The Definite Generalized Eigenvalue Problem: A New Perturbation Theory[1]

Roy Mathias and Chi-Kwong Li
Department of Mathematics, College of William & Mary, Williamsburg, VA 23187
E-mail: ckli@math.wm.edu, mathias@math.wm.edu
October 22, 2004

Dedicated to Pete Stewart on the occasion of his sixty fifth birthday.

Abstract

Let (A, B) be a definite pair of $n \times n$ Hermitian matrices. That is, $|x^*Ax| + |x^*Bx| \neq 0$ for all non-zero vectors $x \in \mathbb{C}^n$. It is possible to find an $n \times n$ non-singular matrix X with unit columns such that

$$X^*(A + iB)X = \mathrm{diag}(\alpha_1 + i\beta_1, \ldots, \alpha_n + i\beta_n)$$

where α_j and β_j are real numbers. We call the pairs (α_j, β_j) *normalized generalized eigenvalues* of the definite pair (A, B). These pairs have not been studied previously. We rework the perturbation theory for the eigenvalues and eigenvectors of the definite generalized eigenvalue problem $\beta Ax = \alpha Bx$ in terms of these normalized generalized eigenvalues and show that they play a crucial rule in obtaining the best possible perturbation bounds. In particular, in existing perturbation bounds, one can replace most instances of the Crawford number

$$c(A, B) = \min\{|x^*(A + iB)x| : x \in \mathbb{C}^n, x^*x = 1\}$$

with the larger quantity

$$d_{\min} = \min\{|\alpha_j + i\beta_j| : j = 1, \ldots, n\}.$$

This results in bounds that can be stronger by an arbitrarily large factor. We also give a new measure of the separation of the jth eigenvalue from the kth:

$$|(\alpha_j + i\beta_j)\sin(\arg(\alpha_j + i\beta_j) - \arg(\alpha_k + i\beta_k))|.$$

This asymmetric measure is entirely new, and again results in bounds that can be arbitrarily stronger than the existing bounds. We show that all but one of our bounds are attainable. We also show that the Crawford number is the infimum of the distance from a definite pencil, *a fortiori* diagonalizable, to a non-diagonalizable pair.

AMS(MOS) 65F15, 65F35, 15A18, 15A60
Keywords Definite Generalized Eigenvalue Problem, Perturbation Theory, Eigenvalue, Eigenvector, Eigenspace, Simultaneous Diagonalization

[1]Both authors were supported in part by NSF grants DMS-9704534, and DMS-0071994. The work was completed while the second author was supported by an Engineering and Physical Sciences Research Council Visiting Fellowship under grant GR/T08739 at the University of Manchester, UK.
Numerical Analysis Report 457, Manchester Centre for Computational Mathematics, October 2004.

1 Introduction

Let (A, B) be a pair of $n \times n$ Hermitian matrices. We say that it is a *definite* pair if $|x^*Ax| + |x^*Bx| \neq 0$ for all non-zero vectors $x \in \mathbb{C}^n$. We say that (α, β) is a generalized eigenvalue of (A, B) with eigenvector $x \neq 0$ in $\mathbb{C}^n$ if

$$\beta Ax = \alpha Bx.$$

For a definite pair (A, B), there exists an invertible matrix X such that

$$X^*(A + iB)X = \mathrm{diag}(\alpha_1 + i\beta_1, \ldots, \alpha_n + i\beta_n)$$

where (α_j, β_j) are the generalized eigenvalues of (A, B) and the jth column x_j of X is the corresponding eigenvector satisfying $\beta_j Ax_j = \alpha_j Bx_j$.

Clearly, if (α_j, β_j) is a generalized eigenvalue for the pair (A, B), then $(d\alpha_j, d\beta_j)$ is also a generalized eigenvalue for any nonzero d. There are at least three ways to normalize the eigenvalues. The perturbation bounds are highly dependent on the normalization. Stewart observed this as early as 1972 [16, pp. 681-2], but later authors have not explicitly noticed this.

First, one may require that

$$\alpha_j^2 + \beta_j^2 = 1. \tag{1.1}$$

In this case we get $\alpha_j + i\beta_j = e^{i\theta_j}$, where θ_j is the argument of $\alpha_j + i\beta_j$ and we call θ_j an *eigenangle*.[2] Second, in the case that B is positive definite, one can require that

$$\beta_j = 1. \tag{1.2}$$

In this case the resulting α_j is the cotangent of the eigenangle and it is a *generalized eigenvalue* of $Ax = \lambda Bx$.

We propose a new normalization – choose (α_j, β_j) so that the columns of X, the diagonalizing matrix, have unit length. In this paper, we shall focus on this third normalization, and call them the resulting pairs *normalized generalized eigenvalues* of (A, B). The quantity

$$d_j \equiv |\alpha_j + i\beta_j|$$

is important in our discussion. It is useful to identify the pair of real numbers (α, β) with the complex number $\alpha + i\beta$, and the Hermitian pair (A, B) with the general complex matrix $A + iB$.

This new normalization is the first key idea in our approach, however, we will have occasion to use each of the other two normalizations.

The *Crawford number* of (A, B):

$$c(A, B) \equiv \min\{|x^*(A + iB)x| : x \in \mathbb{C}^n, x^*x = 1\} \tag{1.3}$$

[2] One may view (1.1) as requiring $\|(\alpha_j,\ \beta_j)\| = 1$, where the norm is the usual Euclidean norm. Then one may consider allowing other norms. Stewart proposed using the max-norm that is, $\max\{\alpha_j, \beta_j\} = 1$ in [16, p. 680]. The resulting (α_j, β_j) are within a factor of $\sqrt{2}$ of those given by (1.1).

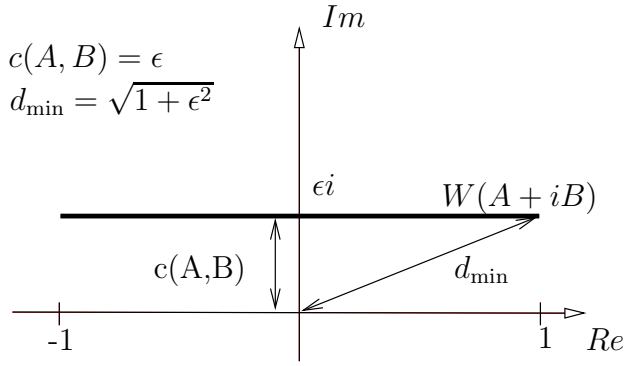

Figure 1: The Crawford number versus $d_{\min}$

is just the distance from the numerical range of $A + iB$ to the origin. A Hermitian pair (A, B) is definite if and only if $c(A, B)$ is positive. We also show (Theorem 2.1) that for a definite pair (A, B), the distance to the boundary of the set of diagonalizable matrices is also $c(A, B)$. Thus one may view $c(A, B)$ as a measure of how definite (A, B) is, and hence, it seems natural that it should appear in perturbation bounds. However, we shall prove perturbation bounds in terms of the normalized generalized eigenvalues and will show that these bounds are stronger than those in terms of the Crawford number. Our contention is that ideally, the Crawford number should not appear in spectral perturbation bounds for the definite generalized eigenvalue problem, though it does determine the "domain of validity" of the error bound.

The second key idea is that since the columns of X are now required to be unit we can reduce almost all perturbation problems to the case of perturbing diagonal pairs where the analysis is much simpler, at the cost of introducing a function of $\|X\| \leq \sqrt{n}$ into the bounds. The condition number of X, that is, $\|X\|\|X^{-1}\|$, which can be very large does not enter our bounds. This diagonalization approach simplifies the analysis, and allows us to replace the Crawford number in bounds by the larger number

$$d_{\min} \equiv \min\{d_j : j = 1, \ldots, n\} = \min\{|\alpha_j + i\beta_j| : j = 1, \ldots, n\}. \tag{1.4}$$

Stewart and Sun [14] prove results in terms of $c(A, B)$, but then observe that these results are not satisfactory , and that the bounds should be in terms of $d_{\min}$, as ours are. To see that $d_{\min}$ can be much larger than the Crawford number consider

Example 1.1 *Take $\epsilon > 0$ and let*

$$A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} \epsilon & 0 \\ 0 & \epsilon \end{pmatrix}.$$

Then $c(A, B) = \epsilon$ while $d_{\min} = \sqrt{1+\epsilon^2} > 1$. Thus, $d_{\min}$ can be larger than $c(A, B)$ by an arbitrarily large factor. (See Figure 1.)

Our bounds will be in terms of the numerical radius:

$$r(T) \equiv \max\{|x^*Tx| : x \in C^n, x^*x = 1\}$$

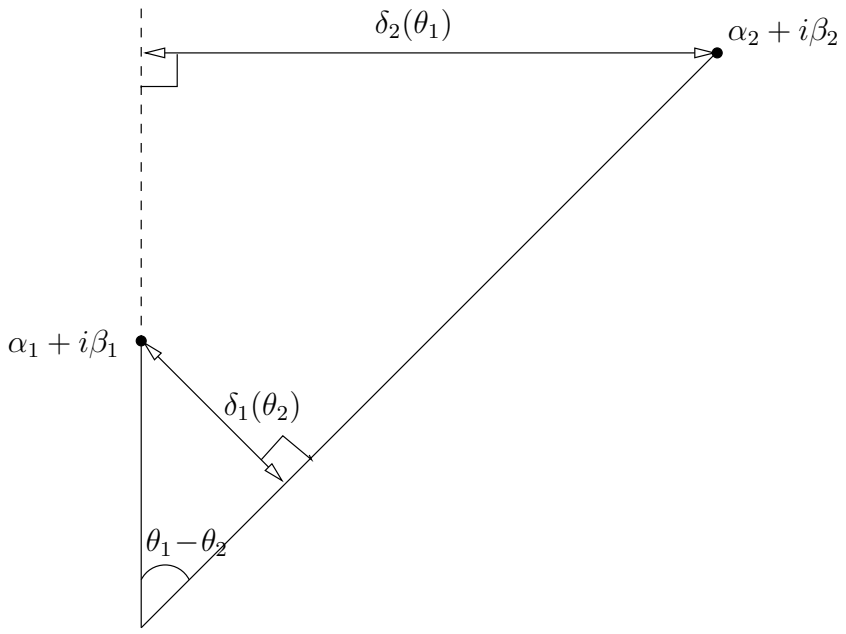

Figure 2: The Separation of $\alpha_1 + i\beta_1$ and $\alpha_2 + i\beta_2$

rather than the more usual spectral norm. This results in cleaner analysis and bounds. The properties of the numerical radius are described in Section 2.1.

Since we are working in terms of the normalized generalized eigenvalues we need a notion of the separation between normalized generalized eigenvalues. We define the separation of $z_j = \alpha_j + i\beta_j = d_j e^{i\theta_j}$ from $z_k = \alpha_k + i\beta_k = d_k e^{i\theta_k}$ to be $|\delta_j(\theta_k)|$ where

$$\delta_j(\theta) \equiv d_j \sin(\theta - \theta_j). \tag{1.5}$$

Notice that $|\delta_j(\theta_k)|$ is *not* symmetric in z_j and z_k. It is the distance from the point z_j in the complex plane to the line through $\pm z_k$. Alternatively,

$$\delta_j(\theta) = \min\{|z| : \arg(z_j + z) = \arg(z_k) (\mathrm{mod} \pi)\}. \tag{1.6}$$

See Figure 2.

Take $\epsilon > 0$ small. Notice that the complex numbers $z_1 = e^{i\epsilon}$ and $z_2 = e^{i(\pi - \epsilon)}$, though well separated on the unit circle are not well separated in our new notion of separation. Note that $|\sin(\theta_j - \theta_k)|$ is just the distance from z_j to z_k in the chordal metric, which is widely used in the analysis of the generalized eigenvalue problem.

We will show that our bounds that use this asymmetric measure of separation are asymptotically optimal in the diagonal case. The usual approach to measuring the separation of z_j from z_k in this context would be to use something like

$$\min\{d_j, d_k\} \, |\sin(\theta_j - \theta_k)|$$

which is smaller that $|\delta_j(\theta_k)|$. Some results are in terms of the even smaller quantity

$$c(A, B) |\sin(\theta_j - \theta_k)|.$$

This is the main difference between our eigenvector perturbation bounds and those in the literature, and is the reason that our bounds are stronger. It appears that $|\delta_j(\theta_k)|$ has not been considered before.

Our approach results in close to optimal perturbation bounds for both eigenangles and eigenvectors. The germ of our bounds for eigenangles was already present in Stewart's 1972 paper, but then later authors were distracted by the apparent elegance of the Crawford number. It appears that our bounds for the perturbation of eigenvectors and eigenspaces are completely new.

In addition to our stronger bounds we answer two open questions from [14, Chapter 6].

Here is a list of the key results and their locations in the paper. They all depend on the use of normalized geneneralized eigenvalues to preserve the "scale of the eigenvalue".

1. The condition number of a simple eigenangle θ_j is d_j^{-1} (Theorems 3.3 and 3.5). Furthermore, d_j^{-1} bounds not only the local sensitivity of θ_j but also the sensitivity to small perturbations, those of size less than a certain easily computable quantity r_j (see the bound (3.6)).

2. We can use $d_{\min}$, or a related quantity involving the d_j's, to replace $c(A, B)$. In fact, $c(A, B)$ does not appear in any of our perturbation bounds! We may sometimes use $c(A, B)$ implicitly in the statement of our bounds to ensure that the perturbation is not so large that we lose definiteness.

3. An ill-conditioned eigenangle, one with a small d_i, can cause a nearby eigenangle to be sensitive to small perturbations. Stewart observed this and called it ill-disposition [16, pp. 685-6]. We explain that the problem is real, but not as serious as one might fear. See Example 3.4 for an instance of "artificial ill-disposition", and Examples 4.2 and 4.3 for instances of true ill-disposition. Theorem 4.6 gives a bound on how severe ill-disposition can be, and Section 4.3 gives an over all discussion of the topic.

4. The separation of the normalized generalized eigenvalue $d_j e^{i\theta_j}$ from $d_k e^{i\theta_k}$ is

$$|\delta_j(\theta_k)| = d_j |\sin(\theta_j - \theta_k)|$$

   not the larger quantity $|d_j e^{-i\theta_j} - d_k e^{-i\theta_k}|$, nor the smaller quantities

$$c(A, B)|\sin(\theta_j - \theta_k)|, \quad \text{or} \quad \min\{d_j, d_k\}|\sin(\theta_j - \theta_k)|$$

   even though these three quantities are symmetric in z_j and z_k. See for example, Theorem 6.1, bound (6.4), and the discussion following the proof of Theorem 6.1.

5. In the diagonal case, the condition number of the jth eigenvector is Δ^{-1} where

$$\Delta = \min\{d_l |\sin(\theta_j - \theta_l)| : \ l \neq j\}$$

   (see Theorem 6.1 and (6.13)). In the general case it is at most $\|X\|^2 \Delta^{-1} \leq n\Delta^{-1}$ (see the discussion at the very end of Section 6).

6. The asymmetry of $\delta_k(\theta_j)$ means that the condition number of two complementary eigenspaces can be very different. This is possible even in the case $n = 2$ (see Example 1.2). In the general case see, for example, Lemma 7.1 and Example 7.5

There are a number of interesting features of the perturbation theory for the definite generalized eigenvalue problem that can be seen even in the 2×2 diagonal case.

Example 1.2 *Consider the Hermitian pair* (A, B) *where*

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \text{ and } B = \begin{pmatrix} 3 & 0 \\ 0 & 10^6 \end{pmatrix}.$$

Its two eigenangles are $\theta_1 = \pi/2$ *and* $\theta_2 = \arctan(3)$*, with normalized generalized eigenvalues* $(0, 10^6)$*, and* $(1, 3)$ *which we identify with the complex numbers* $z_1 = 0 + 10^6 i$ *and* $z_2 = 1 + 3i$*.*

Using the theory developed in this paper we can deduce the following perturbation results

1. The condition number of θ_1 is $|z_1|^{-1} = 10^{-6}$ while the condition number of θ_2 is $|z_2|^{-1} = \frac{1}{\sqrt{10}}$ which is much larger.

2. The condition number of θ_1, the larger eigenangle, is a good measure of the sensitivity of θ_1 for perturbations (E, F) with $r(E + iF) < 1$, however, once $r(E + iF)$ is larger than 1 the condition number of θ_2 is a better measure of the sensitivity of θ_1. This is an instance of "ill-disposition".

3. The condition number of the eigenvector associated with θ_1 is $c_1 \equiv |z_2 \sin(\theta_1 - \theta_2)|$ while the condition number of the eigenvector associated with θ_2 is $c_2 \equiv |z_1 \sin(\theta_1 - \theta_2)|$. Thus, though the eigenvalue θ_1 is better conditioned by a factor of $10^{5.5}$, its eigenvector is worse conditioned by exactly the same factor.

   Consider the perturbation

$$E = \begin{pmatrix} 0 & 0.01 \\ 0.01 & 0 \end{pmatrix}, \quad F = 0.$$

   Then if one computes the eigenvectors of $(A + E, B + F)$ one sees that the eigenvector corresponding to θ_1 changes by about 10^{-2} while the eigenvector corresponding to θ_2 changes by about 3×10^{-8} which is about $10^{-5.5}$ times the change in the first eigenvector.[3]

We shall present several illustrative examples in later sections.

To conclude the introduction here is an outline of the paper. In Section 2 we present some ideas that will be used throughout the paper – the numerical range and numerical radius, rotating the definite generalized eigenvalue problem, the definition, uniqueness and

[3]Of course, this is just a single perturbation, it does not prove the assertion about the conditioning of the eigenvectors.

computation of normalized generalized eigenvalues, the Crawford number, and a discussion of the reduction to the diagonal case.

In Section 3 we present our basic eigenangle perturbation bounds. At the end of the section we present an example of *ill-disposition*–an ill-conditioned eigenvalue causing a nearby eigenvalue to be sensitive to small changes. To better understand this phenomenon we look at perturbation bounds for multiple and clustered eigenangles in Section 4. In Section 5 we quantify the observation that off-diagonal perturbations of a diagonal pair cause only second order perturbations in the eigenangles.

We turn our attention to eigenvector perturbation in Section 6 and eigenspace perturbation in Section 7. Section 7 is rather long, containing an analysis of dif–the analog of sep in the non-symmetric eigenvalue problem–and a version of Stewart's Approximation Theorem ([15, Theorem 3.5], or [14, Theorem V.2.11]), both of which are necessary to obtain our stronger eigenspace perturbation bounds.

2 Preliminaries

Here we present facts about some basic quantities and ideas that arise in this paper: the numerical range, numerical radius, norms, and rotating the problem in Section 2.1; the computation and uniqueness of normalized generalized eigenvalues in Section 2.2, and the reduction to the diagonal case in Section 2.3. Theorem 2.1 gives the distance to non-diagonalizability, and is a new result.

2.1 Numerical range and norms

Recall that the numerical range of an $n \times n$ matrix T is

$$W(T) \equiv \{x^*Tx : x \in \mathbb{C}^n,\ x^*x = 1\},$$

and that the numerical radius of T, $r(T)$, is just the distance of furthest point in the numerical range from the origin:

$$r(T) \equiv \max\{|x^*Tx| : x \in C^n, x^*x = 1\}. \tag{2.1}$$

There are many beautiful results on the numerical range and the numerical radius [2, 3, 5, 6, 9, 10]. We shall use only some elementary results. There are a number of connections between the numerical range and the definite generalized eigenvalue problem:

1. The generalized eigenvalue problem $\beta Ax = \alpha Bx$ (A, B Hermitian) is definite if and only if the numerical range of $A+iB$ does not contain the origin, and by the convexity of the numerical range this is equivalent to the numerical range being contained in an open half plane in the complex plane.

2. If the smallest wedge containing $W(A + iB)$ is $\{z : \theta_1 \le \arg(z) \le \theta_2\}$, and $0 \notin W(A+iB)$, then θ_1 and θ_2 are the extreme eigenangles of (A, B). This is particularly

useful in the 2×2 case since then it is known that $W(A+iB)$ is an elliptical disk with the eigenvalues λ_1 and λ_2 of $A+iB$ as the foci and minor axis of length

$$\sqrt{\mathrm{tr}(A^2+B^2) - |\lambda_1|^2 - |\lambda_2|^2}.$$

See, e.g., [9, Lemma 1.3.3].

3. The normalized generalized eigenvalues are all contained in $W(A+iB)$, and by the previous observation, at least two of them are on the boundary of $W(A+iB)$. (Assuming that the pair (A,B) is at least 2×2.)

4. The Crawford number is the distance from $W(A+iB)$ to the origin, and for a definite pair, it is the distance to the boundary of the set of diagonalizable pairs.

5. If we perturb (A,B) to $(A+E, B+F)$ then $r(E+iF)$ is the appropriate measure of the size of perturbation (E,F).

Numerical analysts tend to prefer to use the *spectral* or *operator norm*, sometimes also called the $2-$*norm*:

$$\|X\| \equiv \max\{\|Xx\| : x \in \mathbb{C}^n,\ x^*x = 1\} = \sqrt{\lambda_{\max}(X^*X)}$$

rather than the numerical radius. The two quantities are very closely related. One can always convert bounds in terms of one into bounds in terms of the other using the fact that for any $n \times n$ matrix A

$$r(A) \le \|A\| \le 2r(A).$$

We prefer the numerical radius because the results one obtains when using it tend to be simpler to state, and only incidentally, very slightly stronger. Here are some other relations between the spectral norm and the numerical radius that we will need. For any Hermitian E and F

$$\|E\| = r(E)$$

and

$$r(E+iF) \le r(E) + r(F) = \|E\| + \|F\| \le \sqrt{2}\left(\|E\|^2 + \|F\|^2\right)^{1/2}.$$

It is easily shown that

$$r(E+iF) = \max\{\|\cos(\theta)E + \sin(\theta)F\| : 0 \le \theta < \pi\},$$

(see, e.g., [9, Section 1.5]) and consequently that for any $\theta \in R$

$$\|\cos(\theta)E + \sin(\theta)F\| \le r(E+iF). \tag{2.2}$$

Note that for any $n \times n$ matrix X, we have

$$r(X^*TX) = \max\{|x^*X^*TXx| : \|x\| = 1\} \le \max\{|y^*Ty| : \|y\| \le \|X\|\} \le \|X\|^2 r(T). \tag{2.3}$$

The numerical radius satisfies the following easily proved submatrix inequalities. Let

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

where A_{11} and A_{22} are square. Then

$$r(\begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}) = \max\{r(A_{11}), r(A_{22})\} \leq r(A), \tag{2.4}$$

and

$$r(\begin{pmatrix} 0 & A_{12} \\ A_{21} & 0 \end{pmatrix}) \leq r(A). \tag{2.5}$$

We will also use the *Frobenius norm*

$$\|X\|_F \equiv \left(\sum_{ij} |x_{ij}|^2 \right)^{1/2} = (\text{tr } X^* X)^{1/2},$$

which is computed more easily than the spectral norm, to which it is related to by

$$\|X\| \leq \|X\|_F \leq \sqrt{n}\|X\|.$$

2.2 Crawford number, distance to non-diagonalizable matrices, rotating pairs

Let us now consider the Crawford number defined in (1.3). One of the important conclusions from this research is that one can replace the Crawford number in bounds by the larger number

$$d_{\min} \equiv \min\{|\alpha_j + i\beta_j| : j = 1, \ldots, n\}.$$

To see that $d_{\min}$ is indeed larger, note that each of the unit generalized eigenvalues $\alpha_j + i\beta_j$ lies in the numerical range of $A + iB$ and the Crawford number is just the distance from $W(A + iB)$ to the origin. Example 1.1 shows that the ratio $c(A, B)/d_{\min}$ can be arbitrarily small for definite pairs.

Another disadvantage of the Crawford number is that it is hard to compute. Until recently, the usual approach was to note that the definition (1.3) is equivalent to

$$c(A, B) = \max_{\theta} \lambda_n(\cos(\theta)A + \sin(\theta)B),$$

and then attempt to solve the single variable maximization problem. Each function evaluation requires the solution of a Hermitian eigenvalue problem – that is, it requires $O(n^3)$ flops[4]. One is interested in bounding $c(A, B)$ away from 0, so if $c(A, B)$ is small one will

[4]A "flop" is a floating point operation.

likely need many function evaluations in order to compute it to a reasonable relative accuracy. Higham, Tisseur and Van Dooren have shown how to compute $c(A, B)$ by bisection, solving an $n \times n$ quadratic eigenvalue problem at each step. [7].

A common approach is to estimate a lower bound on $c(A, B)$ as follows. Suppose that one knows that B is positive definite, and so $c(A, B) \geq \|B^{-1}\|^{-1}$. Further, the first step in the standard direct method for solving the definite generalized eigenvalue problem in this context is to compute a Cholesky factorization of B: $B = LL^*$. Given L one can easily estimate $\|L^{-1}\| = \|B^{-1}\|^{1/2}$, at a cost of $O(n^2)$ flops using a condition estimator.

The Crawford number does however satisfy a very pleasant perturbation bound:

$$c(A + E, B + F) \geq c(A, B) \ - \ r(E + iF),$$

and it is the distance to the boundary of the set of diagonalizable pairs.

Theorem 2.1 *Let (A, B) be a definite pair. Then*

$$c(A, B) = \inf\{r(E + iF) : A + E\text{and } B + F \text{ are not diagonalizable by congruence}\}. \quad (2.6)$$

We need two lemmas to prove Theorem 2.1. The first one is a standard result characterizing simultaneous diagonalizability by congruence.

Lemma 2.2 [8, Table 4.5.15, part 1 (b)] *Let $A, B \in H_n$ with A invertible. Then A and B are simultaneously diagonalizable by congruence if and only if $A^{-1}B$ is diagonalizable and has real eigenvalues.*

By Lemma 2.2, we can prove a limiting case of Theorem 2.1.

Lemma 2.3 *Let $A, B \in H_n$ where $n \geq 2$. If $c(A, B) = 0$ and $W(A + iB)$ is contained in a closed half plane, then for any $\epsilon > 0$ there is a pair $E, F \in H_n$ such that $r(E + iF) < \epsilon$ and the pair $A + E$, $B + F$ is not simultaneously diagonalizable by congruence.*

Proof Let $\phi \in [0, 2\pi)$ be such that $W(e^{i\phi}(A + iB)) = e^{i\phi}W(A + iB)$ lies in the closed upper half plane. Define $\tilde{A}, \tilde{B} \in H_n$ by $\tilde{A} + i\tilde{B} = e^{i\phi}(A + iB)$. We show that there exists $\tilde{E}, \tilde{F} \in H_n$ such that $r(\tilde{E} + i\tilde{F}) < \epsilon$ and the pair $\tilde{A} + \tilde{E}, \tilde{B} + \tilde{F} \in H_n$ are not simultaneously diagonalizable by congruence. Let $E, F \in H_n$ be such that $\tilde{E} + i\tilde{F} = e^{i\phi}(E + iF)$. Then $E, F \in H_n$ will satisfy the desired conclusion. For notational simplicity, we assume that $\phi = 0$, i.e., $(\tilde{A}, \tilde{B}) = (A, B)$.

Note that there is an invertible matrix X such that $X^*AX = \text{diag}(\alpha_1, \ldots, \alpha_n)$ and $X^*BX = \text{diag}(\beta_1, \ldots, \beta_n)$, with $\beta_1 \geq \cdots \geq \beta_n = 0$. We will determine (X^*EX, X^*FX) with

$$r(X^*EX + iX^*FX) \leq \|X\|^2 r(E + iF) < \|X\|^2 \epsilon.$$

Again, for notational simplicity, we may assume that $X = I$.

Note that for two Hermitian matrices $P = P_1 \oplus P_2$ and $Q = Q_1 \oplus Q_2$ with $P_1, Q_1 \in H_k$, P and Q are simultaneously diagonalizable by congruence if and only if P_j and Q_j are

simultaneously diagonalizable by congruence for $j = 1, 2$. Thus, we can focus on a 2×2 submatrix of the diagonal matrix $A + iB$, and find a perturbation of it so that the resulting matrix is not diagonalizable by congruence.

By our reduction, we have $A + iB = \mathrm{diag}(\alpha_1 + i\beta_1, \ldots, \alpha_n + i\beta_n)$, and $W(A + iB)$ is the convex hull of the set $\{\alpha_j + i\beta_j : 1 \le j \le n\}$. We consider two cases:

(a) $W(A+iB)$ touches the real line only at the origin. Then $\alpha_j + i\beta_j = 0$ for $j = 1$ or 2 . Thus, we may permute the rows and columns of $A + iB$ and multiply A by -1 if necessary so that we have $\alpha_1 \ge 0$ and $\alpha_2 + i\beta_2 = 0$.

(b) There is a line segment in $W(A + iB)$ touching the origin. Again, we may permute the rows and columns of $A + iB$ and multiply A by -1 if necessary so that we have $\alpha_1 \ge 0$, $\alpha_2 \le 0$, $\beta_1 = \beta_2 = 0$.

Case (a): Take $\eta > 0$, and set $\tilde{\alpha}_1 = \alpha_1 + \eta$, which is necessarily positive. Set

$$E = \eta \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \text{and} \quad F = \eta \begin{pmatrix} 0 & 1 \\ 1 & (\alpha_1 + \eta)^{-1}\beta_1 \end{pmatrix}.$$

Then

$$C = (A + E)^{-1}(B + F) = \begin{pmatrix} \tilde{\alpha}_1^{-1}\beta_1 & \tilde{\alpha}_1^{-1}\eta \\ -1 & \tilde{\alpha}_1^{-1}\beta_1 \end{pmatrix},$$

which has eigenvalues $\tilde{\alpha}_1^{-1}\beta_1 \pm i(\tilde{\alpha}_1\eta)^{1/2}$. Thus $(A + E, B + F)$ is not diagonalizable. Now take $\eta > 0$ small enough so that $r(E + iF) \le \epsilon$.

Case (b): Take

$$E = \eta \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \text{and} \quad F = \eta \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Then

$$C = (A + E)^{-1}(B + F) = \begin{pmatrix} 0 & (\alpha_1 + \eta)^{-1}\eta \\ (\alpha_2 - \eta)^{-1}\eta & 0 \end{pmatrix}$$

has (non-zero) pure imaginary eigenvalues $\pm\eta[(\alpha_1 + \eta)(\alpha_2 - \eta)]^{-1/2}$ (note that $(\alpha_1 - \eta)(\alpha_2 - \eta) < 0$). By Lemma 2.2 the pair $(A + E, B + F)$ is not simultaneously diagonalizable by congruence. Now choose $\eta > 0$ sufficiently small to ensure $r(E + iF) < \epsilon$. □

Proof of Theorem 2.1 Take $E + iF$ such that $r(E + iF) = c(A, B)$, and $0 \in W((A + E) + i(B + F))$. Set $\tilde{A} = A = E$, and $\tilde{B} = B + F$. Now by Lemma 2.3 there is an arbitrarily small perturbation of $\tilde{A}$ and $\tilde{B}$ that is non-diagonalizable. □

A useful operation is that of *rotating* a pair through an angle ϕ. That is, replace (A, B) by $(\hat{A}, \hat{B}) = (\cos\phi A - \sin\phi B, \sin\phi A + \cos\phi B)$. Note that if A and B are real then so are $\hat{A}$ and $\hat{B}$. One can check that $\hat{A} + i\hat{B} = e^{i\phi}(A + iB)$. From this one can see that rotating a pair through an angle ϕ leaves the eigenvectors unchanged, and that the new eigenangles are $\hat{\theta}_j = \theta_j + \phi$. Clearly, such replacements will not affect our comparison of the values

$(\alpha_j, \beta_j) = (\cos\theta_j, \sin\theta_j)$ and $(\tilde{\alpha}_j, \tilde{\beta}_j) = (\cos\tilde{\theta}_j, \sin\tilde{\theta}_j)$, nor will they change $r(E+iF)$ or $c(A,B)$. In fact, we can choose ϕ so that B is positive definite such that

$$c(A,B) = \lambda_n(B) > r(E+iF),$$

where $\lambda_n(B)$ is the smallest eigenvalue of B. Further, we may arrange α_j and β_j so that

$$-1 < \alpha_1 \le \cdots \le \alpha_n < 1, \quad \text{i.e.,} \quad \pi > \theta_1 \ge \cdots \ge \theta_n > 0.$$

2.3 Computation of normalized generalized eigenvalues and vectors

Proposition 2.4 *Let (A,B) be a Hermitian pair with B is positive definite. Then X is invertible with unit columns satisfying*

$$X^*(A+iB)X = \operatorname{diag}(\alpha_1 + i\beta_1, \ldots, \alpha_n + i\beta_n) \tag{2.7}$$

if and only if $X = B^{-1/2}US$ for some unitary U such that $U^*(B^{-1/2}AB^{-1/2})U$ is diagonal and $S = \operatorname{diag}(1/s_1, \ldots, 1/s_n)$ with $d_j^2 = 1/\beta_j$ is the jth diagonal entry of $U^*B^{-1}U$ for $j = 1, \ldots, n$.

Proof. Suppose X is invertible with unit columns and satisfies (2.7). Comparing the skew-hermitian parts, we see that $X^*BX = \operatorname{diag}(\beta_1, \ldots, \beta_n)$, and hence $U = B^{1/2}XS$ is unitary if $S = (X^*BX)^{-1/2}$. One easily sees that U and S satisfy the asserted conditions. The sufficiency of the proposition can be easily verified. □

Note that the normalized eigenvalues $(\alpha_1, \beta_1), \ldots, (\alpha_n, \beta_n)$ are not uniquely determined for a given definite pair (A,B) with $B > 0$ if $B^{-1/2}AB^{-1/2}$ has repeated eigenvalues. Nevertheless, one may impose an additional assumption on the matrix X so that the resulting normalized eigenvalue pairs will be uniquely determined. We mention a few possibilities in Proposition 2.5. More importantly, in the rest of the paper, we will see that this lack of uniqueness does not compromise the utility of the bounds we derive.

Proposition 2.5 *Let (A,B) be a definite Hermitian pair. Then the diagonal matrix on the right side of (2.7) will be uniquely determined up to permutation if either of the following assumptions is imposed on X.*

1. *X is chosen so that $(\alpha_j, \beta_j) = (\alpha_k, \beta_k)$ whenever $\theta_j = \theta_k$.*
2. *$\det(X^*X)$ has the maximum (or minimum) value among all possible matrices X satisfying (2.7).*

Proof. We may rotate the pair (A,B) to make B positive definite. So we assume that B is positive definite.

Suppose $X = B^{-1/2}US$ satisfies (2.7). If $B^{-1/2}AB^{-1/2}$ has k distinct eigenvalues with multiplicities $n_1, \ldots, n_k$, respectively, we may further assume that $U = [U_1|\cdots|U_k]$ is in

block form so that the columns in U_j span the eigenspace of the jth distinct eigenvalue of $B^{-1/2}AB^{-1/2}$ for $j = 1, \ldots, k$.

For 1, we want to choose U so that $U_j^* B^{-1} U_j$ has constant diagonal entries for $j = 1, \ldots, k$. One can always do that (e.g., see [9, Theorem 1.3.4]) by replacing U_j by a suitable $U_j W_j$, where W_j is an $n_j \times n_j$ unitary matrix, so that $W_j^* U_j^* B^{-1} U_j W_j$ has constant diagonal entries equal to $(\operatorname{tr} U_j^* B^{-1} U_j)/n_j$. Even though the choices of W_j's are not unique, if one modifies X according to U, the resulting diagonal matrix X^*BX will be uniquely determined, namely, the jth diagonal entry equals the reciprocal of that of $X^*B^{-1}X$. One easily checks that the diagonal matrix X^*AX will then be uniquely determined also.

For 2, we want to choose U so that the product of the diagonal entries of $U_j^* B^{-1} U_j$ has minimum or maximum value depending on whether we want $\det(X^*X)$ to be maximum or minimum. It is well-known that the former case happens if $U_j^* B^{-1} U_j$ is in diagonal form, and the latter case happens if $U_j^* B^{-1} U_j$ has constant diagonal entries (as in 1). By arguments similar to those in the preceding paragraph, these cases can be attained and the matrix $X^*(A + iB)X$ will be uniquely determined. □

From the proof of the above result, one sees that the conditions imposed on X can be translated into conditions on the diagonal entries of the matrices $U_j^* B^{-1} U_j$ for $j = 1, \ldots, k$. As a result, one can easily impose other conditions on X that will lead to a unique diagonal matrix $X^*(A + iB)X$.

2.4 Reduction to the diagonal case

One often tries to reduce a problem to the diagonal case since it is much simpler. In the Hermitian eigenvalue problem there is no cost to this transformation since the diagonalization can always be effected using a unitary similarity which preserves norms. In the non-symmetric (or non-Hermitian) eigenvalue problem one cannot always diagonalize the matrix, and even when one can, the similarity may be arbitrarily badly ill-conditioned so the resulting bounds are very weak. We shall see that the definite generalized eigenvalue problem is in-between these two cases, closer to the Hermitian eigenvalue problem since diagonalization introduces a factor of at most n into the bounds. Here are the details.

Let (A, B) be an $n \times n$ definite pair, and let (E, F) be a perturbation such that $r(E+iF) < c(A, B)$. Let X be an invertible $n \times n$ matrix with unit columns such that

$$X^*AX = D_A, \quad \text{and} \quad X^*BX = D_B$$

with D_A and D_B diagonal. Since the columns of X are unit we have

$$\|X\| < \|X\|_F = \sqrt{\operatorname{tr} X^*X} = \sqrt{n}.$$

The inequality is strict because X is required to be invertible.

Now suppose that we want to relate the eigenvalues and eigenvectors of the perturbed pair $(A + E, B + F)$ to those of (A, B). Applying the same transformation to the perturbed pair we have

$$X^*(A + E)X = D_A + X^*EX \equiv D_A + \tilde{E}$$

$$X^*(B+F)X = D_B + X^*FX \equiv D_B + \tilde{F},$$

where

$$\|\tilde{E}\| = \|X^*EX\| \leq \|X\|^2\|E\| \leq n\|E\|$$

$$\|\tilde{F}\| = \|X^*FX\| \leq \|X\|^2\|F\| \leq n\|F\|$$

and

$$r(\tilde{E}+i\tilde{F}) = r(X^*(E+iF)X) \leq \|X\|^2 r(E+iF).$$

Let P be a matrix such that $p_{ii} = 0$ and

$$(I+P)^*(D_A+\tilde{E})(I+P) = D_{\hat{A}}, \quad \text{and} \quad (I+P)^*(D_B+\tilde{F})(I+P) = D_{\hat{B}}$$

and the $D_{\hat{A}}$ and $D_{\hat{B}}$ are diagonal. The eigenangles and eigenvectors of $(D_A+\tilde{E}, D_B+\tilde{F})$ are related as follows:

1. The eigenangles of the pair $(D_A+\tilde{E}, D_B+\tilde{F})$ are the same as those of $(A+E, B+F)$.

2. The generalized eigenvalues of the problem $(D_A+\tilde{E})x = \lambda(D_B+\tilde{F})x$ are the same as those of $(A+E)x = \lambda(B+F)x$.

3. The normalized generalized eigenvalues of the pair $(D_A+\tilde{E}, D_B+\tilde{F})$ *are not necessarily the same as* those of $(A+E, B+F)$.

4. The jth eigenvector of $(D_A+\tilde{E}, D_B+\tilde{F})$ is $e_j + P_j$ while the jth eigenvector of $(A+E, B+F)$ is
$$\tilde{X}_j = X(e_j+P_j) = X_j + \sum_{k\neq j} p_{jk}X_k. \tag{2.8}$$
Here, for any matrix Y let Y_k denote the kth column of Y, and let e_k denote the k the column of I, as is conventional.

Notice, from 1, 2, and 3 above, that eigenangles and generalized eigenvalues are preserved under congruence, but *normalized* generalized eigenvalues are not. The fact that normalized generalized eigenvalues are not preserved under congruence is not a reason not to use them, it is just the price that we must pay to get optimal perturbation bounds. Note also, even if the X_k's are unit, the vector $\tilde{X}_j$ is not necessarily unit. We have the bounds

$$\|X^{-1}\|^{-1}\|(e_j+P_j)\| \leq \|\tilde{X}_j\| \leq \sqrt{n}\|e_j+P_j\| \approx \sqrt{n}.$$

The basic problem is that the set of matrices with unit columns is not closed under multiplication.

We know that we can bound the difference between the eigenvalues and eigenvectors of $(A+E, B+F)$ and those of (A, B) in terms of the difference between those of $(D_A+\tilde{E}, D_B+\tilde{F})$ and those of (D_A, D_B). A natural question arises: are the bounds that we get attainable? They are easily shown to be attainable in the diagonal case. Let $(\tilde{E}, \tilde{F})$ be a perturbation

that attains, or almost attains, the bounds in the diagonal case. We can then back-transform this perturbation to get

$$(E, F) = (X^{-*}\tilde{E}X^{-1}, X^{-*}\tilde{F}X^{-1}).$$

Unfortunately, in doing this we may have greatly increased the norm of the perturbation (E, F) since

$$\|E\| \leq \|X^{-1}\|^2\|\tilde{E}\| \quad \text{and} \quad \|F\| \leq \|X^{-1}\|^2\|\tilde{F}\|$$

and even though we know that X^{-1} exists we have no *a priori* bound on its norm. In short, the attainability of the bounds in the diagonal case does not automatically guarantee the attainability of the bounds in the general case. We use more careful arguments to show that the bounds are approximately attainable.

The basic problem is that the set of matrices with unit columns is not closed under inversion.

Interestingly, ill-conditioning of X is not always bad in the context of perturbation bounds. Suppose that X is invertible, but that all its columns are almost parallel to X_1, its first column. In particular, suppose that

$$X_k = (1 + \epsilon_k^2)^{-1/2}(X_1 + \epsilon_k V_k) \quad k = 2, \cdots, n$$

where V_k is a unit vector orthogonal to X_1, and $\epsilon_k \leq \epsilon$ which is small. We may also assume, without loss of generality, that $p_{ik} \geq 0$. Then the first perturbed eigenvector is

$$\tilde{X}_1 \;=\; X_1 + \sum_{k=2}^{n} \frac{p_{k1}}{\sqrt{1+\epsilon_k^2}} X_1 + \sum_{k=2}^{n} \frac{p_{k1}\epsilon_k}{\sqrt{1+\epsilon_k^2}} V_k = \alpha X_1 + \beta v$$

where

$$\begin{aligned} |\alpha| &= 1 + \sum_{k=2}^{n} |p_{k1}| \geq 1, \\ |\beta| &\leq \sum_{k=2}^{n} |p_{k1}| \leq \sqrt{n}\|P\|\epsilon, \end{aligned}$$

and v is a unit vector, which, being a linear combination of V_2, ..., V_n, is necessarily orthogonal to X_1. Set $\hat{X}_1 \equiv \tilde{X}_1/\alpha = X_1 + (\beta/\alpha)v$. Then, if we let $\theta(u, v)$ denote the angle between the vectors u and v, we have the bound

$$|\tan\theta(X_1, \tilde{X}_1)| = |\tan\theta(X_1, \hat{X}_1)| = \|X_1 - \hat{X}_1\| = |\beta/\alpha| \leq \sqrt{n}\|P\|\epsilon.$$

So, as $\epsilon \to 0$, the matrix X becomes increasing ill-conditioned, but, our bound on the perturbation of the eigenvectors becomes *smaller*.

In the Hermitian or non-Hermitian eigenvalue problems we diagonalize a matrix using a unitary or invertible similarity. Both these classes of matrices are closed under inversion and multiplication. This is not the case for the set of matrices with unit columns that we will use to diagonalize a definite pair. Despite this, the perturbation results we obtain are a powerful argument for the use of diagonalizing matrices with unit columns and the resulting normalized generalized eigenvalues.

3 Eigenangle Perturbation Bounds via min-max

It is known that if θ_j is a simple eigenangle, then its condition number is d_j^{-1}.[5] The condition number gives perturbation bounds for asymptotically small perturbations. In this section we derive perturbation bounds for larger perturbations. A common approach to deriving bounds for larger perturbations from conditioning information is to integrate the condition number. This approach is satisfactory, though not elegant, when the condition number itself does not change dramatically. The d_j's however, can change rapidly:

Example 3.1 *Let*

$$A = \begin{pmatrix} 10^{-4} & 0 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 \\ 0 & 10^2 \end{pmatrix}, \quad E = \begin{pmatrix} -5 \times 10^{-5} & 5 \times 10^{-5} \\ 5 \times 10^{-5} & 0 \end{pmatrix}, \quad \text{and} \quad F = 0.$$

It is easy to see that (A, B) has $d_1 = |\alpha_1 + i\beta_1| \approx 1$ and $d_2 = |\alpha_2 + i\beta_2| = 100$. One can compute $\tilde{d}_1 = |\tilde{\alpha}_1 + i\tilde{\beta}_1| \approx 1.01$ and $\tilde{d}_2 = |\tilde{\alpha}_2 + i\tilde{\beta}_2| \approx 5.10$. Thus a change of the order of 5×10^{-5} in the matrices produces a change of the order of 10^2 in d_2.

Our eigenangle bounds are based on the following basic min-max result, which is just [14, Lemma VI.3.1] with the eigenangles ordered in the opposite order.

Lemma 3.2 *Let (A, B) be a definite pair. Then*

$$\theta_j = \min_{\dim(\mathcal{X})=n-j+1} \max_{x \in \mathcal{X}} \arg(x^*(A + iB)x) \tag{3.1}$$

and

$$\theta_j = \max_{\dim(\mathcal{X})=j} \min_{x \in \mathcal{X}} \arg(x^*(A + iB)x). \tag{3.2}$$

We shall assume throughout this section that A and B are diagonal. This is not a serious assumption since one can always diagonalize them by an X with unit columns. The same transformation applied to the perturbations E and F will increase their norms by a factor of at most $\|X\|^2 < n$. (See Section 2.4.)

Set $r = r(E + iF)$. Let D_j denote the disc in the complex plane centered at $\alpha_j + i\beta_j$ with radius r. One might hope for a Gerschgorin-type result stating that the normalized generalized eigenvalues of $(\tilde{A}, \tilde{B})$ are contained in the union of the D_j's. Example 3.1 shows that this is not the case[6] but these discs do give useful information on the perturbation of eigenangles.

[5] This fact can be derived by considering first order perturbation theory and can be extended beyond the definite generalized eigenvalue problem to the general case generalized eigenvalue problem [14, Section VI.2.1].

[6] One *can* prove a Gerschgorin Theorem for eigenangles as opposed to normalized generalized eigenvalues– see for example [14, Corollary VI.2.5]. Again, this is not a reason to not use normalized generalized eigenvalues. The bound on *eigenangles* given by Theorem 3.3 that involves normalized generalized eigenvalues is actually stronger than a Gerschgorin bound on eigenangles.

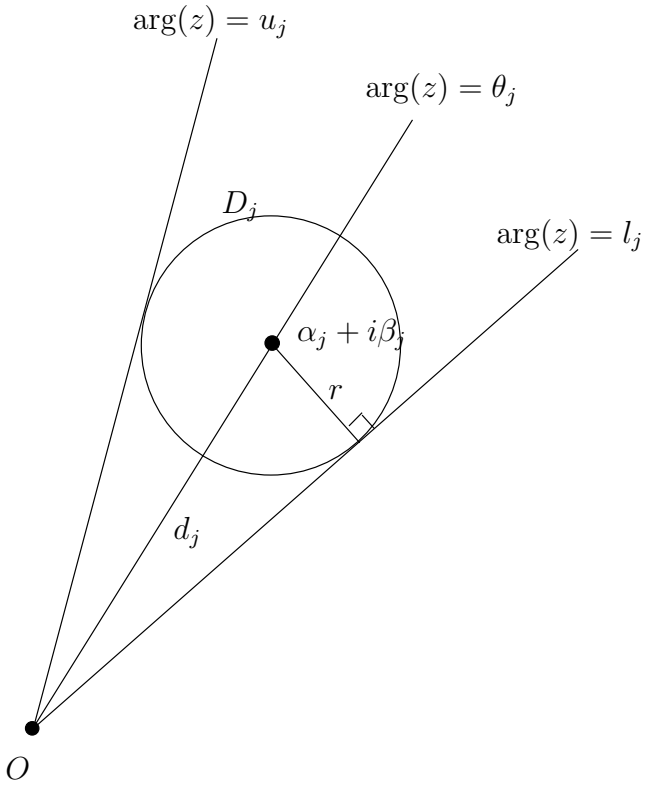

Figure 3: Computation of u_j and l_j

The disc D_j, see Figure 3, is contained in the wedge

$$\{z : l_j \leq \arg z \leq u_j\}$$

where

$$l_j = \theta_j - \sin^{-1}(r/d_j) \tag{3.3}$$

$$u_j = \theta_j + \sin^{-1}(r/d_j). \tag{3.4}$$

We would like all the discs $D_1, \ldots, D_n$ to be contained in an open half plane to ensure that $u_1, \ldots, u_n, l_1, \ldots, l_n$ are contained in an interval of length less than π. This is equivalent to requiring $r < c(A, B)$. Since A and B are diagonal, $c(A, B)$ is easy to compute.

Let us bound the eigenangles $\tilde{\theta}_j$ of $(A + E, B + F)$. Since A and B are diagonalized by $X = I$. Choose $j \in \{1, \ldots, n\}$ and let $\mathcal{I}$ be any subset of $\{1, \ldots, n\}$ of cardinality j. Using the min-max theorem for the arguments of the unit generalized eigenvalues

$$\begin{aligned}
\tilde{\theta}_j &\geq \min_{x \in span\{x_k : k \in \mathcal{I}\}, ||x||=1} \arg x^*((A + iB) + (E + iF))x \\
&= \min_{x \in span\{x_k : k \in \mathcal{I}\}, ||x||=1} \arg[x^*(A + iB)x + x^*(E + iF)x] \\
&\geq \min\{\arg(y + z) : y \in conv\{\alpha_k + i\beta_k : k \in \mathcal{I}\}, z \in W(E + iF)\} \\
&\geq \min\{\arg(y + z) : y \in conv\{\alpha_k + i\beta_k : k \in \mathcal{I}\}, |z| \leq r(E + iF)\} \\
&\geq \min\{\arg w : w \in conv \cup_{k \in \mathcal{I}} D_k\} \\
&= \min\{l_k : k \in \mathcal{I}\}.
\end{aligned}$$

Notice that because the d_k's are not all the same the l_k's are not necessarily in decreasing order. Order them and call the resulting numbers $l_k^{\downarrow}$. Since the analysis above is valid for any index set $\mathcal{I}$ of cardinality j, we have shown

$$\tilde{\theta}_j \geq l_j^{\downarrow}.$$

Let $u_k^{\downarrow}$ denote the ordered u_k's. In the same way we get an upper bound, and hence, we have the result

Theorem 3.3 *Let $A = \mathrm{diag}(\alpha_1, \ldots, \alpha_n)$ and let $B = \mathrm{diag}(\beta_1, \ldots, \beta_n)$. Assume that (A, B) is a definite pair and that $r(E + iF) < c(A, B)$. Then $\tilde{\theta}_1, \ldots, \tilde{\theta}_n$, the eigenangles of $(A + E, B + F)$, satisfy*

$$u_j^{\downarrow} \geq \tilde{\theta}_j \geq l_j^{\downarrow}, \quad j = 1, \ldots, n. \tag{3.5}$$

In Example 3.4 later in the section we present an application of this theorem, and show that it is stronger than a Gerschgorin type theorem.

Notice that the upper and lower bounds $u_j^{\downarrow}$ and $l_j^{\downarrow}$ are completely independent of the Crawford number. The Crawford number appears only in the conditions that ensure the validity of the bound (3.5).

The bounds in (3.5) certainly are easily computable, and given the α_js and β_js the bounds would be easily computable in software. However, the $l_k^{\downarrow}$ notation hides the roles of the size of the perturbation and the moduli of the normalized generalized eigenvalues. Here are two ways to simplify, though slightly weaken, the result and so make it easier to grasp its content.

First suppose that θ_j is a simple eigenangle. (We discuss the condition number of a multiple eigenangle in Section 4.) Recall that $l_j^{\downarrow}$ and $u_j^{\downarrow}$, the ordered l's and u's, defined in (3.3-3.4), are functions of r. Let[7]

$$r_j = \max\{t : l_j^{\downarrow}(r) \geq l_j(r) \ \text{ and } \ u_j^{\downarrow}(r) \leq u_j(r) \ \text{ for all } \ 0 \leq r \leq t\}.$$

Then

$$|\theta_j - \tilde{\theta}_j| \leq \sin^{-1}\left(\frac{r(E + iF)}{d_j}\right), \quad \text{if} \quad r(E + iF) \leq r_j. \tag{3.6}$$

That is, for perturbations (E, F) with $r(E + iF) \leq r_j$, the perturbation in θ_j is bounded by the size of the perturbation multiplied by the condition number. In other words, the condition number, which is based on *purely local* information, gives bounds that are valid in a non-trivial interval.

It is important that the perturbations remain small for $1/d_j$ to give a good measure of the sensitivity of θ_j. We will see this and a number of other points in

[7] It is elementary, though tedious and unenlightening, to give an explicit formula for r_j. The key is the following observation. Suppose that $\theta_j > \theta_i$ and that $d_i < d_j$. Then for small r, $u_j(r) > u_i(r)$, but for large r, $u_j(r) < u_i(r)$. When are they equal? One can check [19] that they are equal when

$$r = \frac{d_i d_j |\sin(\theta_i - \theta_j)|}{d_i^2 + d_j^2 - 2 d_i d_j \cos(\theta_i - \theta_j)}.$$

Example 3.4 *Let*

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1000 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 \\ 0 & 1001 \end{pmatrix}, \quad E = 0, \quad F = \begin{pmatrix} .1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Then (A, B) *has*

$$\theta_1 = \arctan(1.001), \;\; \theta_2 = \arctan(1), \;\; d_1 = \sqrt{1000^2 + 1001^2} \approx 1.4 \times 10^3, \;\; \textit{and} \;\; d_2 = \sqrt{2}.$$

Also, $r = r(E + iF) = .1$. *For* $(\tilde{A}, \tilde{B}) = (A + E, B + F)$ *we have*

$$\tilde{\theta}_1 = \arctan(1.1), \;\; \tilde{\theta}_2 = \arctan(1.001).$$

So

$$|\theta_1 - \tilde{\theta}_1| = 4.7 \times 10^{-2}$$

while $r(E + iF)/d_1 = 7 \times 10^{-5}$ *is much smaller.*

Thus, for perturbations of this size, d_1^{-1} no longer gives a good estimate of the perturbation of θ_1, if fact, d_2^{-1} gives a better estimate: $r(E + iF)/d_2 = 7 \times 10^{-2}$. The ill conditioning of θ_2 has infected the better conditioned eigenangle θ_1. Stewart and Sun observe that this is possible [14, Problem VI.3.1, p324]. Earlier Stewart called this phenomenon *ill-disposition* [16, pp. 682-686, esp. 685].

Plotting the normalized generalized eigenvalues clarifies what has happened here: The eigenangles θ_1 and θ_2 have crossed, and in fact, if we pair $\tilde{\theta}_2$ with θ_1 and $\tilde{\theta}_1$ with θ_2 they do satisfy the bounds in Theorem 3.3:

$$|\tilde{\theta}_2 - \theta_1| = 0 \le \sin^{-1}(.1/d_1), \;\; \text{and} \;\; |\tilde{\theta}_1 - \theta_2| = .048 \le \sin^{-1}(.1/d_2).$$

This suggests that the notion of ill-disposition is just an artifact of the labelling of the eigenangles. In the next section we show that ill-disposition is indeed real – one ill-conditioned eigenangle can cause nearby eigenvalues to be very sensitive to perturbations, and that relabelling eigenangles will not make the problem go away.

Let us compute the bounds on $\tilde{\theta}_1$ and $\tilde{\theta}_2$ from Theorem 3.3 in this example. Evaluating the formulas (3.3-3.4) we get

$$l_1 = 0.785827, \;\; u_1 = 0.785969, \;\; l_2 = 0.714628, \;\; u_2 = 0.856168.$$

Since u_2 is larger than u_1 we have $u_1^{\downarrow} = u_2$ and $u_2^{\downarrow} = u_1$. Theorem 3.3 tells us that

$$0.785827 \le \tilde{\theta}_1 \le 0.856168, \;\; \text{and} \;\; 0.714628 \le \tilde{\theta}_2 \le 0.785969. \tag{3.7}$$

(See Figure 4.) There are several points to make about these bounds. Firstly, the intervals containing $\tilde{\theta}_1$ and $\tilde{\theta}_2$ are not of the same length, nor are they symmetric about θ_1 and θ_2. One can show that each of these intervals taken individually is the smallest possible interval

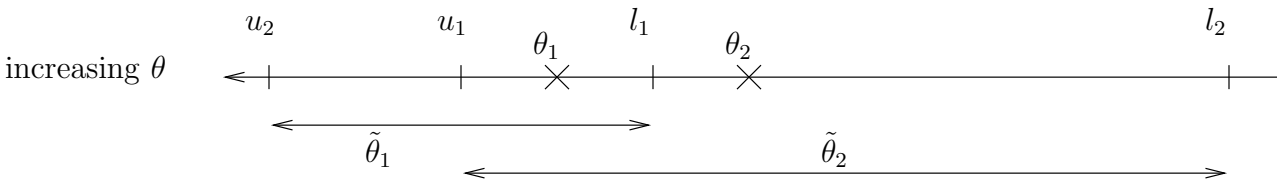

Figure 4: u_j and l_j related to Example 3.4

containing the relevant $\tilde{\theta}_i$ so any intervals that have the "desirable" properties of being the same length or symmetric about the θ_i are necessarily weaker. Secondly, though the intervals $[l_1, u_1]$ and $[l_2, u_2]$ intersect we still get different intervals $[l_1, u_2]$ and $[l_2, u_1]$ that contain $\tilde{\theta}_1$ and $\tilde{\theta}_2$, so Theorem 3.3 is stronger than a Gerschgorin type bound. Thirdly, the bounds (3.7) do not preclude

$$|\tilde{\theta}_1 - \theta_1| \approx |u_2 - \theta_1| = 7 \times 10^{-2}$$

and

$$|\tilde{\theta}_2 - \theta_2| \approx |l_2 - \theta_2| = 7 \times 10^{-2}.$$

That is, at this stage we cannot eliminate the possibility that *both* eigenangles have moved an amount approximately $d_2^{-1} r(E+iF)$. That is, our bounds do not reflect the well conditionedness of θ_1. In the next section we will do a more careful, and more complicated, analysis to show that it is not possible to have both $|\tilde{\theta}_1 - \theta_1| \approx |u_2 - \theta_1|$ and $|\tilde{\theta}_2 - \theta_2| \approx |l_2 - \theta_2|$.

A second approach to simplifying the perturbation bound (3.5) is to notice that

$$l_k^{\downarrow} \geq \theta_k - \sin^{-1}(r/d_{\min}) \quad \text{and} \quad u_k^{\downarrow} \leq \theta_k + \sin^{-1}(r/d_{\min})$$

and hence we have

Theorem 3.5 *Let $A = \operatorname{diag}(\alpha_1, \ldots, \alpha_n)$ and let $B = \operatorname{diag}(\beta_1, \ldots, \beta_n)$. Assume that (A, B) is a definite pair and that $r(E + iF) < c(A, B)$. Then $\tilde{\theta}_1, \ldots, \tilde{\theta}_n$, the eigenangles of $(A + E, B + F)$, satisfy*

$$|\theta_j - \tilde{\theta}_j| \leq \sin^{-1}\left(\frac{r(E+iF)}{d_{\min}}\right). \tag{3.8}$$

Stewart and Sun [14, p. 303] observed that $1/d_j$ is the condition number of θ_j – that is, $1/d_j$ is a measure of θ_j's sensitivity to small perturbations. It appears not to have been observed that $1/d_{\min}$ is a bound on the sensitivity of the eigenangles for larger perturbations.

Stewart and Sun [14, Corollary VI.3.3] give essentially the same result with $d_{\min}$ replaced by $c(A, B)$ which is always smaller and can be much smaller – see Example 1.1. They observe that their results [14, Theorem VI.3.2 and Corollary VI.3.3] are unsatisfactory and that one would expect the bound to involve $d_{\min}$ – just as ours does. Their result is valid for all definite pairs, not just diagonal pairs. We drop the diagonality assumption, at the cost of a factor of n, in Corollary 3.6 and obtain the bound (3.9).

Corollary 3.6 *Assume that (A, B) is a definite pair and that X has unit columns and is such that*

$$X^*(A + iB)X = \mathrm{diag}(\alpha_1 + i\beta_1, \ldots, \alpha_n + i\beta_n).$$

Assume further that

$$r(E + iF)\|X\|^2 < c(\ \mathrm{diag}(\alpha_1, \ldots, \alpha_n), \mathrm{diag}(\beta_1, \ldots, \beta_n)\ ).$$

Then $\tilde{\theta}_1, \ldots, \tilde{\theta}_n$, the eigenangles of $(A + E, B + F)$, satisfy

$$|\theta_j - \tilde{\theta}_j| \leq \sin^{-1}\left(\frac{\|X\|^2 r(E + iF)}{d_{\min}}\right) \leq \sin^{-1}\left(\frac{nr(E + iF)}{d_{\min}}\right). \tag{3.9}$$

4 Multiple and Clustered Eigenangles

We have seen that the condition number of a simple eigenangle θ_j is just d_j^{-1}. We have also seen that if there is an ill-conditioned eigenangle θ_k, with condition number d_k^{-1}, close to θ_j, then the ill-conditioned eigenangle makes the eigenangle θ_j sensitive to perturbations that are (loosely speaking) larger that $d_k|\sin(\theta_j - \theta_k)|$. Let us see what happens when we have a multiple eigenangle.

4.1 Typical perturbations

Consider

Example 4.1

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 2000 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 \\ 0 & 1000 \end{pmatrix}.$$

This pair has a repeated eigenangle $\theta = \arctan(1/2)$. The normalized generalized eigenvalues are not uniquely determined, but we may take them to be $(2 + i)$ and $1000(2 + i)$. When the pair is subjected to random perturbations of norm about .1, one of the two resulting eigenangles differs from θ by about .1, while the other differs by about $.1 \times 10^{-3}$. Thus it would seem that $d_1^{-1} \approx 1$ and $d_2^{-1} \approx 10^{-3}$ are a good indication of the sensitivity of the repeated eigenangle θ. The problem of understanding this phenomenon was mentioned as an open research problem by Stewart and Sun [14, p. 300]. We first show that the issue is more complicated that it appears, and then we explain it.

Take the carefully chosen, non-random perturbations

$$E = \frac{2}{\sqrt{5}}\begin{pmatrix} 0 & .1 \\ 0 & .1 \end{pmatrix}, \quad F = -\frac{1}{\sqrt{5}}\begin{pmatrix} 0 & .1 \\ 0 & .1 \end{pmatrix} \qquad (r(E + iF) = .1). \tag{4.1}$$

This gives two eigenangles $\pi/4 \pm 10^{-3}$. That is, the perturbations in the eigenangles are both of size $10^{-3} = \sqrt{10^{-1} * 10^{-3}}r(E + iF)$, rather than one being $r(E + iF) = 10^{-1}$ and

the other being $10^{-3}r(E+iF) = 10^{-4}$. The perturbation (4.1) is a worst case perturbation, rather than a "typical perturbation". We analyze worst case perturbations in the next subsection, where we also give a more dramatic example of the difference between worst case perturbations and average case perturbations (Example 4.2).

First consider the simple case where *all* the eigenangles of the definite pair (A, B) are identical. Since we may rotate the pair (A, B) and thereby rotate its eigenangles also, we may assume without loss of generality that all the eigenangles of (A, B) are $\pi/2$, or equivalently that their cotangents are 0, which in turn is equivalent to $A = 0$. Recall that the cotangents of the eigenangles are the generalized eigenvalues of the problem $Ax = \lambda Bx$, which are the eigenvalues of the Hermitian matrix $B^{-1/2}AB^{-1/2}$. If ϕ is close to $\pi/2$ then $|\cot(\phi)| \approx |\pi/2 - \phi|$. Thus, for small perturbations of the eigenangles, the perturbation of eigenangles is approximately the same as that of their cotangents. In this section we will look at the perturbation of the cotangents of the eigenangles, that is, the eigenvalues of $B^{-1/2}AB^{-1/2}$.

Since $A = 0$ we may diagonalize the pair (A, B) by a unitary similarity, so assume, without further loss of generality that B is diagonal, and that its diagonal entries are $d_1 \le d_2 \le \cdots \le d_n$, with $d_1 > 0$. The cotangents of the eigenangles of the pair $(A+E, B+F) = (E, B+F)$ are the eigenvalues of $(B+F)^{-1/2}E(B+F)^{-1/2}$, which are the same as those of

$$C = (I + B^{-1/2}FB^{-1/2})^{-1/2}(B^{-1/2}EB^{-1/2})(I + B^{-1/2}FB^{-1/2})^{-1/2}, \tag{4.2}$$

since

$$\begin{aligned} B + F &= B^{1/2}(I + B^{-1/2}FB^{-1/2})B^{1/2} \\ &= [(I + B^{-1/2}FB^{-1/2})^{1/2}B^{1/2}]^*[(I + B^{-1/2}FB^{-1/2})^{1/2}B^{1/2}]. \end{aligned}$$

Ostrowskii's inequality [8, Theorem 4.5.9] tells us

$$\lambda_{\min}(I + B^{-1/2}FB^{-1/2}) \le \frac{\lambda_i(C)}{\lambda_i(B^{-1/2}EB^{-1/2})} \le \lambda_{\max}(I + B^{-1/2}FB^{-1/2}) \tag{4.3}$$

and hence

$$1 - \|B^{-1}\|\|F\| \le \frac{\lambda_i(C)}{\lambda_i(B^{-1/2}EB^{-1/2})} \le 1 + \|B^{-1}\|\|F\|. \tag{4.4}$$

We expect F to be small, so we shall analyze the eigenvalues of $B^{-1/2}EB^{-1/2}$ and then use (4.4) to convert the results to information on the eigenvalues of C. If for example, $r(E+iF) < d_{\min}/2$, then since $\|F\| \le r(E+iF)$, the bound (4.4) tells us that the eigenvalues of C and $B^{-1/2}EB^{-1/2}$ differ by a factor of at most 2. We are now left with the problem of understanding the eigenvalues of the symmetric matrix $B^{-1/2}EB^{-1/2}$.

Notice that if the d_j's vary greatly in magnitude, then $B^{-1/2}EB^{-1/2}$ will be a graded or block graded matrix. Stewart and Zhang have looked at a very similar problem. They explained why, in the non-symmetric eigenvalue problem, it may happen that a multiple eigenvalue, may split into several eigenvalues that lie at different distances from the original

eigenvalue. These distances are, they show, "more a characteristic of the matrix than of the perturbation". They have presented an admirable explanation as to why typically the magnitudes of the eigenvalues of $B^{-1/2}EB^{-1/2}$ will be of the order of

$$d_1^{-1}\|E\| \geq d_2^{-1}\|E\| \geq \cdots \geq d_n^{-1}\|E\|. \tag{4.5}$$

Their results are more precise than this, and they give conditions, that guarantee that this "typical" behavior actually occurs. Their conditions involve certain Schur complements of E [18, Sections 2 and 3]. Since E is the perturbing matrix the conditions describe the perturbations that result in "typical behavior". This typical behavior occurs provided that certain cancellations do not occur, and hence the name "typical" is appropriate.

Let $\tilde{\theta}_i$ denote the eigenangles of $(A+E, B+F)$, ordered in decreasing distance from $\pi/2$, then the results of Stewart and Zhang imply that for *typical small* perturbations

$$|\tilde{\theta}_j - \pi/2| \text{ is of the order of } d_j\|E\|. \tag{4.6}$$

Following Stewart and Zhang we may call the numbers

$$d_1^{-1}, d_2^{-1}, \cdots, d_n^{-1} \tag{4.7}$$

the "(typical) condition numbers of the multiple eigenangle θ", with the understanding that they represent the typical size of perturbations of the multiple eigenangle θ. Notice that we are not saying that the second furthest eigenangle from θ will differ from θ by at most a moderate multiple of $d_2^{-1}r(E+iF)$, it could well differ by much more, as we will see in Example 4.2.

4.2 Worst case perturbation of a cluster

We begin with a example where the typical behavior does not occur.

Example 4.2 *Take*

$$A = 0, \quad B = \begin{pmatrix} 1 & 0 \\ 0 & 10^6 \end{pmatrix}, \quad E = \begin{pmatrix} 0 & 10^{-6} \\ 10^{-6} & 0 \end{pmatrix}, \quad F = 0.$$

The eigenangles of (A, B) are both $\theta = \pi/2$. The eigenangles of $(A+E, B+F) = (E, B)$ are the arc-cotangents of the eigenvalues of

$$B^{-1/2}EB^{-1/2} = \begin{pmatrix} 0 & 10^{-9} \\ 10^{-9} & 0 \end{pmatrix}$$

that is, $\cot^{-1}(\pm 10^{-9}) \approx \pi/2 \pm 10^{-9}$.

Thus, the two eigenangles of $(A+E, B+F)$ both differ from the multiple eigenangle of (A, B) by about $10^{-9} = \sqrt{1 \cdot 10^{-6}} \times \|E\|$ rather than by the "typical" $1 \cdot \|E\| = 10^{-6}$ and $10^{-6}\|E\| = 10^{-12}$. In particular, the closest eigenangle of $(A+E, B+F)$ to $\pi/2$ is much further than 10^{-12} from $\pi/2$.

Since (A, B) above has repeated eigenangles, this example does not, strictly speaking, illustrate ill-disposition. A slight perturbation of it does. Consider

Example 4.3

$$A = \begin{pmatrix} 0 & 0 \\ 0 & 10^{-12} \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 \\ 0 & 10^{6} \end{pmatrix}, \quad E = \begin{pmatrix} 0 & 10^{-6} \\ 10^{-6} & -10^{-12} \end{pmatrix}, \quad F = 0.$$

So let us now address worst case perturbation bounds. We will consider the case when the pair (A, B) has all its eigenangles in a tight cluster. This includes the case where (A, B) has only a single eigenangle. From the preceding analysis one sees that one needs to bound the absolute values of the eigenvalues of $B^{-1/2}EB^{-1/2}$ when $E = E^*$ and B is diagonal. We give a simple partial answer as to what is the best bound on the kth largest eigenvalue (in absolute value) of $B^{-1/2}EB^{-1/2}$.

Proposition 4.4 *Let $E = E^*$ have norm at most 1. Let $B = \mathrm{diag}(d_1, \ldots, d_n)$, where the d_i are diagonal are positive and in increasing order. Let $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$ denote the ordered absolute values of the eigenvalues of $B^{-1/2}EB^{-1/2}$. (In fact, $|\lambda_k|$ is just the kth singular value of $B^{-1/2}EB^{-1/2}$). Then we have the two bounds*

$$|\lambda_k| \leq \left(\prod_{j=1}^{k} d_j^{-1} \right)^{1/k}, \tag{4.8}$$

and

$$|\lambda_k| \leq \left((n-k+1) d_1^{-1} d_k^{-1} \right)^{1/2}. \tag{4.9}$$

Proof Let $\lambda_j(E)$, $j = 1, \ldots, n$ be the eigenvalues of E, also ordered in decreasing absolute value. In other words, $|\lambda_j(E)|$ is just the jth largest singular value of E, $|\lambda_j|$ is the jth largest singular value of $B^{-1/2}EB^{-1/2}$. The singular values of E are at most $\|E\| = 1$ in absolute value. Thus, the standard product inequality for singular values due to A. Horn (see, e.g., [9, Theorem 3.3.4]) yields

$$|\lambda_k|^k \leq \prod_{j=1}^{k} |\lambda_j| \leq \prod_{j=1}^{k} |\lambda_j(E)| d_j^{-1} \leq \prod_{j=1}^{k} d_j^{-1}$$

hence

$$|\lambda_k| \leq \left(\prod_{j=1}^{k} d_j^{-1} \right)^{1/k},$$

which is (4.8).

Now consider (4.9). The j-th column of $B^{-1/2}EB^{-1/2}$ has norm at most $(d_j^{-1} d_1^{-1})^{1/2}$. Thus for any unit vector x with first $k-1$ entries 0 we must have

$$\|B^{-1/2}EB^{-1/2}x\| \leq \sqrt{n-k+1}(d_1^{-1} d_k^{-1})^{1/2},$$

and the bound (4.9) now follows from min-max. □

Sometimes the bound (4.8) is stronger, while sometimes (4.9) is. When $k = 2$ the two bounds are identical, and the example we presented shows that they are attainable.

The second ingredient of our analysis of the worst case perturbation of a multiple eigenangle is the following simple perturbation bound for Hermitian matrices.

Lemma 4.5 *Let A and E be Hermitian matrices. Let $\eta_i, i = 1, \ldots, n$ be the eigenvalues of E, ordered so that $|\eta_1| \geq |\eta_2| \geq \cdots \geq |\eta_n|$. Then at most $k-1$ eigenvalues of $A + E$ lie outside the interval*

$$[\lambda_{\min}(A) - |\eta_k|, \lambda_{\max}(A) + |\eta_k|].$$

Proof Let $E = \sum_{i=1}^{n} \eta_i x_i x_i^*$ be a spectral decomposition of E. We will decompose E as a sum of two matrices – one with norm $|\eta_k|$ and one with rank at most $k-1$ as follows. For $i < k$, let

$$\tilde{\eta}_i = \operatorname{sign}(\eta_i)|\eta_k|, \quad \text{and} \quad \hat{\eta}_i = \eta_i - \tilde{\eta}_i.$$

The $E = E_1 + E_2$ where

$$E_1 = \sum_{i=1}^{n} \tilde{\eta}_i x_i x_i^*, \quad \text{and,} \quad E_2 = \sum_{i=1}^{k-1} \hat{\eta}_i x_i x_i^*.$$

Since $\|E_1\| = |\eta_k|$, all the eigenvalues of $A + E_1$ lie in

$$[\lambda_{\min}(A) - |\eta_k|, \lambda_{\max}(A) + |\eta_k|], \tag{4.10}$$

and since E_2 has rank at most $k-1$, at most $k-1$ eigenvalues of $(A + E_1) + E_2$ lie outside the interval (4.10). □

We can now give worst case bounds on the perturbation of a cluster of eigenangles.

Theorem 4.6 *Let the definite pair (A, B) have all its eigenangles in the interval*

$$[\pi/2 - \bar{\theta}_1, \ \pi/2 + \bar{\theta}_1].$$

Let X be such that

$$X^*AX = \operatorname{diag}(\alpha_1, \ldots, \alpha_n), \quad \textit{and} \quad X^*BX = \operatorname{diag}(\beta_1, \ldots, \beta_n)$$

and $\beta_1 \leq \cdots \leq \beta_n$. Let (E, F) be a perturbation such that

$$r(E + iF) < c(A, B). \tag{4.11}$$

Then at most $k-1$ eigenangles of $(A + E, B + F)$ lie outside

$$[\pi/2 - \bar{\theta}_2, \ \pi/2 + \bar{\theta}_2],$$

where

$$\bar{\theta}_2 = \arctan\left(\frac{\tan(\theta_1) + \epsilon_k \|X\|^2 \|E\|}{1 - \|X\|^2 \|F\| \beta_1^{-1}}\right)$$

and

$$\epsilon_k = \|X\|^2 \min\left\{ \left(\prod_{j=1}^{k} \beta_j^{-1}\right)^{1/k}, \left((n-k+1)\beta_1^{-1}\beta_k^{-1}\right)^{1/2} \right\}.$$

Proof The eigenangles of $(A+E, B+F)$ are the same as those of $(D_A+\tilde{E}, D_B+\tilde{F})$ where $\tilde{E} = X^*EX$ and $\tilde{F} = X^*FX$. The condition (4.11) ensures that (A,B) is definite and hence so is $(D_A+\tilde{E}, D_B+\tilde{F})$. By Lemma 4.5 at most $k-1$ eigenvalues of the Hermitian matrix $D_B^{-1/2}(D_A+\tilde{E})D_B^{-1/2}$ lie outside the interval

$$[-(\tan(\theta_1)+\epsilon_k\|X\|^2\|E\|),\ (\tan(\theta_1)+\epsilon_k\|X\|^2\|E\|)].$$

This fact together with Ostrowskii's Inequality (4.4) tells us that at most $k-1$ eigenvalues of the Hermitian matrix

$$\begin{aligned}(D_B+\tilde{F})^{-1/2}(D_A+\tilde{E})(D_B+F)^{-1/2} = \\ (I+D_B^{-1/2}\tilde{F}D_B^{-1/2})D_B^{-1/2}(D_A+\tilde{E})D_B^{-1/2}(I+D_B^{-1/2}\tilde{F}D_B^{-1/2})^{-1/2}\end{aligned}$$

lie outside the interval

$$\left[-\frac{\tan(\theta_1)+\epsilon_k\|X\|^2\|E\|}{1-\|\tilde{F}\|\beta_1^{-1}},\ \frac{\tan(\theta_1)+\epsilon_k\|X\|^2\|E\|}{1-\|\tilde{F}\|\beta_1^{-1}}\right].$$

Converting this to a bound on eigenangles and bounding $\|\tilde{F}\|$ by $\|X\|^2\|F\|$ gives the desired result. □

Since the preceding result is a worst case bound we have taken pains to give a precise bound – complete with complications. However, the interpretation of this result is straightforward. Typically we would apply this theorem with $\bar{\theta}_1$ and (E,F) both small so that $\beta_i \approx d_i$, $1-\|X\|^2\|F\|\beta_1^{-1} \approx 1$ and $\tan(\theta_1) \approx \theta_1$. Thus, $\theta_2 \approx \theta_1 + \epsilon_k\|X\|^2\|E\|$, and the worst case perturbation of eigenangles in the cluster is bounded in terms of ϵ_k.

Theorem 4.6, as stated, is very specialized but it can easily be applied to the general case of pair that has a cluster of $k<n$ eigenangles around $\theta \neq \pi/2$ as well as $n-k$ other eigenangles.

Since we can rotate the pair (A,B) we may apply the result for values of θ other than $\theta = \pi/2$. This rotation will result in a change in $\|E\|$ and $\|F\|$, so in this case replace these quantities by $r(E+iF)$, $r(E+iF)$ which are rotation invariant and larger. (See Section 2 for a discussion of the rotation a pair.)

Now suppose that (A,B) has a cluster of eigenangles around θ and also other eigenangles well separated from θ. This case can be reduced to the case of a single cluster by block diagonalizing the pair by congruence. This will increase the norm of the perturbation by a factor of at most 2.

4.3 Ill-disposition

What can we say about *ill-disposition*–the phenomenon where an ill conditioned eigenangle causes other nearby eigenangles to be sensitive to perturbations? For simplicity let us assume that the definite pair has eigenangles $\theta_j \neq \theta_k$ that are close to each other and that these two are simple and are well separated from the other eigenangles. Let us assume that d_j^{-1}, the condition number of θ_j is small, but that d_k^{-1}, the condition number of θ_k is large. How can the ill-conditioning of θ_k affect the sensitivity of θ_j?

1. We have seen in (3.6), that provided $r(E+iF) < r_j$, the sensitivity of θ_j depends solely on d_j^{-1}. One can check that if $\epsilon \geq r_j$ then there is a perturbation $E + iF$ such that $r(E + iF) \leq \epsilon$ and the jth eigenangle is multiple. So, ill-disposition can only occur if the perturbation is large enough to allow θ_j to merge with another eigenangle.

2. However, once we have a multiple eigenangle, we have seen that the quantities d_j^{-1} and d_k^{-1} are the "typical condition numbers", measuring the sensitivity of the multiple eigenangle to "typical" perturbations (4.6). So even in this case we will not typically observe the effects of ill-disposition. Sometimes, as in Example 3.4, it is necessary to reorder the eigenangles in order to have the perturbations of the order of $d_j^{-1} r(E+iF)$.

3. Examples 4.2 and 4.3 show that it is possible to construct a particularly bad perturbation so that both θ_j and θ_k are more sensitive than the condition number d_j^{-1} suggests. However, the bound on ϵ_k in Theorem 4.6 shows that the sensitivity of the less sensitive eigenangle is at most
$$\sqrt{d_j^{-1} d_k^{-1}} \quad \text{not} \quad d_k^{-1}.$$

In short, for small perturbations and "typical" perturbations, d_j^{-1} and d_k^{-1} measure the sensitivity of the eigenangles. Even in the worst case, at least one eigenangle has sensitivity bounded by $\sqrt{d_j^{-1} d_k^{-1}}$. In Stewart's words "it is possible to overemphasize the effects of ill-disposition" [16, p. 685].

5 Quadratic Eigenangle bounds

It is well known that if λ is a simple eigenvalue of a diagonal matrix then off-diagonal perturbations of the matrix cause only quadratically small perturbations in the eigenvalue. This has been quantified in the case of the Hermitian eigenvalue problem [17, 12]. We shall use the basic Schur complement method used in [12]: That is, assuming that X is invertible, the Hermitian matrices
$$\begin{pmatrix} X & Y \\ Y^* & Z \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} X & 0 \\ 0 & Z - Y^* X^{-1} Y \end{pmatrix}$$
are congruent, and hence have the same inertia. To see that they are congruent compute
$$\begin{pmatrix} I & -X^{-1}Y \\ 0 & I \end{pmatrix}^* \begin{pmatrix} X & Y \\ Y^* & Z \end{pmatrix} \begin{pmatrix} I & -X^{-1}Y \\ 0 & I \end{pmatrix} = \begin{pmatrix} X & 0 \\ 0 & Z - Y^* X^{-1} Y \end{pmatrix}. \tag{5.1}$$

Let
$$A = \operatorname{diag}(d_1 \cos(\theta_1), \ldots, d_n \cos(\theta_n)), \quad \text{and} \quad B = \operatorname{diag}(d_1 \sin(\theta_1), \ldots, d_n \sin(\theta_n)).$$

We do not assume that the θ_i are ordered. Let $A = A_{11} \oplus A_{22}$ and $B = B_{11} \oplus B_{22}$ where A_{11} and B_{11} are both $m \times m$. Let

$$E = \begin{pmatrix} 0 & R_A \\ R_A^* & 0 \end{pmatrix} \text{ and } F = \begin{pmatrix} 0 & R_B \\ R_B^* & 0 \end{pmatrix}$$

where R_A and R_B are $m \times (n-m)$. Let us relate the eigenangles of (A, B) and $(A+E, B+F)$.

Let θ be different from θ_i, for $i = 1, \ldots, m$. Let $s = \sin(\theta)$ and $c = \cos(\theta)$. Then, by (5.1), $s(A+E) - c(B+F)$ is congruent to

$$\begin{pmatrix} sA_{11} - cB_{11} & 0 \\ 0 & sA_{22} - cB_{22} - (sR_A + cR_B)^*(sA_{11} - cB_{11})^{-1}(sR_A + cR_B) \end{pmatrix},$$

which we may write as

$$\begin{pmatrix} sA_{11} - cB_{11} & 0 \\ 0 & sA_{22} - cB_{22} - Q_{22}(\theta) \end{pmatrix} = \begin{pmatrix} sA_{11} - cB_{11} & 0 \\ 0 & s(A_{22} - sQ_{22}(\theta)) - c(B_{22} + cQ_{22}(\theta)) \end{pmatrix},$$

where

$$Q_{22}(\theta) = (sR_A + cR_B)^*(sA_{11} - cB_{11})^{-1}(sR_A + cR_B).$$

Set

$$Q(\theta) = \begin{pmatrix} 0 & 0 \\ 0 & Q_{22}(\theta) \end{pmatrix}.$$

and

$$\Delta_1(\theta) = \min\{|d_i \sin(\theta_i - \theta)| : i = 1, \ldots, m.\}.$$

Note that since $Q(\theta)$ is Hermitian

$$r(-\sin(\theta)Q(\theta) + i\cos(\theta)Q(\theta)) = r(e^{-i\theta}Q(\theta)) = \|Q(\theta)\|, \tag{5.2}$$

and that

$$\begin{aligned} \|Q(\theta)\| &= \|Q_{22}(\theta)\| \\ &\leq \Delta_1^{-1}(\theta)\|\sin(\theta)R_A + \cos(\theta)R_B\|^2 \\ &= \Delta_1^{-1}(\theta)\|\sin(\theta)E + \cos(\theta)F\|^2 \\ &\leq \Delta_1^{-1}(\theta)r^2(E + iF). \end{aligned} \tag{5.3}$$

We have used (2.2) for the last inequality.

In short, $\sin(\theta)(A+E) - \cos(\theta)(B+F)$ is congruent to

$$\sin(\theta)[A - \sin(\theta)Q(\theta)] - \cos(\theta)[B + \cos(\theta)Q(\theta)].$$

Thus if θ is an eigenangle of $(A+E, B+F)$ it is also an eigenangle of

$$(\ [A - \sin(\theta)Q(\theta)]\ ,\ [B + \cos(\theta)Q(\theta)]\ ),$$

and so θ differs from an eigenangle of (A, B) by $O(\|Q(\theta)\|)$, which is quadratic in $r(E+iF)$.[8] This basic idea has been exploited in a number of ways in the context of the Hermitian eigenvalue problem in [12]. It was easier to exploit the idea in the context of the Hermitian eigenvalue problem because there every eigenvalue is equally well conditioned (in fact, with condition number 1!), and the condition number bounded the sensitivity of both small and large perturbations. Neither of these features is present in the generalized eigenvalue problem – even if we assume definiteness. It is because of these complications that we state only the simplest possible quadratic perturbation bound and do not pursue the generalizations that were relatively straightforward in [12].

If we assume definiteness then there is a natural ordering of the eigenangles and so we can specify which eigenangle of (A, B) is close to the eigenangle $\tilde{\theta}$ of $(A + E, B + F)$.

Theorem 5.1 *Let*

$$A = \mathrm{diag}(d_1 \cos(\theta_1), \ldots, d_n \cos(\theta_n)), \quad \textit{and} \quad B = \mathrm{diag}(d_1 \sin(\theta_1), \ldots, d_n \sin(\theta_n)).$$

Let

$$E = \begin{pmatrix} 0 & R_A \\ R_A^* & 0 \end{pmatrix} \quad \textit{and} \quad F = \begin{pmatrix} 0 & R_B \\ R_B^* & 0 \end{pmatrix}$$

where R_A and R_B are $m \times (n-m)$. Assume that (A, B) is a definite pair. Fix k and let θ_k (respectively, $\tilde{\theta}_k$) be the kth eigenangle of (A, B), (respectively, $(A + E, B + F)$). Define

$$\Delta_1(\tilde{\theta}_k) \equiv \min\{|d_j \sin(\theta_j - \tilde{\theta}_k)| : j = 1, \ldots, m.\},$$

and

$$d_{\min,2} \equiv \min\{d_j : j = m+1, \ldots, n\}.$$

Assume

$$\Delta_1(\tilde{\theta}_k) > 0, \quad \textit{and} \quad r^2(E + iF)\Delta_1^{-1}(\tilde{\theta}_k) < c(A, B).$$

Then

$$|\theta_k - \tilde{\theta}_k| \leq \sin^{-1}\left(\frac{r^2(E + iF)}{\Delta_1(\tilde{\theta}_k) \cdot d_{\min,2}}\right). \tag{5.4}$$

Proof Let $s = \sin(\tilde{\theta}_k)$ and $c = \cos(\tilde{\theta}_k)$. We have seen above that $s(A + E) - c(B + F)$ is congruent to

$$s[A - sQ(\tilde{\theta}_k)] - c[B + cQ(\tilde{\theta}_k)].$$

Since the kth eigenangle of $(A + E, B + F)$ is $\tilde{\theta}_k$ it follows that the kth eigenvalue of $s(A + E) - c(B + F)$ is 0, and so by congruence, the kth eigenvalue of

$$s(A - sQ(\tilde{\theta}_k)) - c(B + cQ(\tilde{\theta}_k))$$

[8] Notice that so far we have not required that B be positive definite.

is also 0, whence it follows that the kth eigenangle of the pair

$$(A - sQ(\tilde{\theta}_k), B + cQ(\tilde{\theta}_k))$$

is $\tilde{\theta}_k$. Now, since Q is 0 except for the $2,2$ block, it follows from (a very slight modification of) Theorem 3.5 that every eigenangle, and in particular the kth eigenangle, of the pair $(A - sQ(\tilde{\theta}_k), B + cQ(\tilde{\theta}_k))$ is within

$$\sin^{-1}\left(\frac{r(sQ(\tilde{\theta}_k) - icQ(\tilde{\theta}_k))}{d_{\min,2}}\right) = \sin^{-1}\left(\frac{r(Q(\tilde{\theta}_k))}{d_{\min,2}}\right)$$

of the corresponding eigenangle of (A, B). The asserted bound (5.4) follows from this and the bound (5.3) we have on Q. $\square$

We could have permuted the blocks of A and B and so exactly the same result is valid with

$$\Delta_1(\tilde{\theta}_k) \equiv \min\{|d_j \sin(\theta_j - \tilde{\theta}_k)| : j = m+1, \ldots, n.\},$$

and

$$d_{\min,2} \equiv \min\{d_j : j = 1, \ldots, m\}.$$

6 Eigenvector perturbation

In this section we look at eigenvector perturbation, and we do not assume that the eigenangles are ordered. Recall that in (1.5) we defined

$$\delta_j(\theta) \equiv d_j \sin(\theta - \theta_j).$$

We begin with the diagonal case.

Theorem 6.1 *Suppose that*

$$A = \mathrm{diag}(d_1 \cos\theta_1, \ldots, d_n \cos\theta_n) \quad \textit{and} \quad B = \mathrm{diag}(d_1 \sin\theta_1, \ldots, d_n \sin\theta_n),$$

that $(\tilde{A}, \tilde{B}) = (A, B) + (E, F)$ *and that* $\sin(\tilde{\theta})\tilde{A} - \cos(\tilde{\theta})\tilde{B}$ *is singular for some* $\tilde{\theta} \in R$. *Let*

$$\Delta = \min\{|\delta_l(\tilde{\theta})|, l \neq j\}.$$

and assume that

$$r(E + iF) < \Delta. \tag{6.1}$$

Let $w \in \mathbb{C}^{n-1}$ *denote the* j*th column of* $\cos(\tilde{\theta})E + \sin(\tilde{\theta})F$ *with its* j*-th element deleted. Then there is a vector* $v \in \mathbb{C}^n$ *with* $v_j = 1$ *that is an eigenvector (not-necessarily unit) corresponding to* $\tilde{\theta}$ *for which we have the norm bound*

$$\|v - e_j\| \;\leq\; \frac{\Delta^{-1}}{1 - \Delta^{-1} r(E + iF)}\|w\| \;\leq\; \frac{\Delta^{-1}}{1 - \Delta^{-1} r(E + iF)} r(E + iF), \tag{6.2}$$

and, for $k \neq j$, the entry-wise bound

$$\begin{aligned} |v_k| &\leq |\delta_k^{-1}(\tilde{\theta})| \left(|w_k| + \frac{\Delta^{-1} r(E+iF)}{1-\Delta^{-1} r(E+iF)} \|w\| \right) && (6.3) \\ &\leq |\delta_k^{-1}(\tilde{\theta})| \left(\sqrt{|e_{kj}|^2 + |f_{kj}|^2} + \frac{\Delta^{-1} r(E+iF)}{1-\Delta^{-1} r(E+iF)} r(E+iF). \right) && (6.4) \end{aligned}$$

We have included the intermediate bounds in terms of w because in Theorem 6.3 we deduce a bound in the general non-diagonal case from the diagonal case, i.e., Theorem 6.1. By using the intermediate bound rather than the right hand bound in (6.2) the bound in Theorem 6.3 contains $\|X\|$, not $\|X\|^2$.

There are two points to note about this result. First, we do not require that $(A+E, B+F)$ is definite. Second, since $v_j = 1$ we know that $v - e_j$ is orthogonal to the unit vector e_j. Consequently, (6.2) implies the "$\tan\theta$" bound:

$$|\tan\theta(e_j, v)| = \|v - e_j\| \leq \frac{\Delta^{-1}}{1-\Delta^{-1} r(E+iF)} \|w\|. \tag{6.5}$$

Proof We prove the result in the case $j = n$. The general case can be reduced to this by permuting rows and columns. Let $\begin{pmatrix} p \\ 1 \end{pmatrix}$ with $p \in \mathbb{C}^{n-1}$ be a (non-unit) eigenvector associated with the eigenangle $\tilde{\theta}$. We will prove the existence of such an eigenvector at the end of the proof: For the time being let us assume it.

Only in this proof, for any matrix Y, we shall let Y_1 denote the matrix obtained from Y by removing its last row and last column. Set $D_1 = \sin(\tilde{\theta}) A_1 - \cos(\tilde{\theta}) B_1$ (we use D_1 rather than $\tilde{D}_1$ for simplicity of notation). Then, since A and B are diagonal,

$$D_1 = \operatorname{diag}(\delta_1(\tilde{\theta}), \ldots, \delta_{n-1}(\tilde{\theta})), \text{ and } \|D_1^{-1}\| = \Delta^{-1}.$$

Let $R = \sin\tilde{\theta} E_1 - \cos\tilde{\theta} F_1$. Let w be the last column of $\sin\tilde{\theta} E - \cos\tilde{\theta} F$ without the last entry. The first $n-1$ rows of the condition

$$(\sin\tilde{\theta}\tilde{A} - \cos\tilde{\theta}\tilde{B}) \begin{pmatrix} p \\ 1 \end{pmatrix} = 0 \quad \in \mathbb{R}^n$$

are

$$w + D_1 p + Rp = 0 \quad \in \mathbb{R}^{n-1}, \tag{6.6}$$

which yields the exact expression for p:

$$p = D_1^{-1}(I + RD_1^{-1})^{-1} w. \tag{6.7}$$

Now, using (2.2) and (2.4), we have

$$\|R\| = \|\cos(\tilde{\theta}) E_1 + \sin(\tilde{\theta}) F_1\| \leq r(E_1 + iF_1) \leq r(E+iF).$$

Thus taking norms in (6.7) we get the left-hand bound in (6.2). To obtain the right hand bound we need only show $\|w\| \le r(E+iF)$. Let e and f denote the last columns of E and F with their last entries deleted. Using (2.2) and (2.5), we have

$$\begin{aligned}
\|w\| &= \left\| \begin{pmatrix} 0 & w \\ w^* & 0 \end{pmatrix} \right\| \\
&= \left\| \cos(\tilde{\theta}) \begin{pmatrix} 0 & e \\ e^* & 0 \end{pmatrix} + \sin(\tilde{\theta}) \begin{pmatrix} 0 & f \\ f^* & 0 \end{pmatrix} \right\| \\
&\le r\left(\begin{pmatrix} 0 & e \\ e^* & 0 \end{pmatrix} + i \begin{pmatrix} 0 & f \\ f^* & 0 \end{pmatrix} \right) \\
&\le r(E+iF)
\end{aligned}$$

For the entry-wise bound we write (6.7) as

$$p = D_1^{-1}(I + \sum_{m=1}^{\infty} (-RD_1^{-1})^m)w = D_1^{-1}w + D_1^{-1}(\sum_{m=1}^{\infty} (-RD_1^{-1})^m w). \tag{6.8}$$

The kth component of this is vector is at most

$$|\delta_k^{-1}(\tilde{\theta})|\ |w_k| + |\delta_k^{-1}(\tilde{\theta})| \frac{\|RD_1^{-1}\|}{1 - \|RD_1^{-1}\|} \|w\|. \tag{6.9}$$

By Cauchy-Schwarz,

$$|w_k| = |\cos(\tilde{\theta}) e_{kn} + \sin(\tilde{\theta}) f_{kn}| \le \sqrt{|e_{kn}|^2 + |f_{kn}|^2}.$$

The asserted entry-wise bounds (6.3) and (6.4) follow from (6.9) and our bounds on D_1^{-1}, R, $|w_k|$, and $\|w\|$.

Finally, let us prove that there is an eigenvector of the form $(p^T \ \ 1)^T$ corresponding to $\tilde{\theta}$. Let $(p^T \ \ t)^T$ be an eigenvector corresponding to $\tilde{\theta}$. If we can show that t must be non-zero then we may divide the vector by it and get an eigenvector with nth component 1. The proof is by contradiction. If t were 0 then the condition

$$(\sin\tilde{\theta}\tilde{A} - \cos\tilde{\theta}\tilde{B}) \begin{pmatrix} p \\ t \end{pmatrix} = 0$$

would imply

$$D_1 p + Rp = 0 \tag{6.10}$$

that is, that $(D_1 - R)$ is singular. However we know that it is not singular. To see this note that (6.1) gives $\Delta - r(E+iF) > 0$, and hence, $0 \notin W(D_1 - R)$. □

Note that in addition to deriving the rigorous perturbation bounds (6.2) and (6.4) we have an exact expression for the perturbation in (6.7).

When E and F are sufficiently small the bounds (6.2) and (6.4) assume the asymptotic form

$$\|v - (e_j^T v)e_j\| \lesssim \max\{|\delta_k^{-1}(\theta_j)|, k \neq j\} r(E + iF) \tag{6.11}$$

and

$$|v_k| \lesssim |\delta_k^{-1}(\theta_j)| \sqrt{|e_{kj}|^2 + |f_{kj}|}. \tag{6.12}$$

The exact expression (6.7) tells us that the coefficients $\max\{|\delta_k^{-1}(\theta_j)|, k \neq j\}$ and $|\delta_k^{-1}(\theta_j)|$ in (6.11) and (6.12) are the best possible. In particular, if the jth eigenangle is simple then the condition number for jth unit eigenvector is

$$\frac{1}{\min\{|\delta_k(\theta_j)| : k \neq j\}} = \frac{1}{\min\{d_j \sin(|\theta_j - \theta_k|) : k \neq j\}}. \tag{6.13}$$

The condition number of the eigenvectors, even in the diagonal case, in the generalized eigenvalue problem has not been explicitly determined before[9]. The quantity in (6.13) involves both the separation of θ_j from the other eigenangles and the magnitudes of the other normalized generalized eigenvalues, *taken together*. Usually authors have, in effect, *independently* measured the separation of θ_j from the other eigenangles (using some kind of *gap*), and then the minimum of the magnitudes of the other normalized generalized eigenvalues (using the Crawford number, $\min\{d_k : k \neq j\}$, or $\|B^{-1}\|$), and finally combined them to give an upper bounds on the condition number, see for example [14, Theorems 3.7-3.10] [1, p 59, first eigenvector bound]. From the form of (6.13) one can see that it is not possible to compute, or even estimate to within a constant factor, the condition number (6.13) using only the separation and the minimum magnitude.

The quantity $|\delta_k(\theta_j)| = d_k |\sin(\theta_j - \theta_k)|$ is not symmetric in j and k, and so the eigenvectors corresponding to two close eigenangles may, perhaps surprisingly, have different condition numbers. Suppose that

$$\delta_{j_2}(\theta_{j_1}) = \min\{|\delta_k(\theta_{j_1})| : k \neq j_1\},$$

that is, the eigenvalue $(\alpha_{j_2}, \beta_{j_2})$ is the closest eigenvalue to the line in the "direction" θ_1, and conversely, suppose also that,

$$\delta_{j_1}(\theta_{j_2}) = \min\{|\delta_k(\theta_{j_2})|, k \neq j_2\}.$$

Then the unit eigenvectors corresponding to θ_{j_1} and θ_{j_2} have condition numbers

$$|\delta_{j_2}(\theta_{j_1})|^{-1} = (d_{j_2} |\sin(\theta_{j_1} - \theta_{j_2})|)^{-1},$$

and

$$|\delta_{j_1}(\theta_{j_2})|^{-1} = (d_{j_1} |\sin(\theta_{j_1} - \theta_{j_2})|)^{-1}$$

[9]The statement on [14, p. 322] about a condition number is not strictly correct, the quantity Stewart and Sun give is *an upper bound* on the condition number, and we show in Section 7 that it can over estimate the true condition number by an arbitrarily large factor.

respectively. Since d_{j_1} is not necessarily the same as d_{j_2} these two condition numbers are not necessarily the same. Let us assume that $d_{j_1} < d_{j_2}$. The unit eigenvector corresponding to the eigenangle θ_{j_2} is better conditioned than the unit eigenvector corresponding to θ_{j_1}, even though the eigenangle θ_{j_1} is better conditioned that θ_{j_2}. Actually, one should not expect a connection between the conditioning of the eigenangle and the eigenvector since, asymptotically, the eigenangle is changed only by diagonal perturbations, while the eigenvector is changed only by off-diagonal perturbations. Here is an example:

Example 6.2 *Let*

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1000 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 \\ 0 & 1001 \end{pmatrix}, \quad E = 0, \quad F = \begin{pmatrix} 0 & 10^{-4} \\ 10^{-4} & 0 \end{pmatrix}.$$

Then for (A, B) we have

$$\alpha_1 + i\beta_1 = 1 + i, \;\; \alpha_2 + i\beta_2 = 1000 + 1001i, \;\; d_1 = \sqrt{2} = 1.4142, \;\; d_2 = 1414.9.$$

The pair (A, B) is diagonalized by $X = I$ while $(A + E, B + F)$ is diagonalized by

$$\tilde{X} = \begin{pmatrix} 1 & -.099501 \\ 9.9999 \times 10^{-5} & 0.99504 \end{pmatrix}.$$

In this example

$$\begin{aligned} |\delta_2(\theta_1)|^{-1} &= |d_2 \sin(\theta_1 - \theta_2)|^{-1} = 1.4, \\ |\delta_1(\theta_2)|^{-1} &= |d_1 \sin(\theta_1 - \theta_2)|^{-1} = 1.4 \times 10^3. \end{aligned}$$

Thus the first generalized eigenvector is better conditioned than the second. Since the size of the perturbation is $r(E + iF) = 10^{-4}$ one would expect the perturbation in the first normalized generalized eigenvector to be about 1.4×10^{-4}, it is in fact about 10^{-4}, and the perturbation in the second normalized generalized eigenvector to be about 1.4×10^{-1}, it is in fact about 10^{-1}.

Demmel has observed that the condition number for a problem is often proportional to the inverse of the norm of the smallest perturbation that makes the problem ill-posed [4]. Let us consider the condition number κ for the problem of computing the eigenvector corresponding to $\alpha_j + i\beta_j = de^{i\theta}$, still assuming that A and B are diagonal. The problem of computing an eigenvector is ill-posed if the eigenangle is not simple. Our analysis shows that the condition number is

$$(\min\{r(E + iF) : \theta \text{ is a multiple eigenangle of } (A + E, B + F)\})^{-1} \tag{6.14}$$

and not the apparently similar, but potentially much larger quantity

$$(\min\{r(E + iF) : \text{the } j\text{th eigenangle of } (A + E, B + F) \text{ is multiple}\})^{-1}. \tag{6.15}$$

The difference is that in (6.14) the jth eigenangle is held fixed at θ and one must move one of the other eigenangles to θ, while in (6.15) one is free to move the jth eigenangle also. The difference between these two quantities can be large if θ_j is ill-conditioned, but the other nearby eigenangles are not. Demmel's principle, while still applicable, must be carefully interpreted here.

Now consider the general, non-diagonal case. The bounds contain $\|X\|^2$, which we may not know. In this case replace it by the larger quantity n.

Theorem 6.3 *Let (A, B) be a definite pair and suppose that $X \in M_n$ has linearly independent unit columns and*

$$X^*(A + iB)X = \mathrm{diag}(\alpha_1 + i\beta_1, \ldots, \alpha_n + i\beta_n).$$

Suppose that $(\tilde{A}, \tilde{B}) = (A, B)+(E, F)$ and that $\sin(\tilde{\theta})\tilde{A}-\cos(\tilde{\theta})\tilde{B}$ is singular for some $\tilde{\theta} \in R$. Let

$$\Delta = \min\{|\delta_l(\tilde{\theta})| : l \neq j\}$$

and assume that

$$\|X\|^2 r(E + iF) < \Delta.$$

Then there is a vector $v \in \mathbb{C}^n$ that is an eigenvector (not-necessarily unit) corresponding to $\tilde{\theta}$ for which we have the norm bound

$$\|v - X_{j\cdot}\| \leq \frac{\Delta^{-1}}{1 - \Delta^{-1}\|X\|^2 r(E + iF)} \|X\|^2 \, r(E + iF). \tag{6.16}$$

Proof We know from Theorem 6.1 (the left hand bound in (6.2)) that the pair

$$(X^*\tilde{A}X, X^*\tilde{B}X) = (\mathrm{diag}(\alpha_1, \ldots, \alpha_n) + X^*EX, \mathrm{diag}(\beta_1, \ldots, \beta_n) + X^*FX)$$

has an eigenvector u, corresponding to $\tilde{\theta}$, such that

$$\|u - e_j\| \leq \frac{\Delta^{-1}}{1 - \Delta^{-1} r(X^*(E + iF)X)} \|w\| \tag{6.17}$$

where w is the jth column of $X^*(\cos(\tilde{\theta})E + \sin(\tilde{\theta})F)X$ with its jth entry deleted. So,

$$\begin{aligned}
\|w\| &\leq \|[X^*(\cos(\tilde{\theta})E + \sin(\tilde{\theta})F)X]_{j\cdot}\| \\
&\leq \|X^*(\cos(\tilde{\theta})E + \sin(\tilde{\theta})F)\| \, \|X_{j\cdot}\| \\
&= \|X^*(\cos(\tilde{\theta})E + \sin(\tilde{\theta})F)\| \\
&\leq \|X^*\| \|\cos(\tilde{\theta})E + \sin(\tilde{\theta})F\| \\
&\leq \|X\| r(E + iF),
\end{aligned} \tag{6.18}$$

again using (2.2) for the last inequality. Now note that $Xe_j = X_{j\cdot}$, while $v = Xu$ is an eigenvector corresponding to $\tilde{\theta}$. The asserted bound (6.16) follows from (6.17) the bound (6.18) and $\|X(e_j - u)\| \leq \|X\|\|e_j - u\|$. $\square$

Let us derive a "$\sin\theta$" theorem. Let P denote the projection onto $X_{j\cdot}$ and let $y = X_{j\cdot} + Pv$. The vector $X_{j\cdot} - v = X(e_j - u)$ is not necessarily orthogonal to $X_{j\cdot}$ so we cannot deduce a $\tan\theta$ bound from (6.16), but we still have that $X_{j\cdot}$ is a unit vector, and so (6.16) yields the "$\sin\theta$" bound

$$|\sin\theta(X_{j\cdot}, v)| \leq \|v - X_{j\cdot}\| \leq \frac{\Delta^{-1}}{1 - \Delta^{-1}\|X\|^2 r(E + iF)}\|X\|^2 r(E + iF).$$

Let us compare our bound (6.16) with [14, Chapter 6, Theorem 3.8] asserting that

$$\|p\| \leq \frac{\sqrt{\|E\|^2 + \|F\|^2}}{\delta c(\tilde{A}, \tilde{B})}, \tag{6.19}$$

where

$$\delta = \min\{|\sin(\theta_1 - \tilde{\theta}_j)| : 2 \leq j \leq n\} > 0.$$

To first order in $\|E + iF\|$, the right hand side of our bound (6.16) is

$$\frac{1}{\Delta} \cdot \|X\|^2 \cdot r(E + iF), \tag{6.20}$$

while the right hand side of (6.19) is

$$\frac{1}{\delta c(A, B)} \cdot (\|E\|^2 + \|F\|^2)^{1/2}. \tag{6.21}$$

(Note that $r(E + iF) \leq \|E\| + \|F\| \leq \sqrt{2}(\|E\|^2 + \|F\|^2)^{1/2}$.) The major difference between these two quantities (6.20) and 6.21 is that we have Δ instead of $\delta c(A, B)$. This is a potentially huge improvement since

$$\delta c(A, B) \leq \delta d_{\min} \leq \Delta$$

and the ratios

$$\frac{\delta c(A, B)}{\delta d_{\min}} = \frac{c(A, B)}{d_{\min}}, \text{ and } \frac{\delta d_{\min}}{\Delta}$$

can be arbitrarily close to 0 – even in the 2×2 case. We see this for the first ratio in Example 1.1. For the second ratio consider Example 6.4 below. Note that when the perturbation is small, taking $j = 1$ for ease of comparison,

$$\delta \approx \min\{|\sin(\theta_1 - \theta_j)| : 2 \leq j \leq n\}$$

and

$$\Delta \approx \min\{|d_j \sin(\theta_1 - \theta_j)| : 2 \leq j \leq n\}.$$

Example 6.4 *Take $\epsilon > 0$. Let*

$$A = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} \epsilon & 0 \\ 0 & 1 \end{pmatrix}.$$

Then

$$\min\{|\sin(\theta_1 - \theta_j)| : 2 \le j \le n\} = 1/\sqrt{2}$$

and

$$\delta d_{\min} \approx \epsilon/\sqrt{2}.$$

However,

$$\Delta \approx \min\{|d_2 \sin(\theta_1 - \theta_j)| : 2 \le j \le n\} = |d_2 \sin(\theta_1 - \theta_2)| = \sqrt{2}/\sqrt{2} = 1$$

independent of ϵ.

We can also get bounds on $\kappa_{X_{j\cdot}}$, the condition number of the unit eigenvector corresponding to θ_j, in the general non-diagonal case. From (6.16) it follows that

$$\kappa_{X_{j\cdot}} \le \|X\|^2 \Delta^{-1}. \tag{6.22}$$

Notice that $\|X^{-1}\|$ does not appear in these bounds – near collinearity of eigenvectors does not produce ill-conditioned eigenvectors. Indeed, as we saw at the end of Section 2.4 it can improve the conditioning of eigenvectors.

7 Eigenspace perturbation

Now let us consider the problem of bounding the perturbation of an eigenspace. We wish to find P and Q such that

$$\begin{pmatrix} I & P \\ Q^* & I \end{pmatrix} \begin{pmatrix} A_{11} & E_{12} \\ E_{12}^* & A_{22} \end{pmatrix} \begin{pmatrix} I & Q \\ P^* & I \end{pmatrix}$$

and

$$\begin{pmatrix} I & P \\ Q^* & I \end{pmatrix} \begin{pmatrix} B_{11} & F_{12} \\ F_{12}^* & B_{22} \end{pmatrix} \begin{pmatrix} I & Q \\ P^* & I \end{pmatrix}$$

are block diagonal. Since both matrices are by construction Hermitian, it is enough to ensure that their (1,2) blocks are both 0. The resulting equations are

$$A_{11}Q + PA_{22} = -(E_{12} + PE_{12}^*Q) \tag{7.1}$$

$$B_{11}Q + PB_{22} = -(F_{12} + PF_{12}^*Q) \tag{7.2}$$

or equivalently,

$$T(P, Q) = -(E_{12} + PE_{12}^*Q, \;\; F_{12} + PF_{12}^*Q), \tag{7.3}$$

where T is defined by

$$T(P, Q) \equiv (A_{11}Q + PA_{22}, \; B_{11}Q + PB_{22}). \tag{7.4}$$

We shall identify the $n \times 2n$ matrix $[X\ Y]$ with the pair (X, Y). Incidentally, if X and Y are real then $\|[X\ Y]\|_F = \|X + iY\|_F$.

We need to bound the norm of the solution to this system of non-linear equations. Stewart and Sun have done this – see for example [14, Theorem VI.2.13]. Their method was to obtain a bound on the solution to the linearized system, they then use this, and a theorem on the norm of the solution of a non-linear equation in terms of the norm of the solution of the linearized version [14, Theorem V.2.11][10]. Unfortunately, they worked in terms of the norm

$$\|(P,\ Q)\|_{\mathcal{F}} \equiv \max\{\|P\|_F, \|Q\|_F\}.$$

Consequently, they obtain the same bound on P and Q. In the last section we have seen that different eigenvectors can have different condition numbers, so we would like to bound P and Q separately. We follow the general approach in [14] but bound P and Q separately.

7.1 dif$(A_{11}, B_{11}, A_{22}, B_{22})$

Let us consider the linearized version of (7.3): $T(P, Q) = (E_{12}, F_{12})$. It has solution $(P, Q) = T^{-1}(E_{12}, F_{12})$. Stewart and Sun wanted to bound $\|(P,\ Q)\|_{\mathcal{F}} \equiv \max\{\|P\|_F, \|Q\|_F\}$ so they defined

$$\tilde{\text{dif}}(A_{11}, B_{11}, A_{22}, B_{22}) \equiv \inf_{\|(P,Q)\|_{\mathcal{F}} \geq 1} \|T(P, Q)\|_{\mathcal{F}} \tag{7.5}$$

which is with in a factor of $\sqrt{2}$ of the perhaps more natural definition

$$\text{dif}(A_{11}, B_{11}, A_{22}, B_{22}) \equiv \inf_{\|(P,Q)\|_F \geq 1} \|T(P, Q)\|_F = \|T^{-1}\|^{-1}. \tag{7.6}$$

We wish to bound P and Q separately, so we define

$$\text{dif}_P(A_{11}, B_{11}, A_{22}, B_{22}) \equiv \inf_{\|P\|_F \geq 1} \|T(P, Q)\|_F = \|\Pi_P T^{-1}\|^{-1} \tag{7.7}$$

$$\text{dif}_Q(A_{11}, B_{11}, A_{22}, B_{22}) \equiv \inf_{\|Q\|_F \geq 1} \|T(P, Q)\|_F = \|\Pi_Q T^{-1}\|^{-1}, \tag{7.8}$$

where Π_P and Π_Q denote the projections defined on $M_{m,n} \oplus M_{m,n}$ given by $\Pi_P(P, Q) = (P, 0)$ and $\Pi_Q(P, Q) = (0, Q)$. Since $\{(P, Q) : \|P\|_F \geq 1\} \subset \{(P, Q) : \|(P, Q)\|_{\mathcal{F}} \geq 1\}$, and $\{(P, Q) : \|Q\|_F \geq 1\} \subset \{(P, Q) : \|(P, Q)\|_{\mathcal{F}} \geq 1\}$, it follows that

$$\text{dif}_P \geq \text{dif}, \quad \text{and} \quad \text{dif}_Q \geq \text{dif}. \tag{7.9}$$

Also,

$$\text{dif}^{-1} = \|T^{-1}\| = \|\Pi_P T^{-1} + \Pi_Q T^{-1}\| \leq \|\Pi_P T^{-1}\| + \|\Pi_Q T^{-1}\| = \text{dif}_P^{-1} + \text{dif}_Q^{-1}. \tag{7.10}$$

So from (7.9) and (7.10) we have

$$\max\{\text{dif}_P^{-1}, \text{dif}_Q^{-1}\} \leq \text{dif}^{-1} \leq \text{dif}_P^{-1} + \text{dif}_Q^{-1}.$$

[10] This approach was first employed by Stewart in 1971, but few others have employed his powerful method [15, Theorem 3.5].

The quantities dif_P and dif_Q will play an important role in our analysis, so we would like to understand them and be able to compute them. Here is a formula for dif_P and dif_Q in the diagonal case.

Lemma 7.1 *Let*

$$A = \mathrm{diag}(d_1\cos(\theta_1), \ldots, d_n\cos(\theta_n)), \quad \textit{and} \quad B = \mathrm{diag}(d_1\sin(\theta_1), \ldots, d_n\sin(\theta_n)).$$

Let A_{11} and B_{11} be $m \times m$. Let P and Q be $m \times (n-m)$ and such that

$$\begin{aligned} A_{11}Q + PA_{22} &= E_{12} && (7.11)\\ B_{11}Q + PB_{22} &= F_{12} && (7.12) \end{aligned}$$

Suppose that

$$E = [e_{rs}]_{r=1,s=m+1}^{m,n}, \quad F = [f_{rs}]_{r=1,s=m+1}^{m,n}, \quad P = [p_{rs}]_{r=1,s=m+1}^{m,n}, \quad \textit{and} \quad Q = [p_{rs}]_{r=1,s=m+1}^{m,n}.$$

Then, for $m+1 \le i \le n$ and $1 \le j \le m$, we have the entry-wise bounds

$$\begin{aligned} |p_{ij}| &\le \frac{1}{|d_j\sin(\theta_i - \theta_j)|}\sqrt{|e_{ij}|^2 + |f_{ij}|^2} && (7.13)\\ |q_{ij}| &\le \frac{1}{|d_i\sin(\theta_i - \theta_j)|}\sqrt{|e_{ij}|^2 + |f_{ij}|^2}. && (7.14) \end{aligned}$$

Furthermore,

$$\begin{aligned} \mathrm{dif}_P &= \max_{i=m+1,\ldots,n,\ j=1,\ldots,m} |\delta_j(\theta_i)^{-1}| && (7.15)\\ \mathrm{dif}_Q &= \max_{i=m+1,\ldots,n,\ j=1,\ldots,m} |\delta_i(\theta_j)^{-1}|. && (7.16) \end{aligned}$$

Proof Take $1 \le i \le m$ and $m+1 \le j \le n$ and consider the i,j entry of the equations (7.11-7.12). For simplicity of notation let $q = q_{ij}$, $p = p_{ij}$, $e = e_{ij}$ and $f = f_{ij}$. The resulting equations are

$$\begin{aligned} d_i\cos(\theta_i)q + d_j\cos(\theta_j)p &= e && (7.17)\\ d_i\sin(\theta_i)q + d_j\sin(\theta_j)p &= f. && (7.18) \end{aligned}$$

Now write these in matrix-vector form and compute the inverse of the matrix to obtain

$$\begin{pmatrix} p \\ q \end{pmatrix} = \frac{1}{\cos(\theta_j)\sin(\theta_i) - \cos(\theta_i)\sin(\theta_j)} \begin{pmatrix} d_j^{-1} & 0 \\ 0 & d_i^{-1} \end{pmatrix} \begin{pmatrix} \sin(\theta_i) & -\cos(\theta_i) \\ -\sin(\theta_j) & \cos(\theta_j) \end{pmatrix} \begin{pmatrix} e \\ f \end{pmatrix}.$$

Hence we obtain the asserted bounds:

$$|p| \le \frac{1}{|d_j\sin(\theta_i - \theta_j)|}\sqrt{|e|^2 + |f|^2}$$

$$|q| \;\leq\; \frac{1}{|d_i \sin(\theta_i - \theta_j)|}\sqrt{|e|^2 + |f|^2}.$$

The entry-wise bound (7.13) on P implies

$$\mathrm{dif}_P \leq \max_{i=m+1,\ldots,n,\ j=1,\ldots,m} |\delta_j(\theta_i)^{-1}|.$$

We must show that this inequality can be attained. Let i, j be a pair of indices for which the maximum is attained. Set all the entries of E and F to be zero except

$$e_{ij} = \sin(\theta_i), \quad \text{and} \quad f_{ij} = -\cos(\theta_i).$$

Then $\|[E,\ F]\|_F = 1$ and

$$\|P\|_F \geq |p_{ij}| = \frac{1}{|d_j \sin(\theta_i - \theta_j)|} = \max_{i=m+1,\ldots,n,\ j=1,\ldots,m} |\delta_j(\theta_i)^{-1}|.$$

Thus

$$\mathrm{dif}_P \geq \max_{i=m+1,\ldots,n,\ j=1,\ldots,m} |\delta_j(\theta_i)^{-1}|,$$

and so we have (7.15). One can prove (7.16) in exactly the same way. □

This lemma not only gives a formula in the diagonal case. It allows us to interpret dif_P: dif_P^{-1} is the minimum perturbation of a normalized generalized eigenvalue of (A_{11}, B_{11}) that will make its argument the same as that of a normalized generalized eigenvalue of (A_{22}, B_{22}) (mod π).

One would like to extend this result to the case where A_{ii} and B_{ii} are not diagonal by saying that the equations (7.11-7.12) are equivalent to

$$(X_1^* A_{11} X_1)(X_1^{-1} Q X_2) + (X_1^* P X_2^{-*})(X_2^* A_{22} X_2) \;=\; X_1^* E_{12} X_2 \tag{7.19}$$

$$(X_1^* B_{11} X_1)(X_1^{-1} Q X_2) + (X_1^* P X_2^{-*})(X_2^* B_{22} X_2) \;=\; X_1^* F_{12} X_2, \tag{7.20}$$

where X_1 and X_2 are $m \times m$ and $(n-m)\times(n-m)$ and diagonalize (A_{11}, B_{11} and (A_{22}, B_{22})). Then using the diagonal result Lemma 7.1 to bound $X_1^{-1} Q X_2$ and $X_1^* P X_2^{-1}$ and then converting these bounds to bounds on Q and P. Unfortunately, in the process, we will introduce terms $\|X_1^{-1}\|$ and $\|X_2^{-1}\|$. That is, ill conditioning of the eigenvectors will result in poor bounds.

If A and B are not diagonal then we can numerically estimate values μ_P by estimating $\|\Pi_P T^{-1}\|$ since we can solve $T(P, Q) = (E, F)$ using (7.19-7.20). The standard approach would be to use the power method on $(\Pi_P T^{-1})^*(\Pi_P T^{-1}) = T^{-1*}\Pi_P T^{-1}$. A very slight improvement would be to use the Lanczos method [13]. Both these methods require the use of the adjoint of the transformation T. This is easily computed. Taking the inner product on $M_{m,n-m} \oplus M_{m,n-m}$ to be

$$< (X, Y), (U, V) > \;\equiv\; \mathrm{tr}(XU^* + YV^*),$$

we have $\|[X\ Y]\|_F^2 = <(X,Y),(X,Y)>$. Now

$$\begin{aligned} < T(P,Q),\ (R,S) > &= \mathrm{tr}(A_{11}QR^* + PA_{22}R^* + B_{11}QS^* + PB_{22}S^*) \\ &= \mathrm{tr}(QR^*A_{11} + PA_{22}R^* + QS^*B_{11} + PB_{22}S^*) \\ &= \mathrm{tr}(P(A_{22}R^* + B_{22}S^*) + Q(S^*B_{11} + R^*A_{11})) \\ &= \mathrm{tr}(P(RA_{22}^* + SB_{22}^*)^* + Q(B_{11}^*S + A_{11}^*R)^*) \\ &= \mathrm{tr}(P(RA_{22} + SB_{22})^* + Q(B_{11}S + A_{11}R)^*) \\ &= < (P,Q),\ (RA_{22} + SB_{22},\ B_{11}S + A_{11}R) > \end{aligned}$$

using the fact that A_{ii} and B_{ii} are self-adjoint in the penultimate step. Thus the adjoint of T is

$$T^*(R,S) = (RA_{22} + SB_{22},\ B_{11}S + A_{11}R).$$

To compute $T^{-1*}(E,F) = T^{*-1}(E,F)$ we must solve $T^*(R,S) = (E,F)$. This is easy since in solving (7.19-7.20) to compute T^{-1} we have already diagonalized (A_{11}, B_{11}) and (A_{22}, B_{22}).

One can avoid the use of the adjoint by using the Monte Carlo estimation: just compute $\|\Pi_P T^{-1}(E,F)\|_F$ for several randomly chosen pairs (E,F) and take the largest. Another nice aspect of the Monte Carlo approach is that the expensive part of computing $\Pi_P T^{-1}(E,F)$ is computing $T^{-1}(E,F)$. Once you have this, you can compute $\Pi_P T^{-1}(E,F)$ and $\Pi_Q T^{-1}(E,F)$ easily, and hence estimate both dif_P and dif_Q at essentially the same computational cost. Kenney and Laub analyze the probabilistic properties of these Monte Carlo estimates [11].

7.2 Eigenspace perturbation bounds

Now let us bound the solution of the non-linear equations, and thereby derive eigenspace perturbation bounds.

Theorem 7.2 *Suppose that (A_{11}, B_{11}) and (A_{22}, B_{22}) are definite pairs. To simplify notation, set*

$$\mu_P = \mathrm{dif}_P,\ \ \mu_Q = \mathrm{dif}_Q,\ \ \text{and}\ \ \eta = \|[E\ \ F]\|_F.$$

Assume that

$$4\eta^2\mu_P\mu_Q < 1. \tag{7.21}$$

Then $T(P,Q) = (E + PE^*Q,\ F + PF^*Q)$ has a unique solution and this solution satisfies the bounds

$$\|P\|_F \ \leq\ \frac{2\eta\mu_P}{1 + \sqrt{1 - 2\eta^2\mu_P\mu_Q}} \leq 2 \cdot \mathrm{dif}_P \cdot \|[E,F]\|_F, \tag{7.22}$$

$$\|Q\|_F \ \leq\ \frac{2\eta\mu_Q}{1 + \sqrt{1 - 4\eta^2\mu_P\mu_Q}} \leq 2 \cdot \mathrm{dif}_Q \cdot \|[E,F]\|. \tag{7.23}$$

Furthermore, suppose that $T(P_1, Q_1) = (E, F)$, that is, (P_1, Q_1) is the solution to the linearized problem, then

$$\begin{aligned} \|P_1\|_F &\leq \mathrm{dif}_P \cdot \|[E, F]\|_F && (7.24)\\ \|Q_1\|_F &\leq \mathrm{dif}_Q \cdot \|[E, F]\|_F && (7.25)\\ \|P - P_1\|_F &\leq \frac{\zeta}{1-\zeta}\mathrm{dif}_Q \cdot \|[E, F]\|_F && (7.26)\\ \|Q - Q_1\|_F &\leq \frac{\zeta}{1-\zeta}\mathrm{dif}_p \cdot \|[E, F]\|_F && (7.27) \end{aligned}$$

where

$$\zeta \equiv \frac{4\eta^2\mu_P\mu_Q}{1 + \sqrt{1 - 4\eta^2\mu_P\mu_Q}}. \tag{7.28}$$

Proof Recall that by definition, $\mu_P = \mathrm{dif}_Q$ and $\mu_Q = \mathrm{dif}_Q$ are such that if

$$T(U, V) = (X,\ Y)$$

then

$$\|U\|_F \leq \mu_P\|[X\ Y]\|_F \quad \text{and} \quad \|V\|_F \leq \mu_Q\|[X\ Y]\|_F. \tag{7.29}$$

Set $P_0 = 0$ and $Q_0 = 0$, and define sequences $P_1, P_2, \ldots$ and $Q_1, Q_2, \ldots$ by

$$T(P_{k+1}, Q_{k+1}) = -(E + P_kE^*Q_k,\ F + P_kF^*Q_k). \tag{7.30}$$

We will show that these sequences converge to the desired P and Q.

The first step is to show that the sequences are bounded. To this end, we derive bounds on the individual terms in the sequences, using (7.29) for the first inequality:

$$\begin{aligned} \|P_{k+1}\|_F &\leq \mu_P\|[(E + P_k^*E^*Q_k)\ \ (F + P_k^*F^*Q_k)]\|_F\\ &\leq \mu_P(\|[E\ \ F]\|_F + \|[P_k^*E^*Q_k\ \ P_k^*F^*Q_k]\|_F)\\ &\leq \mu_P(\eta + \|P_k\|\|Q_k\|\eta)\\ &\leq \mu_P(\eta + \|P_k\|_F\|Q_k\|_F\eta). && (7.31) \end{aligned}$$

In the same way

$$\|Q_{k+1}\|_F \leq \mu_Q(\eta + \|P_k\|_F\|Q_k\|_F\eta). \tag{7.32}$$

Let $p_0 = q_0 = 0$. Set

$$\begin{aligned} p_{k+1} &= \mu_P(\eta + p_kq_k\eta) && (7.33)\\ q_{k+1} &= \mu_Q(\eta + p_kq_k\eta). && (7.34) \end{aligned}$$

Then

$$\|P_0\|_F \leq p_0 \quad \text{and} \quad \|Q_0\|_F \leq q_0,$$

and so by (7.30) (7.31) and (7.32) and induction

$$\|P_k\|_F \le p_k \quad \text{and} \quad \|Q_k\|_F \le q_k \qquad \text{for } \; k = 1, 2, \ldots .$$

From (7.33-7.34) it follows that $q_k = (\mu_Q/\mu_P)p_k$ thus

$$p_{k+1} = \mu_P \eta + \mu_Q \eta p_k^2, f(p_k).$$

where $f(p) = \mu_P \eta + \mu_Q \eta p^2$. The function f is increasing on $[0, \infty)$ and has a fixed point at

$$p^* \equiv \frac{2\eta\mu_P}{1 + \sqrt{1 - 4\eta^2 \mu_P \mu_Q}} > 0.$$

Apply the increasing function f to $0 \le p_0 \le p^*$ to get

$$0 \le f(0) \le f(p_0) \le f(p^*) = p^*$$

and hence $0 \le p_1 \le p^*$. Repeating this we have $0 \le p_k \le p^*$ for $k = 1, 2, \ldots$. So

$$\|P_k\|_F \le p_k \le p^*.$$

Since $q_k = (\mu_Q/\mu_P)p_k$

$$\|Q_k\|_F \le q_k \le q^* \equiv \frac{2\eta\mu_Q}{1 + \sqrt{1 - 4\eta^2 \mu_P \mu_Q}}.$$

Now we shall show that the sequences $P_0, P_1, \ldots$ and $Q_0, Q_1, \ldots$ are Cauchy. To this end we generate sequences Δ_P^k and Δ_Q^k such that

$$\|P_{k+1} - P_k\|_F \le \Delta_P^k, \quad \text{and} \quad \|Q_{k+1} - Q_k\|_F \le \Delta_Q^k, \quad k = 0, 1, 2, \ldots$$

Since $P_0 = 0$ and $Q_0 = 0$, and

$$T(P_1, Q_1) = -(E, \; F)$$

by (7.29), we may take

$$\Delta_P^0 = \mu_P \eta, \quad \text{and} \quad \Delta_Q^0 = \mu_Q \eta.$$

Since T is linear it follows from the definition of T in (7.4), and that of the sequences $\{P_k\}$, $\{Q_k\}$ in (7.30), that

$$T(P_{k+1} - P_k, Q_{k+1} - Q_k) = -(P_{k-1}EQ_{k-1} - P_kEQ_k, P_{k-1}FQ_{k-1} - P_kFQ_k).$$

Let

$$\hat{Q}_{k-1} = \begin{pmatrix} Q_{k-1} & 0 \\ 0 & Q_{k-1} \end{pmatrix} \text{ and } \hat{Q}_k = \begin{pmatrix} Q_k & 0 \\ 0 & Q_k \end{pmatrix}$$

and note that

$$\|\hat{Q}_{k-1} - \hat{Q}_k\| = \|Q_{k-1} - Q_k\| \le \|Q_{k-1} - Q_k\|_F.$$

Then

$$(P_{k-1}EQ_{k-1} - P_kEQ_k, P_{k-1}FQ_{k-1} - P_kFQ_k) = P_{k-1}[E\ \ F]\hat{Q}_{k-1} - P_k[E\ \ F]\hat{Q}_k.$$

From our assumption (7.29) we have

$$\begin{aligned}
\|P_{k+1} - P_k\|_F &\leq \mu_P\|[(P_{k-1}EQ_{k-1} - P_kEQ_k)\ \ (P_{k-1}FQ_{k-1} - P_kFQ_k)]\|_F \\
&= \mu_P\|P_{k-1}[E\ \ F]\hat{Q}_{k-1} - P_k[E\ \ F]\hat{Q}_k\|_F \\
&= \mu_P\|(P_{k-1}[E\ \ F]\hat{Q}_{k-1} - P_k[E\ \ F]\hat{Q}_{k-1}) + (P_k[E\ \ F]\hat{Q}_{k-1} - P_k[E\ \ F]\hat{Q}_k)\|_F \\
&\leq \mu_P(\|P_{k-1}[E\ \ F]\hat{Q}_{k-1} - P_k[E\ \ F]\hat{Q}_{k-1}\|_F + \|P_k[E\ \ F]\hat{Q}_{k-1} - P_k[E\ \ F]\hat{Q}_k\|_F) \\
&\leq \mu_P(\|P_{k-1} - P_k\|\|Q_{k-1}\|\eta\ +\ \|\hat{Q}_{k-1} - \hat{Q}_k\|\|P_k\|\eta) \\
&\leq \mu_P\eta(\|P_{k-1} - P_k\|q^* + \|\hat{Q}_{k-1} - \hat{Q}_k\|p^*) \\
&\leq \mu_P\eta(\|P_{k-1} - P_k\|_Fq^* + \|Q_{k-1} - Q_k\|_Fp^*).
\end{aligned}$$

So if we set

$$\Delta_P^k = \mu_P(q^*\Delta_P^{k-1} + p^*\Delta_Q^{k-1})\eta \tag{7.35}$$

it is indeed a bound on $\|P_{k+1} - P_k\|_F$. Similarly, we may take

$$\Delta_Q^k = \mu_Q(q^*\Delta_P^{k-1} + p^*\Delta_Q^{k-1})\eta \tag{7.36}$$

as an upper bound on $\|Q_{k+1} - Q_k\|_F$.

It follows from (7.35-7.36) that $\mu_Q\Delta_P^k = \mu_P\Delta_Q^k$ for $k = 0, 1, \ldots$. Also, we have seen that $\mu_Qp^* = \mu_Pq^*$. Substituting these into (7.35), and using the definition of ζ in (7.28) we have

$$\Delta_P^k = 2\mu_Qp^*\eta\Delta_P^{k-1} = \zeta\Delta_P^{k-1} = \cdots = \zeta^k\Delta_P^0 = \zeta^k\mu_P\eta.$$

Thus, since $\zeta \leq \rho < 1$ by (7.21), the sequence $P_0, P_1, \ldots$ is Cauchy, and hence converges. Since $\|P_k\|_F \leq p^*$, it follows that the limit of the sequence $P_0, P_1, \ldots$ also has Frobenius norm at most p^*.

We may apply exactly the same argument to $Q_0,\ Q_1, \ldots$.

The bounds on $\|P - P_1\|_F$ and $\|Q - Q_1\|_F$ are easily verified using the Δ_P^k's and Δ_Q^k's we have found:

$$\|P - P_1\|_F \leq \sum_{k=1}^{\infty}\Delta_P^k = \frac{\zeta\mu_P}{1-\zeta}\eta$$

and similarly for Q. □

One may wonder whether the constants in this theorem are the best possible. We shall show that the constants in (7.22) and (7.23) are within a factor of 2 of the optimal constants. The idea is that the first order bounds (7.24-7.25) are attainable and the constants in these bounds are only a factor of 2 less than in the rigorous bounds (7.22) and (7.23). Here are the details for the bound on P. Since $\mu_P = \mathrm{dif}_P$, (defined in (7.7)) there is a choice of non-zero

(E, F) for which the bound (7.24) is attained. Now let $(P(t), Q(t))$ denote the solution to $T(P, Q) = -(tE + tPE^*Q, tF + tPF^*Q)$ for small $t > 0$. Set $\eta_t = \|(tE, tF)\|_F$. Recall that $(P_1(t), Q_1(t))$ is the solution to the linearized problem. Note that

$$\zeta < 4\eta^2\mu_P\mu_Q \leq (4t\|(E\ F)\|_F\ \mu_P\mu_Q)\ \cdot\ \eta_t$$

so

$$\begin{aligned}\|P(t)\| &\geq \|P_1(t)\|_F\ -\ \|P_1(t) - P(t)\|_F \\ &\geq \mu_P\eta_t\ -\ (4t\|(E\ F)\|_F\mu_P\mu_Q)\ \cdot\ \eta_t \\ &= \{\mu_P - 4t\|(E\ F)\|_F\mu_P\mu_Q\}\ \cdot\ \eta_t.\end{aligned}$$

Now if $\|P(t)\| \leq c\eta_t$ for all $t > 0$ then

$$c \geq \lim_{t\downarrow 0}\frac{P(t)}{\eta_t} = \lim_{t\downarrow 0}\mu_P - 4t\|(E\ F)\|_F\mu_P\mu_Q = \mu_P.$$

Thus, our bound $2\mu_P\eta$ in (7.22) is within a factor of 2 of optimality.

The next lemma is of independent interest. It shows how one can deduce a bound on the angle between two subspaces from a norm bound on the difference between two matrices whose columns span the subspaces.

We let $\mathcal{R}(X) = \mathcal{R}X$ denote the range of the matrix X, and $\Theta(\cdot, \cdot)$ denotes the diagonal matrix of canonical angles between two subspaces. See, for example, [14, Section I.5.2] for a formal definition.

Lemma 7.3 *Let $X, \tilde{X} \in M_{n,m}$. Let $|||\cdot|||$ denote any unitarily invariant norm. Assume that $\|X - \tilde{X}\| < \sigma_m(X)$. Then*

$$|||\ \tan(\Theta(\mathcal{R}(X), \mathcal{R}(\tilde{X})))\ ||| \leq \frac{\sigma_m^{-1}(X)|||X - \tilde{X}|||}{1 - \sigma_m^{-1}(X)\|X - \tilde{X}\|}. \tag{7.37}$$

Proof First assume that X has orthonormal columns, and so $\sigma_m(X) = 1$. Let $Q \in M_n$ be a unitary matrix such that

$$QX = \begin{pmatrix} I_m \\ 0 \end{pmatrix}.$$

Let $Y = Q\tilde{X}$ and partition Y as

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}.$$

Note that

$$\|I - Y_1\| \leq \|X - \tilde{X}\| < 1,$$

and that Y and $Y(I + Y_1)^{-1}$ have the same range space. So

$$\begin{aligned}\Theta(\mathcal{R}(X), \mathcal{R}(\tilde{X})) &= \Theta\left(\mathcal{R}\begin{pmatrix} I \\ 0 \end{pmatrix}, \mathcal{R}\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}\right) \\ &= \Theta\left(\mathcal{R}\begin{pmatrix} I \\ 0 \end{pmatrix}, \mathcal{R}\begin{pmatrix} I \\ Y_2(I - Y_1)^{-1} \end{pmatrix}\right).\end{aligned}$$

Thus the singular values of $Y_2(I-Y_1)^{-1}$ are the tangents of the canonical angles between the subspaces

$$\mathcal{R}\begin{pmatrix} I \\ 0 \end{pmatrix} \text{ and } \mathcal{R}\begin{pmatrix} I \\ Y_2(I-Y_1)^{-1} \end{pmatrix},$$

and so

$$\begin{aligned} |||\tan(\Theta(\mathcal{R}(X),\mathcal{R}(\tilde{X})))||| &= |||Y_2(I+Y_1)^{-1}||| \\ &\le |||Y_2|||\ \|(I+Y_1)^{-1}\| \\ &\le \frac{|||Y_2|||}{1-\|Y_1\|} \\ &\le \frac{|||X-\tilde{X}|||}{1-\|X-\tilde{X}\|}. \end{aligned}$$

Since $\sigma_m(X)=1$, this is (7.37).

If X doesn't have orthonormal columns then write $X=QR$ where $Q\in M_{m,n}$ has orthonormal columns. We may then apply the result to $Q=XR^{-1}$ and $\tilde{X}R^{-1}$, which have the same ranges as X and $\tilde{X}$. $\square$.

Theorem 7.4 *Let (A,B) be a definite pair of $n\times n$ Hermitian matrices, and take $m\in\{1,\dots,n-1\}$. Suppose that $X=(X_1\ X_2)$ has unit columns and that $A_{11},\ B_{11}\in M_m$ are such that*

$$X^*AX=\begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix},\ \textit{and}\ \ X^*BX=\begin{pmatrix} B_{11} & 0 \\ 0 & B_{22} \end{pmatrix}.$$

Let

$$X^*(A+E)X=\begin{pmatrix} \tilde{A}_{11} & E_{12} \\ E_{12}^* & \tilde{A}_{22} \end{pmatrix},\ \textit{and}\ \ X^*(B+F)X=\begin{pmatrix} \tilde{B}_{11} & F_{12} \\ F_{12}^* & \tilde{B}_{22} \end{pmatrix}.$$

Set

$$\mu_P = \mathrm{dif}_P(\tilde{A}_{11},\tilde{B}_{11},\tilde{A}_{22},\tilde{B}_{22}) \tag{7.38}$$

$$\mu_Q = \mathrm{dif}_Q(\tilde{A}_{11},\tilde{B}_{11},\tilde{A}_{22},\tilde{B}_{22}). \tag{7.39}$$

Assume that $4\eta^2\mu_P\mu_Q<1$ where $\eta=\|(E_{12}\ \ F_{12})\|_F$.

Then there is a matrix Y such that $Y=(Y_1\ Y_2)$ and $Y^*(A+E)Y$ and $Y^*(B+F)Y$ are block diagonal where

$$Y_1=X_1+X_2P,\ \ Y_2=X_2+X_1Q$$

and

$$\|P\|_F \le 2\mu_P\|(E_{12}\ \ F_{12})\|_F \tag{7.40}$$

$$\|Q\|_F \le 2\mu_Q\|(E_{12}\ \ F_{12})\|_F. \tag{7.41}$$

Consequently,

$$\|\tan(\Theta(\mathcal{R}(X_1),\mathcal{R}(Y_1)))\|_F \le \frac{2\mu_P\sigma_m^{-1}(X_1)\|X_2\|\|(E_{12}\ F_{12})\|_F}{1-2\mu_P\sigma_m^{-1}(X_1)\|X_2\|\|(E_{12}\ F_{12})\|_F} \tag{7.42}$$

and

$$\|\tan(\Theta(\mathcal{R}(X_2),\mathcal{R}(Y_2)))\|_F \leq \frac{2\mu_P\sigma_{n-m}^{-1}(X_2)\|X_1\|\|(E_{12}\ F_{12})\|_F}{1-2\mu_P\sigma_{n-m}^{-1}(X_2)\|X_1\|\|(E_{12}\ F_{12})\|_F}. \tag{7.43}$$

The quantities $\sigma_m^{-1}(X_1)$ and $\|X_2\|$ occur in our tan Θ bound (7.42). Since X_1 and X_2 are required to have unit columns $\sigma_m^{-1}(X_1)$ and $\|X_2\|$ will be minimized if we take X_1 and X_2 to have orthonormal columns. The resulting bound

$$\|\tan(\Theta(\mathcal{R}(X_1),\mathcal{R}(Y_1)))\|_F \leq \frac{2\mu_P\|(E_{12}\ F_{12})\|_F}{1-2\mu_P\|(E_{12}\ F_{12})\|_F}, \tag{7.44}$$

is much cleaner. We did not explicitly make this choice of X_1 and X_2 in the theorem because the choice of X_1 and X_2 also determines $\tilde{A}_{22}$ and $\tilde{B}_{22}$ which occur in the definition of T and may thus adversely affect the value of μ_P.

7.3 Comparison with results of Stewart and Sun

Now let us compare our bounds with those in [14, Chapter 6, Section 3].

In [14, Theorem VI.3.7] Stewart and Sun give a bound on

$$\max\{\|P\|_F, \|Q\|_F\} \tag{7.45}$$

We have seen that in the definite generalized eigenvalue problem it is possible for one of a pair of complementary eigenspaces to be much better conditioned than the other[11]. A bound on (7.45) cannot detect this while our separate bounds on P and Q in (7.22) and (7.23) can. It is hard to compare our bounds directly with those in [14, Theorem VI.3.7] because of the different notation. When E and F are sufficiently small the bound in [14, Theorem VI.3.7] on P assumes the form

$$\|P\|_F \lesssim \frac{\|X_1\|_F\|X_2\|_F\max\{\|E\|,\|F\|\}}{\delta} \tag{7.46}$$

where $X = (X_1\ X_2)$ is such that

$$X^*AX = \mathrm{diag}(\cos(\theta_1),\ldots,\cos(\theta_n)),\quad X^*BX = \mathrm{diag}(\sin(\theta_1),\ldots,\sin(\theta_n))$$

(note: this choice of X corresponds to the normalization (1.1) and is not the normalization we use in this paper) and

$$\delta = \frac{1}{\sqrt{2}}\min_{i=1,\ldots,m,\ j=m+1,\ldots,n}|\sin(\theta_i-\theta_j)|.$$

The columns of X have lengths $\sqrt{d_i^{-1}}$.

[11] This is in contrast to the symmetric eigenvalue problem where both pairs are have the same condition number because in the symmetric eigenvalue problem complementary eigenspaces are necessarily orthogonal complements.

There are a couple of nice aspects of (7.46) as compared with other results in the literature. Firstly, the Crawford number does not appear–rather we see the d_i^{-1}'s in the form of $\|X_1\|_F$ and $\|X_2\|_F$. Secondly, if it should happen that though $d_{\min}$ is small, the quantities $d_1, \ldots, d_m$ are all large then $\|X_1\|_F$ will be small. That is, this bound can exploit the fact that the normalized generalized eigenvalues are large in one group.

Never-the-less, our bounds in Theorem 7.2 are stronger. As before (7.46) bounds the separation by looking at the angular separation, i.e. $\min |\sin(\theta_i - \theta_j)|$, and the magnitude, i.e. $d_{\min}$, separately, whereas we look at them together in μ_P and μ_Q. In the diagonal case our bound is demonstrably stronger. In the general case one would expect our bound to be stronger. Our bound can be weaker by a factor of at most 2, since it is within a factor of 2 of the optimal bound.

Here is an example to show that even when μ_P and μ_Q are equal our bounds (7.22) and (7.23) can be much stronger than (7.46).

Example 7.5 *Let $t > 0$ and set*

$$A = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & -1 & \\ & & & -1 \end{pmatrix}, \quad B = \begin{pmatrix} 2 & & & \\ & t & & \\ & & 2 & \\ & & & t \end{pmatrix}.$$

The normalized generalized eigenvalues of (A, B) are the pairs (1,2), (1,t), (-1,2), (-1,t). The reader may want to plot these pairs on the plane.

Let us take $m = 2$ and think of t as being large. Then, since in (7.46) the columns of X have norms d_i^{-1} we have

$$\|X_1\|_F = \|X_2\|_F = \sqrt{(1 + 2^2)^{-1} + (1^2 + t^2)^{-1}} \approx 1/\sqrt{5}$$

and

$$\delta \approx 2/t$$

so

$$\frac{\|X_1\|_F \|X_2\|_F}{\delta} \approx \frac{t}{10}.$$

This is the factor that multiplies the norm on the perturbation (E, F) in (7.46). It is linear in t.

In our bounds the error is multiplied by μ_P or μ_Q. Plotting the normalized generalized eigenvalues, and using the fact that $\delta_j(\theta_i)$ is the distance from the jth generalized unit eigenvalue to the line through the origin and the kth generalized eigenangle, one can see that

$$\mu_P = \max_{i=3,4,\ j=1,2} |\delta_j(\theta_i)|^{-1} \geq 1,$$

independent of t.

Now let us compare (7.22-7.23) with [14, Theorem VI.3.10]:

$$\|\sin\Theta(\mathcal{R}(X_1),\mathcal{R}(\tilde{X}_1))\|_F \le \frac{\sqrt{\|A^2+B^2\|}}{c(A,B)}\frac{\sqrt{\|EX_1\|_F^2+\|FX_1\|_F^2}}{\sqrt{2}\;\delta\;c(\tilde{A},\tilde{B})}, \tag{7.47}$$

where the matrices X and $\tilde{X}$ *block* diagonalize (A,B) and $A+E,B+F)$ and $X=(X_1,\ X_2)$ is such that X_1 and X_2 have orthonormal columns, and δ is as before.

There are two essential differences. Firstly, Stewart and Sun's bound contains an additional factor

$$\frac{\sqrt{\|A^2+B^2\|}}{c(A,B)}$$

(which is roughly the ratio of the distance of furthest point from the origin in $W(A+iB)$ to the closest point) where we have the factor n. Secondly, the bound in (7.47) uses $\delta c(\tilde{A},\tilde{B})$ where we use dif_P^{-1} or dif_Q^{-1}. It is not hard to construct examples where the ratio

$$\frac{\delta c(A,B)}{\mathrm{dif}_P^{-1}}$$

is arbitrarily close to 0. Thus (7.22) can be stronger than (7.47). In the diagonal case the formula (7.15) shows that $\delta c(A,B)\le \mathrm{dif}_P$. One might expect that in the general case $\delta c(A,B)\le c(n)\mathrm{dif}_P$ for some moderate function $c(n)$ but we have not been able to prove this.

Stewart and Sun observe that the presence of the two Crawford numbers in the denominator of (7.47) is troubling and ask whether both of them should be there. In (7.42) and (7.44) have replaced the two Crawford numbers with a single μ_P^{-1}. In the diagonal case our bound is stronger.

8 Conclusions

We have proposed the use of normalized generalized eigenvalues in the perturbation theory for the definite generalized eigenvalue problem. The use of normalized generalized eigenvalues, rather than eigenangles (normalization (1.1)) or generalized eigenvalues (normalization (1.2)), preserves the scale of the eigenvalue rather than normalizing the eigenvalue for the purposes of esthetics/uniqueness. The resulting perturbation bounds are generally stronger than the existing bounds. They can be stronger by an arbitrarily large factor, and are never much weaker.

In the case of eigenangles, qualitative forms of our quantitative bounds were given by Stewart in 1972, and are proposed as error estimates in LAPACK [1, eigenangle bound on p. 60]. However in the case of eigenvectors, the observation that

$$|\delta_j(\theta_k)| = d_j|\sin(\theta_j-\theta_k)|$$

is the appropriate gap appears to be completely novel. It provides a great advance over existing analyzes, in that the bounds it yields are much stronger than the bounds in the literature, and in fact they are asymptotically optimal. They allow one to determine the condition number of eigenvectors, and show that complementary eigenspaces can have very different condition numbers.

Our approach has resolved two open question in [14, Chapter VI]. It has also provided some insight as to why ill-disposed eigenangles do not usually compromise the accuracy of other eigenangles. Finally, our approach has provided clean quadratic bounds for off diagonal perturbations.

Our approach has been to first analyze the diagonal case and then extend the results to the general case by introducing a factor of $\|X\|^2$ into the bounds. While this is certainly simple, and gives fairly good bounds here because $\|X\|^2 \leq n$, it is not elegant. A direct analysis of the general case would be ideal.

References

[1] 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*. Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[2] F.F. Bonsall and J. Duncan. *Numerical ranges of operators on normed spaces and of elements of normed algebras*. Cambridge University Press, London-New York, 1971. London Mathematical Society Lecture Note Series, No. 2.

[3] F.F. Bonsall and J. Duncan. *Numerical ranges II*. Cambridge University Press, London-New York, 1973. London Mathematical Society Lecture Note Series, No. 10.

[4] J.W. Demmel. On condition numbers and the distance to the nearest ill-posed problem. *Numerische Math.*, 51:251–289, 1987.

[5] K.R. Gustafson and D.K.M. Rao. *Numerical range: The field of values of linear operators and matrices*. Univeritext, Springer-Verlag, New York, 1997.

[6] P. R. Halmos. *A Hilbert Space Problem Book*. Springer-Verlag, New York, second edition, 1982.

[7] Nicholas J. Higham, Francoise Tisseur, and Paul Van Dooren. Detecting a definite hermitian pair and a hyperbolic or elliptic quadratic eigenvalue problem, and associated nearness problems. Linear Algebra Appl., 2002.

[8] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, New York, 1985.

[9] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, New York, 1991.

[10] B. Istratescu. *Introduction to Linear Operator Theory*. Marcel Dekker, New York, 1981.

[11] C. Kenney and A. J. Laub. Small-sample statistical condition estimates for general matrix functions. *SIAM J. Sci. Comp.*, 15(1):36–61, 1994.

[12] R. Mathias. Quadratic residual bounds for the Hermitian eigenvalue problem. *SIAM J. Matrix Anal. Appl.*, 19(2):541–559, 1998.

[13] B. N. Parlett, H. Simon, and L. M. Springer. On estimating the largest eigenvalue with the Lanczos algorithm. *Math. of Comp.*, 38(157):153–165, 1982.

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

[15] G. W. Stewart. Error boudns for approximate invariant subspaces for closed linear operators. *SIAM J. Num. Anal.*, 8:796–808, 1971.

[16] G. W. Stewart. On the sensitivity of the eigenvalue problem $Ax = \lambda Bx$. *SIAM J. Num. Anal.*, 9:669–686, 1972.

[17] G. W. Stewart. Two simple residual bounds for the eigenvalues of a Hermitian matrix. *SIAM J. Matrix Anal. Appl.*, 12:205–208, 1991.

[18] G. W. Stewart and G. Zhang. Eigenvalues of graded matrices and the condition numbers of a multiple eigenvalue. *Numerische Math.*, 58:703–712, 1991.

[19] M. W. Trosset. private communication. 2000.