{ij}|,$$

$$\|T\|_E \equiv \left(\sum_{i,j=1}^{n} |T_{ij}|^2 \right)^{1/2}.$$

We will use the following well-known inequalities:

$$\|A\| \leqslant \|A\|_E \leqslant \sqrt{n}\cdot\|A\|,$$

$$\|AB\| \leqslant \|A\|\cdot\|B\|,$$

$$\|AB\|_E \leqslant \|A\|_E\cdot\|B\|,$$

$$\|AB\|_E \leqslant \|A\|\cdot\|B\|_E.$$

The condition number of T is $\kappa(T) \equiv \|T\|\cdot\|T^{-1}\|$; and $\kappa_1(T)$, $\kappa_\infty(T)$, and $\kappa_E(T)$ denote the condition numbers computed using the other norms of T. The symbol $\sigma_{\min}(T)$ will denote the smallest singular value of the matrix T.

$A \otimes B$ will denote the Kronecker product of two matrices: $A\otimes B \equiv [A_{ij}\cdot B]$. Finally, col A will denote the column vector formed by taking the columns of A and stacking them atop one another from left to right. Thus if A is m by n, col A is mn by 1 with its first m entries being column 1 of A, its second m entries being column 2 of A, and so on.

3. STABLE DECOMPOSITIONS

In this section we define our criterion for being able to stably decompose $T = S\,\mathrm{diag}(\theta_1, \theta_2)S^{-1}$ into two pieces for T in $\mathrm{T}(\epsilon)$, and show how to use this criterion to decompose $T = S\,\mathrm{diag}(\theta_1,\dots,\theta_b)S^{-1}$ into as fine a partition as possible. Let $\mathrm{T}(\epsilon)$ be defined by a norm $|||\cdot|||$:

$$\mathrm{T}(\epsilon) \equiv \{T' : |||T - T'||| \leqslant \epsilon\}.$$

The following definitions depend on this choice of norm. As T' varies continuously inside $\mathrm{T}(\epsilon)$, its eigenvalues will also vary continuously. Let $\sigma_i = \sigma(\theta_i)$, $i = 1,2$, be a partition of T's spectrum into disjoint pieces.

DEFINITION 3.1. The *dissociation* $\mathrm{diss}(\sigma_1, \sigma_2, T, |||\cdot|||)$ of σ_1 and σ_2 ($\sigma_1 \cap \sigma_2 = \varnothing$) [or just $\mathrm{diss}(\sigma_1, \sigma_2)$ if T and $|||\cdot|||$ are clear from context] is the smallest perturbation $T + E$ of T (measured as $|||E|||$) that makes an eigenvalue $\lambda_1 \in \sigma_1$ coalesce with $\lambda_2 \in \sigma_2$.

To understand this definition, think of λ_1 and λ_2 as continuous functions of E, where $\lambda_i(E)$ is an eigenvalue of $A + E$ and $\lambda_i(0) = \lambda_i$. As $|||E|||$ is allowed to increase, the $\lambda_i(E)$'s can sweep out larger and larger but initially disjoint regions of the complex plane. The smallest $|||E|||$ such that $\lambda_1(E) = \lambda_2(E)$ is $\mathrm{diss}(\sigma_1, \sigma_2)$.

In particular, if we can show that the two regions swept out by $\lambda_1(E)$ and $\lambda_2(E)$ are disjoint for $|||E||| < x$, then we will have the lower bound $x \leqslant \mathrm{diss}(\sigma_1, \sigma_2)$. This is how we will obtain a lower bound on $\mathrm{diss}(\sigma_1, \sigma_2)$ later. However, the meeting of these two regions does not guarantee that two eigenvalues from σ_1 and σ_2 can be made to meet, since they needn't occupy the same region of the plane for the same E. Indeed, by choosing the norm $|||\cdot|||$ to have a sufficiently elongated unit ball, it is easy to find an example where the regions swept out by $\lambda_1(E)$ and $\lambda_2(E)$ meet for an ϵ arbitrarily smaller than that required to cause a double eigenvalue. [Consider $T = \mathrm{diag}(1,2)$, and let the unit ball of $|||\cdot|||$ be highly elongated in the direction of the identity matrix.]

For the norms $\|\cdot\|$ and $\|\cdot\|_E$ we do not know if the regions containing σ_1 and σ_2 can overlap before a multiple eigenvalue appears. To get an upper bound, then, we need to exhibit a perturbation that causes T to have a multiple eigenvalue formed from one eigenvalue from each σ_i. This is how we will obtain an upper bound on $\mathrm{diss}(\sigma_1, \sigma_2)$ later.

We can use the dissociation to partition $\sigma(T)$ into arbitrarily many pieces by noting that if $\mathrm{diss}(\sigma_1, \sigma_2) > \epsilon$ and $\mathrm{diss}(\sigma_1', \sigma_2') > \epsilon$, then the dissociations of the intersections of the σ_i and σ_j' also exceed ϵ:

$$\mathrm{diss}\left(\sigma_i \cap \sigma_j', \sigma - \sigma_i \cap \sigma_j'\right) > \epsilon .$$

This holds because if no perturbation of size ϵ can make an eigenvalue from σ_i coalesce with an eigenvalue from σ_j nor an eigenvalue from σ_j' coalesce with one from σ_i', then no perturbation of that size can make an eigenvalue from $\sigma_i \cap \sigma_j'$ coalesce with one in its complement.

This immediately yields

LEMMA 3.2. *The decomposition* $\sigma = \bigcup_{i=1}^{b} \sigma_i$ *is stable if and only if* $\mathrm{diss}(\sigma_i, \sigma - \sigma_i) > \epsilon$ *for* $i = 1, \dots, b$.

Henceforth $\mathrm{diss}(\sigma_1, \sigma_2)$ will be measured with respect to the $\|\cdot\|_E$ norm.

4. PROJECTIONS, sep, AND sep_λ

In this section we summarize the properties of projections, sep, and sep_λ we will need for our bounds. We will refer to the literature for most proofs.

By Schur's lemma [7] we can put any matrix into upper or lower triangular form by a unitary similarity: $A = QUQ^*$ where $QQ^* = I$ and U is upper (or lower) triangular. Furthermore we may select the order in which the eigenvalues of A appear on the diagonal of U [12]. Such a unitary change of coordinates changes neither distances between matrices (measured with $\|\cdot\|$ or $\|\cdot\|_E$) nor angles between subspaces, so we may assume given any T that it is in the block upper triangular form

$$T=\begin{bmatrix} A & C \\ 0 & B \end{bmatrix} \tag{4.1}$$

where $\sigma(A)=\sigma_1$, $\sigma(B)=\sigma_2$, and A and B may themselves be either upper or lower triangular. We assume $\sigma_1 \cap \sigma_2 = \varnothing$ as above, with $\dim(A)=n_1$ and $\dim(B)=n_2$.

The projection P corresponding to σ_1 is the matrix satisfying $P^2=P$, $TP=PT$, and which projects onto the invariant subspace of A. We exhibit it by solving

$$\begin{bmatrix} I & R \\ 0 & I \end{bmatrix}\begin{bmatrix} A & C \\ 0 & B \end{bmatrix}\begin{bmatrix} I & R \\ 0 & I \end{bmatrix}^{-1}=\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} \tag{4.2}$$

for R, yielding $AR-RB=C$, a set of $n_1 n_2$ linear equations in as many unknowns. We can rewrite $AR-RB=C$ as

$$\left(I_{n_2}\otimes A - B^T\otimes I_{n_1}\right)\operatorname{col} R = \operatorname{col} C. \tag{4.3}$$

By choosing A upper triangular and B lower triangular as described above, we see that $I_{n_2}\otimes A - B^T\otimes I_{n_1}$ will be upper triangular with eigenvalues $\lambda_i(A)-\lambda_j(B)$ for all pairs of eigenvalues of A and B. Therefore Equation (4.3) is solvable for arbitrary C if and only if $\sigma_1 \cap \sigma_2 = \varnothing$ as we assumed.

Given R, we can express the projection P associated with σ_1 as

$$P=\begin{bmatrix} I & R \\ 0 & 0 \end{bmatrix}.$$

The projection associated with σ_2 is $I-P$. Clearly $\|P\|=\|I-P\|$. The left and right invariant subspaces of T belonging to A are spanned by the rows of $[I|R]$ and columns of $[I|0]^T$ respectively. The left and right invariant subspaces of T belonging to B are spanned by the rows of $[0|I]$ and columns of $[-R^T|I]^T$ respectively.

In (4.2) we see that

$$S=\begin{bmatrix} I & -R \\ 0 & I \end{bmatrix}=\begin{bmatrix} I & R \\ 0 & I \end{bmatrix}^{-1}$$

is a block diagonalizing similarity of T which transforms T into a matrix $S^{-1}TS$ whose invariant subspaces corresponding to σ_1 and σ_2 are orthogonal. In [2] the author proved

THEOREM 4.1. *If S is a block diagonalizing similarity of T as described above, then*

$$\kappa(S) \geqslant \|P\| + \left(\|P\|^2 - 1\right)^{1/2} = \left(\|R\|^2 + 1\right)^{1/2} + \|R\|.$$

It is easy to prove

LEMMA 4.2. *Let*

$$S = \begin{bmatrix} I & -R/\|P\| \\ 0 & I/\|P\| \end{bmatrix}.$$

Then S is a block diagonalizing similarity of T whose condition number $\kappa(S)$ attains the lower bound in Theorem 4.1.

We can now define $\operatorname{sep}(A, B)$:

DEFINITION 4.3.

$$\operatorname{sep}(A, B) \equiv \sigma_{\min}\left(I_{n_2} \otimes A - B^T \otimes I_{n_1}\right)$$
$$= \inf_{R \neq 0} \frac{\|AR - RB\|_E}{\|R\|_E}.$$

NOTE. Stewart in [14] defines sep as the smallest singular values for the slightly different operator $A \otimes I_{n_2} - I_{n_1} \otimes B^T$, but it is easy to show that by reordering rows and columns this last matrix is orthogonally similar to the transpose of the matrix in the Definition 4.3, so the definitions are equivalent.

A simple consequence of Definition 4.3 and Equation (4.3) is

LEMMA 4.4.

$$\|P\| \leqslant \left(1 + \frac{\|C\|_E^2}{\operatorname{sep}^2(A, B)}\right)^{1/2}.$$

Some simple properties of sep, proven in [14], are

LEMMA 4.5 (Stewart).

$$\frac{\operatorname{sep}(A,B)}{\kappa(S)\kappa(Q)} \leqslant \operatorname{sep}(SAS^{-1}, QBQ^{-1}) \leqslant \operatorname{sep}(A,B)\kappa(S)\kappa(Q).$$

By choosing S and Q unitary in this lemma, we see that given T, sep(A, B) depends only on σ_1 and σ_2, not the choice of basis. Thus if T is known from context, we may write sep(σ_1, σ_2), or even just sep if σ_1 and σ_2 are known as well.

LEMMA 4.6 (Stewart). *If* $A = \operatorname{diag}(A_1, \dots, A_k)$ *and* $B = \operatorname{diag}(B_1, \dots, B_l)$ *are block diagonal, then*

$$\operatorname{sep}(A,B) = \min_{i,j} \operatorname{sep}(A_i, B_i).$$

Lemmas 4.5 and 4.6 show how to reduce the problem of computing sep to solving a sequence of smaller subproblems. In particular, if A and B are diagonal, sep$(A, B) = \min_{i,j}|\lambda_i(A) - \lambda_j(B)|$, an expression which is clearly also an upper bound on sep(A, B).

Now we define

DEFINITION 4.7. $\operatorname{sep}_\lambda(A, B) = \inf_\lambda \max(\sigma_{\min}(A - \lambda I), \sigma_{\min}(B - \lambda I))$.

NOTE. Varah [16] defines $\operatorname{sep}_\lambda(A, B)$ as the sum of the two singular values rather than the max, so his version of $\operatorname{sep}_\lambda$ differs from ours be at most a factor of 2. Our version allows sharper versions of the bounds on diss(σ_1, σ_2) later.

In words, $\operatorname{sep}_\lambda(A, B)$ is the size of the smallest perturbations E and F that make $A + E$ and $B + F$ have a common eigenvalue. We may relate sep and $\operatorname{sep}_\lambda$ with the following lemma [16]:

LEMMA 4.8.

$$\operatorname{sep}_\lambda(A,B) \leqslant \inf_{\substack{R \neq 0 \\ \operatorname{rank}(R)=1}} \frac{\|AR - RB\|_E}{\|R\|_E} \leqslant 2\operatorname{sep}_\lambda(A,B).$$

Comparing this with the definition of sep we immediately see:

COROLLARY 4.9. $\mathrm{sep}(A, B) \leqslant 2\,\mathrm{sep}_\lambda(A, B)$. *If in addition* $n_1 = \dim(A) = 1$ *or* $n_2 = \dim(B) = 1$ (*i.e. R is a row vector or column vector*), *then*

$$\mathrm{sep}_\lambda(A, B) \leqslant \mathrm{sep}(A, B) \leqslant 2\,\mathrm{sep}_\lambda(A, B). \tag{4.4}$$

Proof. The first inequality follows from Lemma 4.8 and Definition 4.3 of sep. The inequality (4.4) holds because if $n_1 = 1$ or $n_2 = 1$ then R is necessarily of rank one. ∎

In general, sep may be much smaller than sep_λ; this will be important later. One may easily prove the following lemmas, which are analogous to Lemmas 4.5 and 4.6.

LEMMA 4.10.

$$\frac{\mathrm{sep}_\lambda(A, B)}{\max(\kappa(S), \kappa(Q))} \leqslant \mathrm{sep}_\lambda(SAS^{-1}, QBQ^{-1}) \leqslant \mathrm{sep}_\lambda(A, B)\max(\kappa(S), \kappa(Q)).$$

As for sep, Lemma 4.10 implies that given T, σ_1, and σ_2, $\mathrm{sep}_\lambda(A, B)$ is determined independently of the choice of basis that makes T block upper triangular. Thus we may write $\mathrm{sep}_\lambda(\sigma_1, \sigma_2)$ if T is known from context or sep_λ if σ_1 and σ_2 are known as well.

LEMMA 4.11. *If* $A = \mathrm{diag}(A_1, \dots, A_k)$ *and* $B = \mathrm{diag}(B_1, \dots, B_l)$ *are block diagonal, then*

$$\mathrm{sep}_\lambda(A, B) = \min_{i,j} \mathrm{sep}_\lambda(A_i, B_j).$$

Thus, if A and B are diagonal, $\mathrm{sep}_\lambda(A, B) = \min_{i,j} |\lambda_i(A) - \lambda_j(B)|/2$, an expression which is clearly also an upper bound on $\mathrm{sep}_\lambda(A, B)$.

A third characterization of $\mathrm{sep}_\lambda(A, B)$ is as a "structured singular value" of the matrix $I_{n_2} \otimes A - B^T \otimes I_{n_1}$. Recall that this last matrix is singular if and only if A and B have a common eigenvalue. If we allow arbitrary perturbations of $I_{n_2} \otimes A - B^T \otimes I_{n_1}$, the distance to the nearest singular matrix is $\mathrm{sep}(A, B)$. If we only allow perturbations of the form $I_{n_2} \otimes \delta A - \delta B^T \otimes I_{n_1}$, the distance is $\mathrm{sep}_\lambda(A, B)$. Perturbations of this form have the same zero structure as $I_{n_2}\otimes A - B^T\otimes I_{n_1}$ and generally [unless $\dim(A)=1$ or $\dim(B)=1$] cannot have the same structure as the perturbation whose size is $\mathrm{sep}(A, B)$. Thus, sep will generally be smaller than sep_λ, which is why, as we will see in the next section, our lower bound on $\mathrm{diss}(\sigma_1, \sigma_2)$ is larger than previous ones. The concept of structured singular values is also discussed in [4].

5. LOWER BOUNDS ON $\mathrm{diss}(\sigma_1, \sigma_2)$

We give a brief history of lower bounds on $\mathrm{diss}(\sigma_1, \sigma_2)$ to compare them with ours. These earlier bounds are by-products of inclusion theorems on the eigenvalues of a perturbed matrix.

The three earliest theorems, due to Gershgorin, to Dunford and Schwartz, and to Bauer and Fike, all assume T is completely diagonalizable: $STS^{-1} = \Lambda$, where $\Lambda = \mathrm{diag}(\lambda_i)$.

THEOREM 5.1 [1]. *If λ' is an eigenvalue of $T + E$, then λ' is contained in one of the circles*

$$|\lambda' - \lambda| \leqslant \inf_S \kappa(S)\cdot\|E\|,$$

where the infimum is over all S such that $STS^{-1} = \Lambda$.

NOTE. The Bauer-Fike result with $\|E\|$ replaced by $\|E\|_1$ or $\|E\|_\infty$ and $\kappa(S)$ replaced by $\kappa_1(S)$ or $\kappa_\infty(S)$ may be derived by applying Gershgorin's theorem to $S(T+E)S^{-1}$. The Dunford-Schwartz result [5] replaces $\inf_S \kappa(S)$ with $4\max_i\|P_i\|$, where P_i is the one dimensional projection corresponding to λ_i. In [2] the author showed that

$$\max_i\|P_i\| \leqslant \inf_S \kappa(S) \leqslant \dim(T)\max_i\|P_i\|,$$

so the Dunford-Schwartz result is equivalent to the Bauer-Fike except for a constant depending on dimensionality.

By insisting that the circles around eigenvalues in σ_1 do not intersect with circles around eigenvalues in σ_2, we easily derive the following lower bound on diss(σ_1, σ_2):

COROLLARY 5.2.

$$\operatorname{diss}(\sigma_1,\sigma_2) \geqslant \frac{\min_{\lambda_i \in \sigma_i} |\lambda_1 - \lambda_2|}{2 \inf_S \kappa(S)},$$

$$\operatorname{diss}(\sigma_1,\sigma_2) \geqslant \frac{\min_{\lambda_i \in \sigma_i} |\lambda_1 - \lambda_2|}{8 \max_i \|P_i\|}.$$

The disadvantage of this result is that $\kappa(S)$ (or equivalently $\max_i\|P_i\|$) may be very large because the matrix may be nearly nondiagonalizable even if the eigenvalues are well separated. For example, consider

$$T = \begin{bmatrix} A & C \\ 0 & B \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1+\epsilon \end{bmatrix}, \qquad \sigma_1 = \{0\}, \qquad \sigma_2 = \{1, 1+\epsilon\}. \quad (5.1)$$

The lower bounds in Corollary 5.2 approach 0 as ϵ does, even though (as implied by Theorem 5.4 below) diss(σ_1, σ_2) is at least 0.3 for small ϵ.

Stewart gives another lower bound:

THEOREM 5.3 [14]. *If T is as in Equation* (4.1) *and P is the projection corresponding to either A or B, then*

$$\operatorname{diss}(\sigma_1,\sigma_2) \geqslant \frac{\operatorname{sep}(\sigma_1,\sigma_2)}{4\|P\|}.$$

[This theorem will be proven in Section 7 in the course of bounding the condition number of S in (1.1).] Stewart's bound does not suffer from the problem illustrated by the example in (5.1), in which case the lower bound of Theorem 5.3 is approximately 0.154 for small ϵ. One can, however, construct examples where the lower bound in Corollary 5.2 is stronger (larger) than the one in Theorem 5.3:

$$T = \begin{bmatrix} A & C \\ 0 & B \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1+\epsilon & 1 \\ 0 & 0 & 1+2\epsilon \end{bmatrix}, \qquad \sigma_1 = \{1\}, \qquad \sigma_2 = \{1+\epsilon, 1+2\epsilon\}.$$

(5.2)

For small ϵ, the Bauer-Fike lower bound is easily shown to be about $\epsilon^3/\sqrt{18}$, whereas Stewart's is approximately ϵ^4, smaller by a factor of $\sqrt{18}\,\epsilon$. In practice, of course, one would not likely want to split the cluster of three eigenvalues in T anyway.

Finally, we present our new bound, which is similar in spirit to results in [17] and [11].

THEOREM 5.4. *Let T be as in Theorem 5.3. Then*

$$\operatorname{diss}(\sigma_1, \sigma_2) \geqslant \frac{\operatorname{sep}_\lambda(\sigma_1, \sigma_2)}{\|P\| + (\|P\|^2 - 1)^{1/2}}.$$

Proof. If λ is an eigenvalue of $T + E$ but not of T, then we may write

$$0 = \det(\lambda I - T - E) = \det\left([\lambda I - T]\left[I - (\lambda I - T)^{-1}E\right]\right)$$
$$= \det\left(I - [\lambda I - T]^{-1}E\right).$$

Therefore, choosing S according to Lemma 4.2, we get

$$1 \leqslant \left\|(\lambda I - T)^{-1}E\right\| = \left\|S\begin{bmatrix}\lambda I - A & 0\\ 0 & \lambda I - B\end{bmatrix}S^{-1}E\right\|$$
$$\leqslant \|S\| \max\left(\left\|(\lambda I - A)^{-1}\right\|, \left\|(\lambda I - B)^{-1}\right\|\right) \cdot \|S^{-1}\| \cdot \|E\|$$
$$\leqslant \frac{\kappa(S)\|E\|}{\min\left(\left\|(\lambda I - A)^{-1}\right\|^{-1}, \left\|(\lambda I - B)^{-1}\right\|^{-1}\right)}$$
$$= \frac{\left[\|P\| + (\|P\|^2 - 1)^{1/2}\right]\|E\|}{\min(\sigma_{\min}(\lambda I - A), \sigma_{\min}(\lambda I - B))},$$

where we have used the value of $\kappa(S)$ from Lemma 4.2 and the fact that $\|X^{-1}\|^{-1} = \sigma_{\min}(X)$. Therefore, the eigenvalues of $T + E$ lie in one of two regions, one where

$$\|E\| \geqslant \frac{\sigma_{\min}(\lambda I - A)}{\|P\| + (\|P\|^2 - 1)^{1/2}}$$

and one where

$$\|E\| \geqslant \frac{\sigma_{\min}(\lambda I - B)}{\|P\| + \left(\|P\|^2 - 1\right)^{1/2}}.$$

In order for an eigenvalue from $A(\sigma_1)$ and eigenvalue from $B(\sigma_2)$ to coalesce, these regions must overlap at some λ, yielding

$$\|E\|_E \geqslant \|E\| \geqslant \frac{\max\left(\sigma_{\min}(\lambda I - A), \sigma_{\min}(\lambda I - B)\right)}{\|P\| + \left(\|P\|^2 - 1\right)^{1/2}}.$$

Taking the infimum of the right hand side over all λ yields the result. ∎

Now we compare this new bound with the bounds discussed earlier. From Corollary 4.9 we see that our bound is stronger (larger) than Stewart's bound:

LEMMA 5.5.

$$\frac{\operatorname{sep}(\sigma_1, \sigma_2)}{4 \cdot \|P\|} \leqslant \frac{\operatorname{sep}_\lambda(\sigma_1, \sigma_2)}{\|P\| + \left(\|P\|^2 - 1\right)^{1/2}}.$$

When $n_1 = \dim(A) = 1$ or $n_2 = \dim(B) = 1$, the inequality (4.4) shows the two bounds are almost equivalent:

LEMMA 5.6. *When* $n_1 = \dim(A) = 1$ *or* $n_2 = \dim(B) = 1$, *then*

$$\frac{1}{4} \cdot \frac{\operatorname{sep}(\sigma_1, \sigma_2)}{4\|P\|} \leqslant \frac{\operatorname{sep}_\lambda(\sigma_1, \sigma_2)}{\|P\| + \left(\|P\|^2 - 1\right)^{1/2}} \leqslant \frac{\operatorname{sep}(\sigma_1, \sigma_2)}{4\|P\|}.$$

How much better (larger) can the lower bound in Theorem 5.4 be than the one in Theorem 5.3? We give an example here, and return to the question in the next section. If

$$A = \begin{bmatrix} \epsilon & 1 & & & \\ & \cdot & \cdot & & \\ & & \cdot & \cdot & \\ & & & \cdot & 1 \\ & & & & \epsilon \end{bmatrix} \quad \text{and} \quad B = -A, \tag{5.3}$$

then a simple calculation shows $\operatorname{sep}(A, B)$ is proportional to ϵ^{2n-1} and $\operatorname{sep}_\lambda(A, B)$ is proportional to ϵ^n for small ϵ. Thus, $\operatorname{sep}(A, B)$ is almost the square of $\operatorname{sep}_\lambda(A, B)$.

We present three more examples to show when the lower bound on $\operatorname{diss}(\sigma_1, \sigma_2)$ in Theorem 5.4 is sharp.

LEMMA 5.7. *If T is a 2 by 2 matrix, $\operatorname{diss}(\sigma_1, \sigma_2)$ equals its lower bound in Theorem* 5.4.

Proof. Let

$$T = \begin{bmatrix} a & c \\ 0 & b \end{bmatrix}.$$

There is a unitary matrix Q such that

$$QTQ^* = \begin{bmatrix} \dfrac{a+b}{2} & \gamma \\ \dfrac{b-a}{2\left[p+\left(p^2-1\right)^{1/2}\right]} & \dfrac{a+b}{2} \end{bmatrix},$$

where $p = \|P\|$ and γ is complicated and unimportant. Setting the (2,1) entry of this matrix to zero (a perturbation of the desired minimal size) results in an upper triangular matrix with a double eigenvalue at $(a+b)/2$. ■

COROLLARY 5.8. *If T is normal, then* $\operatorname{diss}(\sigma_1, \sigma_2) = \min_{\lambda_i \in \sigma_i} |\lambda_1 - \lambda_2|/2$.

Proof. This follows from Lemmas 5.7 and 4.11. ■

The last example is when $n_1 = 1$ and B is diagonal, in which case we can show the lower bound in Theorem 5.4 can only be too small, by a factor depending on n_2.

LEMMA 5.9. *If T is in the form of* (4.1) *with $A = [a]$, $B = \operatorname{diag}(b_1, \dots, b_{n_2})$, and $C = [c, \dots, c_{n_2}]$, then*

$$\operatorname{diss}(\sigma_1, \sigma_2) \leqslant 4(1+\sqrt{2})\sqrt{n_2}\, \frac{\operatorname{sep}_\lambda(\sigma_1, \sigma_2)}{\|P\| + \left(\|P\|^2 - 1\right)^{1/2}}.$$

Sketch of proof. Let s minimize $|a - b_s|$; then $\text{sep}_\lambda(\sigma_1, \sigma_2) = |a - b_s|/2$. Let i maximize $p_i \equiv |c_i/(b_i - a)|$; then $\|P\|$ is not too far in value from p_i. There are several cases. If $s = i$, the result follows from Lemma 5.7. If $s \neq i$ and $p_i < 1$, then the result follows by perturbing the submatrix

$$\begin{bmatrix} a & c_s \\ 0 & b_s \end{bmatrix}$$

to have a double eigenvalue at $(a + b_s)/2$. If $s \neq i$ and $p_i > 1$, then the result follows by perturbing the submatrix

$$\begin{bmatrix} a & c_p \\ 0 & b_p \end{bmatrix}$$

to move the eigenvalue a to b_s. ∎

6. UPPER BOUNDS ON $\text{diss}(\sigma_1, \sigma_2)$

As in the last section, we discuss previous upper bounds before presenting ours and comparing. We also discuss how far apart our upper and lower bounds can be.

It is convenient to assume $\|T\|_E = 1$, so that our bounds are for the relative change in T needed to make eigenvalues coalesce. We assume as before that T is in the form (4.1) with $\sigma_1 = \sigma(A)$, $\sigma_2 = \sigma(B)$, $n_1 = \dim(A)$, and $n_2 = \dim(B)$. P will denote the projection corresponding either to σ_1 or σ_2.

Previous upper bounds are due to Kahan [10], Wilkinson [19], and Ruhe [13].

THEOREM 6.1 (Kahan). *If* $\|P\| > (n_2 + 1)^{1/2}$ *then*

$$\text{diss}(\sigma_1, \sigma_2) \leqslant \frac{1.22}{\left(\|P\|^2 - 1\right)^{1/2n_2}}.$$

Kahan's proof yields the further insight that if the singular values of the R attaining the infimum in Definition 4.3 of $\text{sep}(A, B)$ are well separated (i.e., some near $\|R\|$ and the rest near 0), then the exponent $1/2n_2$ in this bound can be replaced by $\frac{1}{2}$, so his upper bound behaves like $1/\|P\|$ for large $\|P\|$.

Wilkinson's result is only for the case $n_2 = 1$, but removes the 1.22 in the numerator of Kahan's result.

Ruhe's bound is not precisely on $\text{diss}(\sigma_1, \sigma_2)$, but rather on the distance from T to the nearest matrix with any double eigenvalue $(\min_i \text{diss}(\{\lambda_i\}, \sigma - \{\lambda_i\}))$:

THEOREM 6.2 [Ruhe]. *Let T be completely diagonalizable with eigenvalues λ_i and projections P_i, $i = 1, \ldots, n$. Then the distance from T to the nearest matrix with a double eigenvalue is at most*

$$\frac{n \max_{i \neq j} |\lambda_i - \lambda_j|}{4\left(\max_i \|P_i\|^{2/(n-1)} - 1\right)}.$$

Kahan's bound immediately yields the following bound on the distance from T to the nearest matrix with a double eigenvalue:

$$\min_i \text{diss}(\{\lambda_i\}, \sigma - \{\lambda_i\}) \leqslant \frac{1.22}{\left(\max_i \|P_i\|^2 - 1\right)^{1/2}}$$

which is sometimes better than Ruhe's bound and sometimes worse.

Our upper bound is a trivial consequence of the definition of sep_λ:

LEMMA 6.3. $\sqrt{2}\,\text{sep}_\lambda(\sigma_1, \sigma_2) \geqslant \text{diss}(\sigma_1, \sigma_2)$.

Combining this with Theorem 5.4 we get

COROLLARY 6.4.

$$\sqrt{2}\,\text{sep}_\lambda(\sigma_1, \sigma_2) \geqslant \text{diss}(\sigma_1, \sigma_2) \geqslant \frac{\text{sep}_\lambda(\sigma_1, \sigma_2)}{\|P\| + \left(\|P\|^2 - 1\right)^{1/2}}.$$

If $T = \text{diag}(A, B)$, i.e. block diagonal, then $\|P\| = 1$ and we know $\text{diss}(\sigma_1, \sigma_2)$ to within a factor of $\sqrt{2}$:

COROLLARY 6.5. *If $T = \text{diag}(A, B)$ then $\sqrt{2}\,\text{sep}_\lambda(A, B) \geqslant \text{diss}(\sigma_1, \sigma_2) \geqslant \text{sep}_\lambda(A, B)$.*

If we were measuring $\text{diss}(\sigma_1, \sigma_2)$ with respect to the $\|\cdot\|$ norm instead of the $\|\cdot\|_E$ norm, we would have $\text{diss}(\sigma_1, \sigma_2) = \text{sep}_\lambda(A, B)$, although we will not prove this here. In this case one could furthermore choose the smallest perturbation to have the same block diagonal structure as T.

To compare our upper and lower bounds, we need

THEOREM 6.6. *If* $\|T\|_E = 1$ *and* $m = \min(n_1, n_2)$, *then*

$$2\,\mathrm{sep}_\lambda(A,B) \geqslant \mathrm{sep}(A,B) \geqslant m^{-1/2} \cdot 2^{1-m}\,\mathrm{sep}_\lambda^m(A,B).$$

Proof. The left hand inequality is just Corollary 4.9, and when $m = 1$, Equation (4.4). Now we assume without loss of generality that $m = n_2 \leqslant n_1$ and B is lower triangular. From Definition 4.3 of sep we know

$$\mathrm{sep}(A,B) = \sigma_{\min}\left(I_{n_2} \otimes A - B^T \otimes I_{n_1}\right) = \left\| \left(I_{n_2} \otimes A - B^T \otimes I_{n_1}\right)^{-1} \right\|^{-1}.$$

When $n_2 = 2$, for example,

$$\left(I_{n_2} \otimes A - B^T \otimes I_{n_1}\right)^{-1} = \begin{bmatrix} \left(A - B_{11} I_{n_1}\right)^{-1} & -B_{21}\left(A - B_{11} I_{n_1}\right)^{-1}\left(A - B_{22} I_{n_1}\right)^{-1} \\ 0 & \left(A - B_{22} I_{n_1}\right)^{-1} \end{bmatrix}.$$

We will bound the norm of this matrix from above (thus bounding sep from below) using the norm $|||X||| \equiv \max_i \sum_j \|X_{ij}\|$, where X contains n_2^2 square subblocks X_{ij}. It is not hard to show

$$\|X\| \leqslant \sqrt{n_2}\,|||X|||.$$

Now consider the n_1 by n_1 block entries of $(I_{n_2} \otimes A - B^T \otimes I_{n_1})^{-1}$. They consist of sums of products whose factors are scalars B_{ij} (which all satisfy $|B_{ij}| \leqslant 1$, since $\|T\|_E = 1$) and matrices $(A - B_{ii} I_{n_1})^{-1}$. Clearly $\sigma_{\min}(A - B_{ii} \cdot I_{n_1})$ is an upper bound on $\mathrm{sep}_\lambda(A, B)$, so

$$\left\| \left(A - B_{ii} \cdot I_{n_1}\right)^{-1} \right\| = \frac{1}{\sigma_{\min}\left(A - B_{ii} I_{n_1}\right)} \leqslant \frac{1}{\mathrm{sep}_\lambda(A,B)}.$$

Using these estimates, it is not hard to show by induction that block (i, j) $(j > i)$ is bounded in $\|\cdot\|$ norm by

$$\mathrm{sep}_\lambda^{-2}(A,B) \cdot \left[\mathrm{sep}_\lambda^{-1}(A,B) + 1\right]^{j-i-1}.$$

Thus

$$\operatorname{sep}^{-1}(A,B)=\left\|\left(I_{n_2}\otimes A-B^T\otimes I_{n_1}\right)^{-1}\right\|\leqslant\sqrt{n_2}\left|\left|\left|\left(I_{n_2}\otimes A-B^T\otimes I_{n_1}\right)^{-1}\right|\right|\right|$$

$$\leqslant\sqrt{n_2}\left(\operatorname{sep}_\lambda^{-1}(A,B)+\sum_{i=0}^{n_2-2}\operatorname{sep}_\lambda^{-2}(A,B)\cdot\left[\operatorname{sep}_\lambda^{-1}(A,B)+1\right]^i\right)$$

$$\leqslant\sqrt{n_2}\operatorname{sep}_\lambda^{-1}(A,B)\cdot\left[1+\operatorname{sep}_\lambda^{-1}(A,B)\right]^{n_2-1}$$

$$\leqslant\sqrt{n_2}\cdot 2^{n_2-1}\operatorname{sep}_\lambda^{-n_2}(A,B)$$

[since $\|T\|_E=1$ implies $\operatorname{sep}_\lambda(A,B)\leqslant 1$]. Thus

$$\operatorname{sep}(A,B)\geqslant n_2^{-1/2}\cdot 2^{1-n_2}\operatorname{sep}_\lambda^{n_2}(A,B),$$

as desired. ■

Since our upper bound is derived by only considering possible perturbations to the A and B blocks of T, we would not expect it to be very strong. Nonetheless, we now show it to be at least approximately as strong as Kahan's and Wilkinson's results (and sometimes much stronger). From Theorem 6.6 and Lemma 4.4 we have (where we assume without loss of generality that $n_2\leqslant n_1$)

$$\operatorname{diss}(\sigma_1,\sigma_2)\leqslant\sqrt{2}\operatorname{sep}_\lambda(A,B)\leqslant n_2^{1/2n_2}\cdot 2^{\frac{3}{2}-1/n_2}\operatorname{sep}^{1/n_2}(A,B)$$

$$\leqslant n_2^{1/2n_2}\cdot 2^{\frac{3}{2}-1/n_2}\left(\frac{\|C\|_E}{\left(\|P\|^2-1\right)^{1/2}}\right)^{1/n_2}$$

$$\leqslant\frac{2^{3/2}\left(\sqrt{n_2}/2\right)^{1/n_2}}{\left(\|P\|^2-1\right)^{1/2n_2}}\qquad(\text{since}\quad\|C\|_E\leqslant 1)$$

$$\leqslant\frac{2^{3/2}\left(\sqrt{11}/2\right)^{1/11}}{\left(\|P\|^2-1\right)^{1/2n_2}}$$

$$\leqslant\frac{2.97}{\left(\|P\|^2-1\right)^{1/2n_2}},$$

which is within a constant factor of 2.97 of Kahan's and Wilkinson's results. The point is that $\mathrm{sep}_\lambda(A, B)$ can be small even when $\|P\| = 1$, so our result can be much stronger.

In the example in (5.3), $\mathrm{sep}_\lambda(A, B)$ behaved like ϵ^n and $\mathrm{sep}(A, B)$ like ϵ^{2n-1}, or almost the square. Experience in constructing examples led us to believe this situation is worst case:

CONJECTURE. If $\|T\|_E = 1$, then there is a constant K such that

$$K\,\mathrm{sep}_\lambda^2(A, B) \leqslant \mathrm{sep}(A, B).$$

As it is, we can only show that for this relation to be violated, both A and B must nearly have a common triple derogatory eigenvalue, a rare situation indeed.

We can also use Theorem 6.6 and Lemma 4.4 to compare our upper bound and lower bound in Corollary 6.4. Abbreviating the upper bound $\sqrt{2}\,\mathrm{sep}_\lambda(\sigma_1, \sigma_2)$ by u.b. and the lower bound $\mathrm{sep}_\lambda(\sigma_1, \sigma_2)/[\|P\| + (\|P\|^2 - 1)^{1/2}]$ by l.b., we can prove

THEOREM 6.7. *If* $m = \min(n_1, n_2)$, *then*

$$\text{u.b.} \geqslant \text{l.b.} \geqslant \left(\sqrt{m}\cdot 2^{3(m+1)/2}\right)^{-1}(\text{u.b.})^{m+1}.$$

(Note that $\|T\|_E = 1$, so that u.b. $\leqslant 1$.)

Proof. From Lemma 4.4 and the facts that $\|C\|_E \leqslant 1$ and $\mathrm{sep}(A, B) \leqslant 2$ (since $\|T\|_E = 1$), we see

$$\|P\| + \left(\|P\|^2 - 1\right)^{1/2} \leqslant 1 + \frac{2\|C\|_E}{\mathrm{sep}(A, B)} \leqslant \frac{4}{\mathrm{sep}(A, B)}.$$

From Theorem 6.6 then

$$\|P\| + \left(\|P\|^2 - 1\right)^{1/2} \leqslant \left(\sqrt{m}\cdot 2^{m+1}\right)\mathrm{sep}_\lambda^{-m}(A, B),$$

so

$$\text{l.b.} = \frac{\text{sep}_\lambda(A,B)}{\|P\| + \left(\|P\|^2 - 1\right)^{1/2}} \geqslant \left(\sqrt{m}\cdot 2^{3(m+1)/2}\right)^{-1} (\text{u.b.})^{m+1},$$

as desired. ■

Therefore, although the upper and lower bounds can be quite far apart, they cannot be arbitrarily far apart. In the last section we saw that the lower bound was a good estimator of $\text{diss}(\sigma_1, \sigma_2)$ in a number of interesting cases. We close with an example where the upper and lower bound are almost as far apart as permitted by the last theorem and where the upper bound is a better estimator of $\text{diss}(\sigma_1, \sigma_2)$ than the lower bound:

$$T = \begin{bmatrix} A & C \\ 0 & B \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ & & \epsilon & 1 & & & \cdot \\ & & & \cdot & \cdot & & \cdot \\ & & & \cdot & \cdot & & \cdot \\ & & & & & \cdot & 1 \\ & & & & & & \epsilon \end{bmatrix}. \tag{6.1}$$

It is easy to see that $\text{sep}_\lambda(A, B)$ is near ϵ^{n_2} and $\|P\|$ is near ϵ^{-n_2} for small ϵ, so our upper bound on $\text{diss}(\sigma_1, \sigma_2)$ is near ϵ^{n_2} and our lower bound is near ϵ^{2n_2}. We claim that $\text{diss}(\sigma_1, \sigma_2)$ is near ϵ^{n_2+1}, so while neither upper nor lower bound is asymptotically correct, the upper bound is a better approximation of $\text{diss}(\sigma_1, \sigma_2)$. We just sketch the proof of this claim. To show $\text{diss}(\sigma_1, \sigma_2)$ is no larger than ϵ^{n_2+1}, perturb the matrix in the $(n_2+1, 1)$ entry; a perturbation of size proportional to ϵ^{n_2+1} causes the eigenvalue at 0 and an eigenvalue initially at ϵ to coalesce at $\epsilon/(n_2+1)$. To show that the perturbation must be at least this big, consider a general perturbation E of T each of whose entries is bounded by a constant times ϵ^x, where $x > n_2 + 1$. Computing the characteristic polynomial of $T + E$ yields (where we change variables to $\lambda = \epsilon u$)

$$\det(\epsilon u I - T - E) = \epsilon^{n_2+1}\left[u\cdot(u-1)^{n_2} + \epsilon^{x-n_2-1} p(u)\right],$$

where $p(u)$ is a polynomial of degree at most $n_2 + 1$ with bounded coefficients. If $x > n_2 + 1$, we see that the eigenvalues of $T + E$ remain in a cluster near $u = 1$ ($\lambda = \epsilon$) and $u = 0$ ($\lambda = 0$). Therefore, if $x > n_2 + 1$, the eigenvalue at 0 cannot coalesce with the eigenvalue at ϵ.

7. BOUNDS ON INVARIANT SUBSPACES AND CONDITION NUMBERS OF BLOCK DIAGONALIZING SIMILARITIES

Insisting for a stable decomposition $T = S\,\mathrm{diag}(\theta_1,\ldots,\theta_b)S^{-1}$ that S and θ_i be continuous functions of T for $T \in \mathrm{T}(\epsilon)$ puts no constraint on the condition number of S, or equivalently, the smallness of the angles between invariant subspaces of T which are spanned by the columns of S acted on by each θ_i. [The first $\dim(\theta_1)$ columns of S span the right invariant subspace of T belonging to σ_1, the next $\dim(\theta_2)$ columns of S span the right invariant subspace of T belonging to σ_2, and so on.] For example, if

$$T_x = \begin{bmatrix} 0 & 1 \\ x & 0 \end{bmatrix}$$

then $\sigma(T_x) = \{+\sqrt{x}, -\sqrt{x}\}$ and $\mathrm{diss}(\{+\sqrt{x}\},\{-\sqrt{x}\}) = x$ by Lemma 5.7. As we decrease x to 0, $\|P\| \approx 1/\sqrt{x} \to \infty$, so the condition number of the best diagonalizing similarity $(\|P\| + (\|P\|^2 - 1)^{1/2})$ from Lemma 4.2 goes to infinity, and the angle between the invariant subspaces corresponding to $\sqrt{x}$ and $-\sqrt{x}$ goes to zero. It thus becomes numerically difficult to compute S accurately [2, 8, 14, 18]. Thus a further condition we might impose on a stable decomposition is a bound on $\kappa(S)$ for all $T \in \mathrm{T}(\epsilon)$, or a bound on how far the invariant subspaces spanned by the columns of S may vary.

To compute such a decomposition we use the approach of Stewart [14], which only works when $\|E\|_E < \mathrm{sep}(\sigma_1, \sigma_2)/4\|P\|$. For completeness, we review some basic definitions and lemmas in [14] first.

Definition 7.1. The largest angle between two subspaces **X** and **Y**, written $\theta_{\max}(\mathbf{X},\mathbf{Y})$, is the largest (acute) angle $\theta(x,y)$ between any two nonzero vectors $x \in \mathbf{X}$ and $y \in \mathbf{Y}$:

$$\theta_{\max}(\mathbf{X},\mathbf{Y}) \equiv \sup_{\substack{x\in \mathbf{X}\\ x\neq 0}} \inf_{\substack{y\in \mathbf{Y}\\ y\neq 0}} \theta(x,y).$$

The smallest angle between **X** and **Y**, written $\theta_{\min}(\mathbf{X},\mathbf{Y})$, is the smallest angle $\theta(x,y)$ between any two nonzero vectors $x \in \mathbf{X}$ and $y \in \mathbf{Y}$:

$$\theta_{\min}(\mathbf{X},\mathbf{Y}) \equiv \inf_{\substack{x\in \mathbf{X}\\ x\neq 0}} \inf_{\substack{y\in \mathbf{Y}\\ y\neq 0}} \theta(x,y).$$

We may compute $\theta_{\max}(\mathbf{X},\mathbf{Y})$ and $\theta_{\min}(\mathbf{X},\mathbf{Y})$ as follows:

LEMMA 7.2 [14]. *Let* $\mathbf{X}$ *be spanned by the columns of* $[I|0]^T$, *and* $\mathbf{Y}$ *by the columns of* $[I|Z]^T$. *Suppose that* $\dim(\mathbf{X}) = \dim(\mathbf{Y})$. *Then*

$$\theta_{\max}(\mathbf{X},\mathbf{Y}) = \arctan\|Z\| = \operatorname{arcsec}\left[\left(1+\|Z\|^2\right)^{1/2}\right].$$

LEMMA 7.3 [2]. *Let* $\mathbf{X}$ *be spanned by the columns of* $[I|0]^T$, *and* $\mathbf{Y}$ *by the columns of* $[Z|I]^T$. *Suppose that* $\dim(\mathbf{X}) = \dim(\mathbf{Y})$. *Then*

$$\theta_{\min}(\mathbf{X},\mathbf{Y}) = \operatorname{arccot}\|Z\| = \operatorname{arccsc}\left[\left(1+\|Z\|^2\right)^{1/2}\right].$$

Recalling the discussion in Section 4, where for

$$T = \begin{bmatrix} A & C \\ 0 & B \end{bmatrix}$$

the left and right invariant subspaces corresponding to A were spanned by $[I|R]$ and $[I|0]^T$ respectively, and the left and right subspaces corresponding to B were spanned by $[0|I]$ and $[-R^T|I]^T$ respectively, we have

COROLLARY 7.4. *The smallest angle between right (or left) subspaces corresponding to A and B is* $\operatorname{arccot}\|R\| = \operatorname{arccsc}\|P\|$, *where P is a projection. The largest angle between left and right subspaces corresponding to A (or to B) is* $\arctan\|R\| = \operatorname{arcsec}\|P\|$.

Thus, as $\|P\|$ gets large, the smallest angle between subspaces of A and B gets small, and the largest angle between left and right subspaces of A (or of B) gets large (close to $\pi/2$).

We will also need to relate norms of projections to condition numbers of block diagonalizing similarities:

THEOREM 7.5 [2]. *Let* $\sigma(T)$ *be partitioned into disjoint sets* $\sigma(T) = \bigcup_{i=1}^{b}\sigma_i$, *where the projection* P_i *corresponds to* σ_i. *Let S be a block diagonalizing similarity as in* (1.1). *Then*

$$\max_{1\leqslant i\leqslant b}\left[\|P_i\|+\left(\|P_i\|^2-1\right)^{1/2}\right] \leqslant \inf_S \kappa(S) \leqslant b \max_{1\leqslant i\leqslant b}\|P_i\|, \tag{7.1}$$

where the infimum is over all block diagonalizing S *such that* $\sigma(\theta_i)=\sigma_i$. *Furthermore, if* S' *is chosen so that the* $\dim(\theta_i)$ *columns spanning the invariant subspaces corresponding to* σ_i *are orthonormal,* $\kappa(S')$ *also satisfies the bounds in* (7.1).

Slight improvements of this theorem may be found in [6] and [15].

The approach taken by Stewart and slightly generalized here is to look for a similarity of the form

$$S=\begin{bmatrix} I & -Z \\ -Q & I \end{bmatrix}\begin{bmatrix} (I-ZQ)^{-1} & 0 \\ 0 & (I-QZ)^{-1} \end{bmatrix}, \qquad S^{-1}=\begin{bmatrix} I & Z \\ Q & I \end{bmatrix} \tag{7.2}$$

such that

$$S^{-1}\begin{bmatrix} A' & C' \\ D' & B' \end{bmatrix}S$$

is block diagonal. (Stewart chooses $Z=-Q$ in his approach and seeks block upper triangularity.) Writing out this last matrix product, we get that

$$\begin{bmatrix} A'+ZD'-C'Q-ZB'Q & -A'Z-ZD'Z+C'+ZB' \\ QA'+D'-QC'Q-B'Q & -QA'Z-D'Z+QC'+B' \end{bmatrix}$$
$$\times\begin{bmatrix} (I-ZQ)^{-1} & 0 \\ 0 & (I-QZ)^{-1} \end{bmatrix} \tag{7.3}$$

should be block diagonal, implying

$$QA'-B'Q=-D'+QC'Q \quad \text{and} \quad A'Z-ZB'=C'-ZD'Z. \tag{7.4}$$

Stewart gives a sufficient condition for the existence of a solution of (7.4):

THEOREM 7.6 [14]. *Suppose* $\|D'\|_E\cdot\|C'\|_E \leqslant \operatorname{sep}^2(A',B')/4$. *Then both equations in* (7.4) *are solvable with*

$$\|Q\|_E \leqslant \frac{2\cdot\|D'\|_E}{\operatorname{sep}(A',B')} \quad \textit{and} \quad \|Z\|_E \leqslant \frac{2\cdot\|C'\|_E}{\operatorname{sep}(A',B')}.$$

Sketch of proof. Stewart shows that if the conditions of the theorem are satisfied, then the iterations

$$Q_{i+1}A' - B'Q_{i+1} = -D' + Q_iC'Q_i \quad \text{and} \quad A'Z_{i+1} - Z_{i+1}B' = -C' + Z_iD'Z_i$$

are contractions and converge to solutions bounded as above.

The equations (7.4) let us rewrite (7.3) as

$$S^{-1}\begin{bmatrix} A' & C' \\ D' & B' \end{bmatrix} S = \begin{bmatrix} A' + ZD' & 0 \\ 0 & B' + QC' \end{bmatrix}, \tag{7.5}$$

which is block diagonal as desired. Thus $\begin{bmatrix} A' & C' \\ D' & B' \end{bmatrix}$ has right invariant subspaces spanned by the columns of $[I|-Q^T]^T$ and $[-Z^T|I]^T$ as long as $I - ZQ$ and $I - QZ$ are invertible, which we must check. ■

Finally, we need

LEMMA 7.7 [14]. $\operatorname{sep}(A+E, B+F) \geqslant \operatorname{sep}(A, B) - \|E\|_E - \|F\|_E$.

We are now is a position to prove the following theorem, which is a slight extension of Theorem 4.11 in [14]:

LEMMA 7.8. *Let*

$$T = \begin{bmatrix} A & C \\ 0 & B \end{bmatrix}$$

with $\sigma_1 = \sigma(A)$ *and* $\sigma_2 = \sigma(B)$ *disjoint. Suppose* E *satisfies* $\|E\|_E < \operatorname{sep}(\sigma_1, \sigma_2)/(4\|P\|)$. *Let* $x = 4\|P\| \cdot \|E\|_E / \operatorname{sep}(\sigma_1, \sigma_2) < 1$. *Then the largest angle* $\theta_{\max}$ *between the right (or left) invariant subspaces of* T *and* $T + E$ *corresponding to* σ_1 *is bounded by*

$$\theta_{\max} \leqslant \arctan\left\{x \cdot \left[\|P\| + \left(\|P\|^2 - 1\right)^{1/2}\right]\right\} < \frac{\pi - \theta}{2},$$

where $\theta = \operatorname{arccsc}(\|P\|)$ *is the largest angle between left and right subspaces of* T *belonging to* σ_1 *(or* σ_2*).*

Proof. Choose S_0 as in Lemma 4.2, so that $S_0^{-1}TS_0 = \mathrm{diag}(A, B)$. Thus

$$S_0^{-1}(T+E)S_0 = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} + \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}, \tag{7.6}$$

where $\|F\|_E \leqslant \|P\|\cdot\|E\|_E$. In order to apply Theorem 7.6 to the matrix in (7.6), we need

$$\|F_{12}\|_E \cdot \|F_{21}\|_E \leqslant \tfrac{1}{4}\left[\mathrm{sep}(A + F_{11}, B + F_{22})\right]^2,$$

which will be satisfied if

$$\|P\|\cdot\|E\|_E \leqslant \tfrac{1}{2}\left[\mathrm{sep}(A, B) - 2\|P\|\cdot\|E\|_E\right],$$

or

$$\|E\|_E \leqslant \frac{\mathrm{sep}(A, B)}{4\|P\|},$$

which we have assumed. Furthermore,

$$\begin{aligned} \|Q\|_E &\leqslant \frac{2\|P\|\cdot\|E\|_E}{\mathrm{sep}(A, B) - 2\|P\|\cdot\|E\|_E} \\ &\leqslant \frac{2\|P\|\cdot\|E\|_E}{\mathrm{sep}(A, B) - 2\|P\|\mathrm{sep}(A, B)/4\|P\|} \\ &= x < 1, \end{aligned}$$

and similarly $\|Z\|_E \leqslant x < 1$, implying $I - QZ$ and $I - ZQ$ are invertible. Thus $T + E$ is block diagonalized by the similarity $S_0 S$, where S and S_0 are given in Equation (7.2) and Lemma 4.2.

Thus, $T + E$ has a right invariant subspaces spanned by the first $\dim(A)$ columns of $S_0 S$:

$$\begin{bmatrix} I - \dfrac{RQ}{\|P\|} \\ -\dfrac{Q}{\|P\|} \end{bmatrix} (I - ZQ)^{-1}.$$

Postmultiplying these columns by $(I - ZQ)(I - RQ/\|P\|)^{-1}$ yields the equivalent basis

$$\begin{bmatrix} I \\ -Q\|P\|^{-1}\left(I - \frac{RQ}{\|P\|}\right)^{-1} \end{bmatrix}.$$

By Lemma 7.2, the largest angle $\theta_{\max}$ between this last subspace and the unperturbed spaces spanned by $[I|0]^T$ is

$$\begin{aligned} \theta_{\max} &\leqslant \arctan\left[\left\| -Q\cdot\|P\|^{-1}\left(I - \frac{RQ}{\|P\|}\right)^{-1}\right\|\right] \\ &\leqslant \arctan\left(\frac{\|Q\|}{(1 - \|R\|\cdot\|Q\|/\|P\|)\cdot\|P\|}\right) \\ &\leqslant \arctan\left(\frac{x}{(1 - \|R\|/\|P\|)\cdot\|P\|}\right) \\ &\leqslant \arctan\left\{x\cdot\left[\|P\| + \left(\|P\|^2 - 1\right)\right]\right\}, \end{aligned}$$

as claimed. Since $x < 1$

$$\arctan\left\{x\cdot\left[\|P\| + \left(\|P\|^2 - 1\right)\right]\right\} \leqslant \arctan\left[\|P\| + \left(\|P\|^2 - 1\right)\right] = \frac{\pi - \theta}{2}.$$

This completes the proof for the right invariant subspaces corresponding to σ_1. The other cases are analogous. ■

We may also use this technique to estimate the condition number of the similarity which block diagonalizes $T + E$:

LEMMA 7.9. *Let T, E, and x be as in Theorem 7.8. Then if P_E is the projection corresponding to σ_1 for $T + E$, we have*

$$\|P_E\| \leqslant 2\cdot\frac{1 + x}{1 - x}\cdot\|P_0\|.$$

Let $S_E = S_0 S$ be the diagonalizing similarity of Lemma 7.8. Then

$$\kappa(S_E) \leqslant \frac{1+x}{1-x} \cdot \kappa(S_0).$$

Proof. We begin by estimating $\kappa(S_E)$. Using (7.2), we see

$$\begin{aligned}\kappa(S_E) &= \kappa(S_0 S) \leqslant \kappa(S_0)\kappa(S) \\ &\leqslant \kappa(S_0)\frac{[1+\max(\|Z\|, \|Q\|)]^2}{1-[\max(\|Z\|, \|Q\|)]^2} \\ &\leqslant \kappa(S_0)\frac{1+x}{1-x},\end{aligned}$$

as claimed. The inequality for $\|P_E\|$ follows from Theorem 7.5:

$$\|P_E\| \leqslant \kappa(S_E) \leqslant \kappa(S_0)\frac{1+x}{1-x} \leqslant 2 \cdot \frac{1+x}{1-x} \cdot \|P_0\| \qquad \blacksquare$$

We may now summarize our results on computing stable decompositions:

THEOREM 7.10. *Let $\sigma(T_0) = \bigcup_{i=1}^{b} \sigma_i$ be a disjoint partitioning of T_0's spectrum. Let P_i denote the projection corresponding to σ_i. Then if*

$$x \equiv \epsilon \max_{1 \leqslant i \leqslant b} \frac{4\|P_i\|}{\operatorname{sep}(\sigma_i, \sigma - \sigma_i)} < 1, \tag{7.7}$$

$\bigcup_{i=1}^{b} \sigma_i$ is a stable decomposition of $\mathrm{T}(\epsilon)$, and the block diagonalizing similarity S in (1.1) can be chosen such that

$$\kappa(S) \leqslant 2b\kappa(S_0)\frac{1+x}{1-x} \tag{7.8}$$

for all $T \in \mathrm{T}(\epsilon)$. (Here S_0 is any block diagonalizing similarity of T_0.)

If instead of (7.7) the weaker condition

$$\epsilon \max_{1 \leqslant i \leqslant b} \frac{\|P_i\| + \left(\|P_i\|^2 - 1\right)^{1/2}}{\operatorname{sep}_\lambda(\sigma_i, \sigma - \sigma_i)} < 1 \tag{7.9}$$

holds, then $\bigcup_{i=1}^{b}\sigma_i$ *is also a stable decomposition of* $\mathrm{T}(\epsilon)$, *but no bound on* $\kappa(S)$ *can be asserted.*

Proof. That the decomposition $\bigcup_i \sigma_i$ is stable if either condition (7.7) or (7.9) holds is just a simple consequence of Theorems 5.3 and 5.4 and Lemma 3.2. It remains to prove (7.8). This follows from the inequality (7.1) and Lemma 7.9. Letting P_{Ei} denote the projection for $T + E$ corresponding to σ_i, we have

$$\kappa(S_E) \leqslant b \max_i \|P_{Ei}\| \leqslant 2b \cdot \frac{1+x}{1-x} \cdot \|P_i\| \leqslant 2b \cdot \frac{1+x}{1-x} \cdot \kappa(S_0).$$

as desired. ■

REFERENCES

1 F. L. Bauer and C. T. Fike, Norms and exclusion theorems, *Numer. Math.* 2:137–141 (1960).
2 J. Demmel, The condition number of equivalence transformations that block diagonalize matrix pencils, *SIAM J. Numer. Anal.* 20(3):599–610 (June 1983).
3 J. Demmel, A numerical analyst's Jordan canonical form, Ph.D. Dissertation, Univ. of California, Berkeley, June 1983.
4 J. Doyle, Analysis of feedback systems with structured uncertainty, *Proc. IEE — D* 29(6):242–250 (Nov. 1982).
5 Nelson Dunford, Spectral operators, *Pacific J. Math.* 4(2):321–354 (Sept 1954).
6 L. Elsner, Block scaling with optimal Euclidean condition, *Linear Algebra Appl.* 58:69–73 (1984).
7 E. Isaacson and H. B. Keller, *Analysis of Numerical Methods*, Wiley, 1966.
8 B. Kågström, Numerical computation of matrix functions, Report UMINF-58.77, Dept. of Information Processing, Umeå Univ., Umeå, Sweden, 1977.
9 B. Kågström and A. Ruhe, An algorithm for numerical computation of the Jordan normal form of a complex matrix, Report UMINF-51.74, Dept. of Information Processing, Umeå Univ., Umeå, Sweden, 1974.
10 W. Kahan, Conserving confluence curbs ill-condition, Technical Report 6, Computer Science Dept., Univ. of California, Berkeley, 4 Aug. 1972.
11 D. Meyer and K. Veselic, New inclusion theorems for partitioned matrices, *Numer. Math.* 34:431–437 (1980).
12 A. Ruhe, An algorithm for numerical determination of the structure of a general matrix, *BIT* 10:196–216 (1970).
13 A. Ruhe, Properties of a matrix with a very ill-conditioned eigenproblem, *Numer. Math.* 15:57–60 (1970).
14 G. W. Stewart, Error and perturbation bounds for subspaces associated with certain eigenvalue problems, *SIAM Rev.* 15(4):752 (Oct. 1973).

15 P. Van Dooren and P. Dewilde, Minimal cascade factorization of real and complex rational transfer matrices, *IEEE Trans. Circuits and Systems*, CAS-28(5):395 (May 1981).
16 J. M. Varah, On the separation of two matrices, *SIAM J. Numer. Anal.* 16, No. 2 (Apr. 1979).
17 David G. Feingold and Richard S. Varga, Block diagonally dominant matrices and generalizations of the Gerschgorin circle theorem, *Pacific J. Math.* 12(4):1241–1250 (Winter 1962).
18 J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford U.P., 1965.
19 J. H. Wilkinson, Note on matrices with a very ill-conditioned eigenproblem, *Numer. Math.* 19:176–178 (1972).

Received 14 December 1984; revised 3 June 1985