ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

MATHICSE
Mathematics Institute of Computational Science and Engineering
School of Basic Sciences - Section of Mathematics

MATHICSE Technical Report
Nr. 09.2013
March 2013

On a perturbation bound for invariant subspaces of matrices

Michael Karow, Daniel Kressner

http://mathicse.epfl.ch

Phone: +41 21 69 37648

Address:
EPFL - SB - MATHICSE (Bâtiment MA)
Station 8 - CH-1015 - Lausanne - Switzerland

Fax: +41 21 69 32545

On a Perturbation Bound for Invariant Subspaces of Matrices

Michael Karow* Daniel Kressner†

March 7, 2013

Abstract

Given a nonsymmetric matrix A, we investigate the effect of perturbations on an invariant subspace of A. The result derived in this paper is less restrictive on the norm of the perturbation and provides a potentially tighter bound compared to Stewart's classical result. Moreover, we provide norm estimates for the remainder terms in well-known perturbation expansions for invariant subspaces, eigenvectors, and eigenvalues.

1 Introduction

A subspace $\mathcal{X} \subset \mathbb{C}^n$ of a matrix $A \in \mathbb{C}^{n\times n}$ is called *invariant* if it satisfies

$$A\mathcal{X} \subset \mathcal{X}. \tag{1.1}$$

In this paper, we reconsider the classical question of estimating the impact of perturbations in A on $\mathcal{X}$.

Suppose that the columns $X \in \mathbb{C}^{n\times k}$ form an orthonormal basis of $\mathcal{X}$. Then (1.1) implies the existence of $A_{11} \in \mathbb{C}^{k\times k}$ such that $AX = XA_{11}$. The eigenvalues of A_{11} are independent of the choice of basis and constitute the spectrum of the restriction of A to $\mathcal{X}$. Extending X to a unitary matrix $[X, X_\perp]$ leads to the block Schur decomposition

$$A\big[X, X_\perp\big] = \big[X, X_\perp\big] \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}. \tag{1.2}$$

Note that this implies $\Lambda(A) = \Lambda(A_{11}) \cup \Lambda(A_{22})$, where $\Lambda(\cdot)$ denotes the spectrum of a matrix. Throughout this paper, we will assume

$$\Lambda(A_{11}) \cap \Lambda(A_{22}) = \emptyset. \tag{1.3}$$

This is a necessary and sufficient condition for the Lipschitz continuity of $\mathcal{X}$ with respect to perturbations in A [7, Thm 15.5.1]. (Note that continuity requires a substantially weaker condition [7,

*Institut für Mathematik, TU Berlin, Germany. karow@math.tu-berlin.de
†ANCHP, MATHICSE, EPF Lausanne, Switzerland, daniel.kressner@epfl.ch

Thm 15.2.1].) Hence, if (1.3) holds, adding a small perturbation $A \mapsto A + E$ implies a change in the invariant subspace that is asymptotically proportional to $\|E\|$.

Various bounds on the change of invariant subspaces under perturbations of A have been derived, notably by Davies and Kahan [5], Stewart [15, 16], Demmel [6], Sun [18], and many others. In the general nonsymmetric case, these bounds are only valid as long as $\|E\|$ remains sufficiently small. A minimal requirement is that the separation condition (1.3) remains valid under perturbations. In the language of pseudospectra, this means that $\|E\|$ should stay below the critical perturbation level ε for which the components of the ε-pseudospectrum containing $\Lambda(A_{11})$ and $\Lambda(A_{22})$ first meet each other [1]. It turns out that existing perturbation results are unnecessarily restrictive; they require $\|E\|$ to stay much below this critical level. The main contribution of this paper consists of a bound that imposes less restrictions on E, see Theorem 3.1 below. To derive this bound, we employ pseudospectral techniques in the analysis of a quadratic matrix equation.

The rest of this paper is organized as follows. In Section 2, we introduce the basic tools for the perturbation analysis of invariant subspaces and recall some existing results. Section 3 contains the statement and proof of our main result, a new perturbation bound for invariant subspaces. In Section 3.2, it is shown that this bound is sharp for a 2×2 example. Section 3.3 provides a comparison to existing perturbation bounds, while Section 3.4 discusses a variation of the main result based on the block diagonalization of A. In Section 4, the latter is used to quantify existence conditions and remainder terms for well-known eigenvalue and eigenvector expansions.

2 Preliminaries and existing results

The goal of this section is to summarize some existing perturbation results for invariant subspaces and introduce notation, needed in the rest of the paper, on the way. Let us first recall some basic tools used in the perturbation analysis.

2.1 Separation between matrices

The condition (1.3) can be quantified by the *separation* between A_{11} and A_{22}. Based on Varah's original definition [21], Demmel [6] has proposed

$$\begin{aligned} \operatorname{sep}_\lambda(A_{11}, A_{22}) := \sup\{\varepsilon > 0 \mid \quad & \Lambda(A_{11} + E_{11}) \cap \Lambda(A_{11} + E_{11}) = \emptyset \\ & \text{for all } E_{11}, E_{22} \text{ with } \max\{\|E_{11}\|_2, \|E_{22}\|_2\} \leq \varepsilon\}. \end{aligned} \tag{2.1}$$

This definition has an important interpretation in terms of ε-pseudospectra, defined as

$$\Lambda_\varepsilon(M) = \big\{z \in \mathbb{C} \mid z \in \Lambda(M + E) \text{ for some } E \in \mathbb{C}^{n\times n} \text{ with } \|E\|_2 \leq \varepsilon\big\}$$

for a matrix $M \in \mathbb{C}^{n\times n}$ [20]. The separation (2.1) gives the smallest value such that $\Lambda_\varepsilon(A_{11})$ and $\Lambda_\varepsilon(A_{22})$ do not intersect for any $\varepsilon < \operatorname{sep}_\lambda(A_{11}, A_{22})$. This interpretation yields the expression

$$\operatorname{sep}_\lambda(A_{11}, A_{22}) = \inf_{\lambda\in\mathbb{C}} \max\big\{\sigma_{\min}(A_{11} - \lambda I), \sigma_{\min}(A_{22} - \lambda I)\big\},$$

where $\sigma_{\min}(\cdot)$ denotes the smallest singular value of a matrix. Based on this expression, an algorithm for computing $\operatorname{sep}_\lambda$ has been developed by Gu and Overton [9].

Stewart [16] has introduced a different notion of separation based on the observation that (1.3) is satisfied if and only if the Sylvester operator

$$\mathbb{T} : \mathbb{C}^{(n-k)\times k} \to \mathbb{C}^{(n-k)\times k}, \qquad \mathbb{T} : Z \mapsto ZA_{11} - A_{22}Z.$$

is nonsingular. We will formulate Stewart's definition in a general norm setting. Let $\|\cdot\|$ denote a consistent family of unitarily invariant norms which dominates the spectral norm $\|\cdot\|_2$. This means that $\|\cdot\|$ is defined for matrices of any size, $\|X\| = \|UXV\|$ for all unitary matrices U, V of compatible size, $\|X\| = \|\begin{bmatrix} X & 0 \end{bmatrix}\| = \left\|\begin{bmatrix} X^\top & 0 \end{bmatrix}^\top\right\|$, and $\|X\|_2 \le \|X\|$. Examples for such families are the Schatten p-norms, in particular the Frobenius norm and the spectral norm. Note that

$$\|PXQ\| \le \|P\|_2 \, \|X\| \, \|Q\|_2 \tag{2.2}$$

for all matrices P, X, Q of compatible size.

The separation of A_{11} and A_{22} with respect to $\|\cdot\|$ is defined as

$$\operatorname{sep}(A_{11}, A_{22}) := \inf_{\|Z\|=1} \|\mathbb{T}(Z)\| = \inf_{\|Z\|=1} \|ZA_{11} - A_{22}Z\|. \tag{2.3}$$

In the case of the Frobenius norm, an efficient algorithm for estimating $\operatorname{sep}(A_{11}, A_{22})$ can be derived from the inverse power method [3]. Moreover, the inequality $\operatorname{sep}(A_{11}, A_{22}) \le 2 \cdot \operatorname{sep}_\lambda(A_{11}, A_{22})$ has been shown for the Frobenius norm in [6]. The following lemma generalizes this result.

Lemma 2.1 *Let* $\operatorname{sep}(A_{11}, A_{22})$ *be defined as in* (2.3)*, where* $\|\cdot\|$ *is a consistent family of unitarily invariant norms dominating the spectral norm. Then* $\operatorname{sep}(A_{11}, A_{22}) \le 2 \cdot \operatorname{sep}_\lambda(A_{11}, A_{22})$.

Proof. Let $s = \operatorname{sep}(A_{11}, A_{22})$ and let $E_{11} \in \mathbb{C}^{k\times k}, E_{22} \in \mathbb{C}^{(n-k)\times(n-k)}$ be such that $\|E_{11}\|_2 < s/2$, $\|E_{22}\|_2 < s/2$. Then for all $Z \in \mathbb{C}^{(n-k)\times k}$,

$$\|Z(A_{11} + E_{11}) - (A_{22} + E_{22})Z\| \ge (s - \|E_{11}\|_2 - \|E_{22}\|_2)\|Z\| > 0.$$

Hence, the perturbed Sylvester operator $Z \mapsto Z(A_{11} + E_{11}) - (A_{22} + E_{22})Z$ is nonsingular, which implies $\Lambda(A_{11} + E_{11}) \cap \Lambda(A_{22} + E_{22}) = \emptyset$. Consequently, $s/2$ is a lower bound for $\operatorname{sep}_\lambda(A_{11}, A_{22})$. □

Examples in [6, 21] show that the quantity $\operatorname{sep}(A_{11}, A_{22})$ can be magnitudes smaller than $\operatorname{sep}_\lambda(A_{11}, A_{22})$.

2.2 Invariant subspaces and a quadratic matrix equation

Let us consider a general matrix $A \in \mathbb{C}^{n\times n}$ and partition

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \qquad A_{11} \in \mathbb{C}^{k\times k}, \qquad A_{22} \in \mathbb{C}^{(n-k)\times(n-k)}.$$

For given $Z \in \mathbb{C}^{k\times(n-k)}$, consider the similarity transformation

$$\begin{bmatrix} I & 0 \\ -Z & I \end{bmatrix} A \begin{bmatrix} I & 0 \\ Z & I \end{bmatrix} = \begin{bmatrix} A_{11} + A_{12}Z & A_{12} \\ A_{21} + A_{22}Z - ZA_{11} - ZA_{12}Z & A_{22} - ZA_{12} \end{bmatrix} \tag{2.4}$$

which becomes block upper triangular if and only if the quadratic matrix equation

$$0 = f(A, Z) := A_{21} + A_{22}Z - ZA_{11} - ZA_{12}Z \tag{2.5}$$

is satisfied. This implies Lemma 2.2 below. Note that a subspace $\mathcal{Y}$ of row vectors is called a *left invariant subspace* if $\mathcal{Y}A \subset \mathcal{Y}$.

Lemma 2.2 *Using the notation introduced above, the following statements are equivalent.*

(i) The columns of $[I \;\; Z^\top]^\top$ *span a (right) invariant subspace of* A *such that*

$$A \begin{bmatrix} I \\ Z \end{bmatrix} = \begin{bmatrix} I \\ Z \end{bmatrix} (A_{11} + A_{12}Z).$$

(ii) The rows of $[-Z \;\; I]$ *span a left invariant subspace of* A *such that*

$$\begin{bmatrix} -Z & I \end{bmatrix} A = (A_{22} - ZA_{12}) \begin{bmatrix} -Z & I \end{bmatrix}.$$

(iii) The quadratic matrix equation (2.5) *is satisfied.*

2.3 An asymptotic result

Perturbation bounds that are asymptotically valid as $\|E\| \to 0$ can be obtained in a relatively straightforward way from truncating perturbation expansions. For invariant subspaces, such expansions have been discussed in [4, 12, 19]. In the following, we will illustrate this approach.

Lemma 2.3 *Given* $A \in \mathbb{C}^{n\times n}$*, suppose that there exists* $Z \in \mathbb{C}^{(n-k)\times k}$ *such that* $f(A, Z) = 0$*, with* f *defined as in* (2.5)*. If* $\Lambda(A_{11} + A_{12}Z) \cap \Lambda(A_{22} - ZA_{12}) = \emptyset$ *then there exist an open neighborhood* $\mathcal{E} \subset \mathbb{C}^{n\times n}$ *of* 0 *and an open neighborhood* $\mathcal{Z} \subset \mathbb{C}^{(n-k)\times k}$ *of* Z *such that for each* $E \in \mathcal{E}$ *the equation* $f(A + E, Z_E) = 0$ *has a unique solution* $Z_E \in \mathcal{Z}$*. Moreover,* Z_E *depends analytically on* E *and admits the first-order expansion*

$$Z_E = Z + \mathbb{T}^{-1}\big(E_{21}\big) + O(\|E\|^2),$$

with the Sylvester operator $\mathbb{T} : \Delta Z \mapsto \Delta Z(A_{11} + A_{12}Z) - (A_{22} - ZA_{12})\Delta Z$*.*

Proof. Clearly, f is an analytic function in the entries of A and Z. The derivative of f with respect to the variable Z equals the Sylvester operator $-\mathbb{T}$. Since $A_{22} - ZA_{12}$ and $A_{11} + A_{12}Z$ have disjoint spectra, the operator $\mathbb{T}$ is invertible. Thus, the lemma follows from the implicit function theorem. □

The way Lemma 2.3 is stated will be convenient for later purposes. However, for the sake of an asymptotic result, we may assume without loss of generality that the unperturbed matrix A is already in block triangular form,

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \qquad A_{11} \in \mathbb{C}^{k\times k}, \qquad A_{22} \in \mathbb{C}^{(n-k)\times(n-k)}. \tag{2.6}$$

By Lemma 2.2, this is equivalent to requiring that $X = \begin{bmatrix} I_k \\ 0 \end{bmatrix}$ spans an invariant subspace of A. Combining the statement of Lemma 2.3 with Lemma 2.2 then yields the following result.

Corollary 2.4 *Let A be in block triangular form (2.6) and assume $\Lambda(A_{11}) \cap \Lambda(A_{22}) = \emptyset$. Then, for every E with $\|E\|$ sufficiently small, there exists an invariant subspace $\mathcal{X}_{\mathcal{E}} = \operatorname{span}\begin{bmatrix} I \\ Z_E \end{bmatrix}$ of $A + E$, such that Z_E admits the first-order expansion*

$$Z_E = \mathbb{T}^{-1}\big(E_{21}\big) + O(\|E\|^2), \qquad \mathbb{T} : Z \mapsto ZA_{11} - A_{22}Z. \tag{2.7}$$

Once we have obtained Z_E, there are different ways of comparing the two invariant subspaces

$$\mathcal{X} = \operatorname{span}\begin{bmatrix} I_k \\ 0 \end{bmatrix}, \qquad \mathcal{X}_E = \operatorname{span}\begin{bmatrix} I_k \\ Z_E \end{bmatrix}$$

of the matrices A and $A + E$, respectively. If $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k$ denote the singular values of Z_E then the ith canonical angle between $\mathcal{X}$ and $\mathcal{X}_E$ is given by $\theta_i(\mathcal{X}, \mathcal{X}_E) = \arctan \sigma_i$. Defining $\Theta(\mathcal{X}, \mathcal{X}_E) := \operatorname{diag}(\theta_1, \ldots, \theta_k)$, it is well known [17, Sec II.4] that $\|\sin(\Theta(\mathcal{X}, \mathcal{X}_E))\|$ generates a metric on the k-dimensional subspaces of $\mathbb{C}^n$ for any unitarily invariant matrix norm $\|\cdot\|$. However, it is sometimes more convenient to simply use

$$\|Z_E\| = \|\tan(\Theta(\mathcal{X}, \mathcal{X}_E))\| \tag{2.8}$$

for measuring the distance, which remains close to $\|\sin(\Theta(\mathcal{X}, \mathcal{X}_E))\|$ as long as $\|Z_E\|$ is small. The first-order result

$$\|\tan(\Theta(\mathcal{X}, \mathcal{X}_E))\| = \|Z_E\| = \frac{\|E_{21}\|}{\operatorname{sep}(A_{11}, A_{22})} + O(\|E\|^2) \tag{2.9}$$

is now readily obtained from Corollary 2.4. This also confirms that $\operatorname{sep}(A_{11}, A_{22})^{-1}$ is the condition number of $\mathcal{X}$ [2].

2.4 Non-asymptotic results

The derivation of non-asymptotic results requires a more careful study of the quadratic matrix equation $f(A + E, Z_E) = 0$, with f as in (2.5) and

$$A + E = \begin{bmatrix} A_{11} + E_{11} & A_{12} + E_{12} \\ E_{21} & A_{22} + E_{22} \end{bmatrix}. \tag{2.10}$$

In particular, A is assumed to be block upper triangular.

Lemma 2.5 ([15, Thm 3.5]) *Let $A + E \in \mathbb{C}^{n\times n}$ be partitioned as in (2.10) and set*

$$\hat{s} := \operatorname{sep}(A_{11} + E_{11}, A_{22} + E_{22}).$$

If $\|E_{21}\|\,\|A_{12} + E_{12}\| < \hat{s}^2/4$ then there exists a solution Z_E of $f(A + E, Z_E) = 0$ satisfying

$$\|Z_E\| \leq \frac{2\|E_{21}\|}{\hat{s} + \sqrt{\hat{s}^2 - 4\|E_{21}\|\,\|A_{12} + E_{12}\|}} < 2\frac{\|E_{21}\|}{\hat{s}}.$$

Lemma 2.5 invokes the quantities $A_{11}+E_{11}$ and $A_{22}+E_{22}$ in the separation, but these matrices are usually not known. There are different ways of reformulating Lemma 2.5 in terms of the original matrix A. Stewart [15, 16] uses the relations

$$\begin{aligned} \|A_{21}+E_{21}\| &\leq \|A_{21}\|+\|E_{21}\| \\ \operatorname{sep}(A_{11}+E_{11},A_{22}+E_{22}) &\geq \operatorname{sep}(A_{11},A_{22})-\|E_{11}\|-\|E_{22}\| \end{aligned}$$

to turn Lemma 2.5 into the following result.

Theorem 2.6 ([15, Thm 4.1]) *Let $A+E\in\mathbb{C}^{n\times n}$ be partitioned as in* (2.10) *and set*

$$s_E := \operatorname{sep}(A_{11},A_{22})-\|E_{11}\|-\|E_{22}\|.$$

If $s_E>0$ and

$$\|E_{21}\|(\|A_{12}\|+\|E_{12}\|) < s_E^2/4 \tag{2.11}$$

then there exists a solution Z_E of $f(A+E,Z_E)=0$ satisfying

$$\|Z_E\| \leq \frac{2\|E_{21}\|}{s_E+\sqrt{s_E^2-4\|E_{21}\|(\|A_{12}\|+\|E_{12}\|)}} < 2\frac{\|E_{21}\|}{s_E}.$$

Somewhat surprisingly, Theorem 2.6 can be derived directly from the Newton-Kantorovich theorem, see Appendix A.

A more refined analysis for the case of the Frobenius norm has been given by Demmel [6], see also [13].

Theorem 2.7 ([13, Thm 1.15]) *Using the notation of Theorem 2.6, assume that*

$$\|E\|_F < \frac{\operatorname{sep}(A_{11},A_{22})}{4\|P\|_2}, \tag{2.12}$$

where P denotes the spectral projector of A belonging to $\Lambda(A_{11})$ and sep *is defined in terms of the Frobenius norm. Then*

$$\|Z_E\|_F < \frac{4\|E\|_F}{\operatorname{sep}(A_{11},A_{22})-4\|P\|_2\|E\|_F}.$$

The quantity $\|P\|_2$ in Theorem 2.7 can be much smaller than $1/\operatorname{sep}(A_{11},A_{22})$. In particular, this is the case when $\|A_{12}\|_F$ is small, as can be seen from the inequality

$$\|P\|_2 \leq \|P\|_F \leq \sqrt{1+\frac{\|A_{12}\|_F^2}{\operatorname{sep}^2(A_{11},A_{22})}}, \tag{2.13}$$

see [6, Lemma 4.4]. Note that the latter inequality can always be attained for some A_{12}. Using (2.13) we can replace the condition (2.12) on the norm of E by the (possibly stronger) condition

$$\|E\|_F < \frac{\operatorname{sep}^2(A_{11},A_{22})}{4\sqrt{\operatorname{sep}^2(A_{11},A_{22})+\|A_{12}\|_F^2}}. \tag{2.14}$$

Comparison It is not straightforward to compare the results of Theorem 2.6 and Theorem 2.7. On the one hand, Theorem 2.6 is valid for any unitarily invariant norm and correctly reproduces the first-order bound $\|Z_E\| \lesssim \|E_{21}\|/\operatorname{sep}(A_{11}, A_{22})$ as $\|E\| \to 0$. On the other hand, for $\|P\|_2 \ll 1/\operatorname{sep}(A_{11}, A_{22})$, the condition (2.11) is significantly more restrictive than (2.12). In turn, the set of admissible perturbations is significantly larger for Theorem 2.7. In the important case of $A_{12} = 0$ (which holds, e.g., for a normal matrix A), we have $\|P\|_2 = 1$ and condition (2.12) becomes $\|E\|_F < \operatorname{sep}(A_{11}, A_{22})/4$. As we will see below, even this condition is not optimal.

2.5 Hermitian matrices

Assuming that A is a Hermitian matrix does not affect the results of Section 2.4 to a large extent, except that $\operatorname{sep}(A_{11}, A_{22})$ reduces to the distance between the sets $\Lambda(A_{11})$ and $\Lambda(A_{22})$. To attain stronger results, one needs to impose additional assumptions on the eigenvalue locations. For example, suppose that all eigenvalues of A_{11} are contained in an interval $[\alpha, \beta]$ (e.g., the convex hull of $\Lambda(A_{11})$) and that $\Lambda(A_{22}) \cap [\alpha, \beta] = \emptyset$. Then the well-known Davies-Kahan $\sin 2\Theta$ theorem [5] states

$$\|\sin(2\Theta(\mathcal{X}, \mathcal{X}_E))\| \leq 2\frac{\|E\|}{\operatorname{sep}(A_{11}, A_{22})}$$

for any unitarily invariant norm $\|\cdot\|$. Most importantly, no assumption on the size of $\|E\|$ is needed. Many extensions and modifications of this result have been proposed and a faithful account of the literature, which would do justice to these important developments, is beyond the scope of this paper.

3 Main result

The following theorem contains the main result of this paper.

Theorem 3.1 *Consider the block triangular matrix*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \qquad A_{11} \in \mathbb{C}^{k\times k}, \qquad A_{22} \in \mathbb{C}^{(n-k)\times(n-k)},$$

and assume that $\Lambda(A_{11}) \cap \Lambda(A_{22}) = \emptyset$. *Let*

$$s := \operatorname{sep}(A_{11}, A_{22}) = \inf_{\|Z\|=1} \|ZA_{11} - A_{22}Z\|,$$

where $\|\cdot\|$ *denotes a consistent family of unitarily invariant norms which dominates the spectral norm* $\|\cdot\|_2$. *For* $\varepsilon \geq 0$ *define* $g(\varepsilon) = \sqrt{\varepsilon(\varepsilon + \|A_{12}\|_2)}$ *and let* $\rho \geq 0$ *be such that* $g(\rho) = \frac{s}{2}$, *i.e.,* $\rho = \frac{1}{2}(\sqrt{s^2 + \|A_{12}\|_2^2} - \|A_{12}\|_2)$. *Finally, let* $\mathcal{B}_\rho := \{\, E \in \mathbb{C}^{n\times n} \mid \|E\| < \rho\}$. *Then the following assertions hold:*

(a) For all $\varepsilon \geq 0$, $\Lambda_\varepsilon(A) \subseteq \Lambda_{g(\varepsilon)}(A_{11}) \cup \Lambda_{g(\varepsilon)}(A_{22})$.

(b) If $\varepsilon < \rho$ *then* $\Lambda_{g(\varepsilon)}(A_{11}) \cap \Lambda_{g(\varepsilon)}(A_{22}) = \emptyset$.

(c) There exists a unique analytic function

$$\mathcal{B}_\rho \ni E \longmapsto Z_E \in \mathbb{C}^{m\times l}$$

with the following properties.

(i) The columns of $[I \;\; Z_E^\top]^\top$ *span a right invariant subspace* $\mathcal{X}_E$ *of* $A+E$.

(ii) The rows of $[-Z_E \;\; I]$ *span a left invariant subspace* $\mathcal{Y}_E$ *of* $A+E$.

(iii) The spectrum of the restriction of $A+E$ *to* $\mathcal{X}_E$ *is contained in the pseudospectrum* $\Lambda_{g(\|E\|_2)}(A_{11})$. *The spectrum of the restriction of* $A+E$ *to* $\mathcal{Y}_E$ *is contained in the pseudospectrum* $\Lambda_{g(\|E\|_2)}(A_{22})$.

(iv) The matrix Z_E *satisfies*

$$\|Z_E\| \le \frac{2\|E\|}{s+\sqrt{s^2-4\|E\|(\|E\|+\|A_{12}\|_2)}} \le \frac{2}{s}\,\|E\| \tag{3.1}$$

as well as

$$\|Z_E - \mathbb{T}^{-1}\big(E_{21}\big)\| \le \frac{6}{s^2}\|E\|^2, \tag{3.2}$$

where $\mathbb{T} : Z \mapsto ZA_{11} - A_{22}Z$.

Remark 3.2 *The inequality (3.2) gives a bound for the remainder* $O(\|E\|^2)$ *in the first-order expansion* (2.7).

Remark 3.3 *The first bound in* (3.1) *looks quite complicated. A slightly weaker but more appealing estimate for* $\|Z_E\|$ *is*

$$\|Z_E\| \le \frac{\|E\|}{s}\left(1+\frac{\|E\|}{\rho}\right), \tag{3.3}$$

which can be derived as follows. Setting $\psi(\epsilon) = 2/\left(s+\sqrt{s^2-4\epsilon(\epsilon+\|A_{12}\|)}\right)$ *the first bound in (3.1) can be written as* $\|Z_E\| \le \psi(\|E\|)\,\|E\|$. *A direct computation yields* $\psi''(\epsilon) \ge 0$. *Thus,* ψ *is a convex function. It follows that* $\psi(\epsilon) \le \psi(0) + (\epsilon/\rho)(\psi(\rho)-\psi(0)) = s^{-1}(1+(\epsilon/\rho))$. *This shows (3.3).*

3.1 Proof of Theorem 3.1

Statement (a) of Theorem 3.1 has been shown by Grammont and Largillier [8, Proposition 3.1], see also [11]. Statement (b) is a consequence of (a) and Lemma 2.1. We are now going to prove statement (c). For notational convenience we write Z instead of Z_E. Then, by applying Lemma 2.2 to $A+E$, each of the conditions (ci) and (cii) of Theorem 3.1 is equivalent to

(α') $f(A+E, Z) = 0$.

Again by Lemma 2.2, condition (ciii) of Theorem 3.1 is equivalent to the statements

(β') $\Lambda(A_{11}+E_{11}+(A_{12}+E_{12})Z) \subseteq \Lambda_{g(\|E\|_2)}(A_{11})$,

(γ') $\Lambda(A_{22}+E_{22}-Z(A_{12}+E_{12}))\subseteq\Lambda_{g(\|E\|_2)}(A_{22})$.

For $\varepsilon\geq 0$ with $\varepsilon(\varepsilon+\|A_{12}\|_2)\leq s^2/4$ (equivalently $\varepsilon\leq\rho$), we define

$$\begin{aligned} r_+(\varepsilon) &= 2\varepsilon/\big(s+\sqrt{s^2-4\varepsilon(\varepsilon+\|A_{12}\|_2)}\big),\\ r_-(\varepsilon) &= 2\varepsilon/\big(s-\sqrt{s^2-4\varepsilon(\varepsilon+\|A_{12}\|_2)}\big). \end{aligned}$$

This allows to rewrite the left inequality in Theorem 3.1 (civ) as

(δ) $\|Z\|\leq r_+(\|E\|)$.

Lemma 3.4 *With the notation introduced above, the following assertions hold.*

(i) $r_+(\varepsilon)$ *and* $r_-(\varepsilon)$ *are the zeros of the quadratic polynomial* $p_\varepsilon(r):=(\varepsilon+\|A_{12}\|_2)r^2-s\,r+\varepsilon$. *We have* $p_\varepsilon(r)<0$ *if and only if* $r_+(\varepsilon)<r<r_-(\varepsilon)$.

(ii) If $0\leq\varepsilon_1<\varepsilon_2\leq\rho$ *then* $r_+(\varepsilon_1)<r_+(\varepsilon_2)\leq r_-(\varepsilon_2)<r_-(\varepsilon_1)$.

(iii) If $\|E\|\leq\rho$ *and* $f(A+E,Z)=0$ *then* $\|Z\|\leq r_+(\|E\|)$ *or* $\|Z\|\geq r_-(\|E\|)$.

Proof. (i) and (ii) can be verified by direct computation.

(iii). Since $\Lambda(A_{11})\cap\Lambda(A_{22})=\emptyset$ the Sylvester operator $\mathbb{T}:Z\mapsto ZA_{11}-A_{22}Z$ is nonsingular, and hence $s=\|\mathbb{T}^{-1}\|^{-1}$. The equation $f(A+E,Z)=0$ can be written as

$$\mathbb{T}(Z)=[-Z\ \ I]\,E\,\begin{bmatrix}I\\Z\end{bmatrix}-ZA_{12}Z$$

or, equivalently,

$$Z=\mathbb{T}^{-1}\left([-Z\ \ I]\,E\,\begin{bmatrix}I\\Z\end{bmatrix}-ZA_{12}Z\right).$$

Using (2.2) and the fact that $\|[-Z\ \ I]\|_2=\|\,[I\ \ Z^\top]^\top\,\|_2=\sqrt{1+\|Z\|_2^2}\leq\sqrt{1+\|Z\|^2}$, we conclude that $\|Z\|\leq\big(\|E\|(1+\|Z\|^2)+\|A_{12}\|_2\,\|Z\|^2\big)/s$, which is equivalent to $0\leq p_{\|E\|}(\|Z\|)$. Thus, the claim follows from (i). □

For $E\in\mathcal{B}_\rho$ let $\mathbb{I}_E$ denote the set of $t\in[0,1]$ such that there exists a matrix Z with the following properties:

(α) $f(A+tE,Z)=0$;

(β) $\Lambda_1(t)\subseteq\Lambda_{g(\|E\|_2)}(A_{11})$, where $\Lambda_1(t):=\Lambda(A_{11}+t(E_{11}+E_{12}Z))$;

(γ) $\Lambda_2(t)\subseteq\Lambda_{g(\|E\|_2)}(A_{22})$, where $\Lambda_2(t):=\Lambda(A_{22}+t(E_{22}-ZE_{21}))$;

(δ) $\|Z\|\leq r_+(\|E\|)$.

Observe that the conditions (α')-(γ') coincide with the conditions (α)-(γ) for $t = 1$. We now show that $1 \in \mathbb{I}_E$ via analytic continuation.

Claim 1. $0 \in \mathbb{I}_E$.

Proof. The conditions (α)-(δ) hold for $t = 0$ and $Z = 0$. □

Claim 2. Suppose $\hat{t} \in \mathbb{I}_E$ and $\hat{t} < 1$. Then there exists $\varepsilon > 0$ such that $[\hat{t}, \hat{t} + \varepsilon) \subset \mathbb{I}_E$.

Proof. Let $\hat{Z}$ be such that $f(A+\hat{t}E, \hat{Z}) = 0$. The pseudospectra $\Lambda_{g(\|E\|_2)}(A_{11})$ and $\Lambda_{g(\|E\|_2)}(A_{22})$ are disjoint by Theorem 3.1 (b). Thus $\Lambda_1(\hat{t})$ and $\Lambda_2(\hat{t})$ are disjoint, too. Hence, Lemma 2.3 applied to $A + \hat{t}E$ implies that there exist an $\varepsilon > 0$ and an analytic function

$$[\hat{t}, \hat{t} + \varepsilon) \ni t \longmapsto Z_t \in \mathbb{C}^{m\times l}$$

such that $f(A + tE, Z_t) = 0$ and $Z_0 = \hat{Z}$. We may assume that $\hat{t} + \varepsilon < 1$. The set $\Lambda_1(t)$ is the spectrum of $A + tE$ restricted to the right invariant subspace $[I \;\; Z_t^\top]^\top$. Thus,

$$\Lambda_1(t) \subset \Lambda(A + tE) \subset \Lambda_{g(\|E\|)}(A_{11}) \cup \sigma_{g(\|E\|)}(A_{22}).$$

However, since the latter pseudospectra are disjoint closed sets it follows from $\Lambda_1(\hat{t}) \subseteq \Lambda_{g(\|E\|)}(A_{11})$ and the continuity of eigenvalues that $\Lambda_1(t) \subseteq \Lambda_{g(\|E\|)}(A_{11})$ for all $t \in [\hat{t}, \hat{t} + \varepsilon)$. Analogously, we conclude that $\Lambda_2(t) \subseteq \Lambda_{g(\|E\|)}(A_{22})$ for all $t \in [\hat{t}, \hat{t}+\varepsilon)$. It remains to verify property (δ). By Lemma 3.4 (iii), we have for every t that $\|Z_t\| \le r_+(t\|E\|) \le r_+(\|E\|)$ or $\|Z_t\| \ge r_-(t\|E\|) \ge r_-(\|E\|)$. Since the former inequality holds for $t = \hat{t}$ and $r_+(\|E\|) < r_-(\|E\|)$, the continuity of $t \mapsto Z_t$ implies that $\|Z_t\| \le r_+(\|E\|)$ for all $t \in [\hat{t}, \hat{t} + \varepsilon)$. □

Claim 3. The set $\mathbb{I}_E$ is closed.

Proof. Let (t_j) be a sequence in $\mathbb{I}_E$ with limit t_*. Then there exists a sequence (Z_j) such that the pairs (t_j, Z_j) satisfy (α)-(δ). In particular $\|Z_j\| \le r_+(\|E\|)$ for all j. By compactness the sequence (Z_j) has a convergent subsequence. Let Z_* denote its limit. Then (t_*, Z_*) satisfies (α)-(δ). In particular, the properties (β) and (γ) for (t_*, Z_*) are consequences of the continuity of eigenvalues and the closedness of the pseudospectra $\Lambda_{g(\|E\|_2)}(A_{11}), \Lambda_{g(\|E\|_2)}(A_{22})$. □

From Claims 1–3 it follows that $1 = \sup \mathbb{I}_E \in \mathbb{I}_E$. Thus, we have shown the existence of Z satisfying $(\alpha')-(\gamma')$. Next we prove uniqueness. For this purpose, suppose that $f(A+E, Z) = f(A+E, \tilde{Z}) = 0$. Then

$$\begin{aligned}
0 &= f(A + E, Z) - f(A + E, \tilde{Z}) \\
&= E_{21} + (A_{22} + E_{22})Z - Z(A_{11} + E_{11}) - Z(A_{12} + E_{12})Z \\
&\quad -\big[E_{21} + (A_{22} + E_{22})\tilde{Z} - \tilde{Z}(A_{11} + E_{11}) - \tilde{Z}(A_{12} + E_{12})\tilde{Z}\big] \\
&= \underbrace{(A_{22} + E_{22} - \tilde{Z}(A_{12} + E_{12}))}_{=:\tilde{A}_{22}}(Z - \tilde{Z}) - (Z - \tilde{Z})\underbrace{(A_{11} + E_{11} + (A_{12} + E_{12})Z)}_{=:\tilde{A}_{11}}.
\end{aligned}$$

Since we also assume that Z and $\tilde{Z}$ satisfy (β') and (γ'), the spectra $\Lambda(M_E)$ and $\Lambda(L_E)$ are disjoint. Thus, $Z - \tilde{Z} = 0$.

In summary, we have shown the existence and uniqueness of a function $E \mapsto Z_E$ satisfying the conditions (i)–(iii) in Theorem 3.1. The analyticity of this function is a consequence of Lemma 2.3. It remains to prove the inequality (3.2). The relation $f(A+E, Z_E) = 0$ is equivalent to

$$Z_E = \mathbb{T}^{-1}(E_{21} + E_{22}Z_E - Z_E E_{11} - Z_E E_{12} Z_E).$$

Furthermore, by (3.1) we have $\|Z_E\| \leq 2\|E\|/s < 2\rho/s \leq 1$. Thus,

$$\begin{aligned} \|Z_E - \mathbb{T}^{-1}\big(E_{21}\big)\| &= \|\mathbb{T}^{-1}(E_{22}Z_E - Z_E E_{11} - Z_E E_{12} Z_E)\| \\ &\leq \frac{1}{s}(2\|E\|\,\|Z_E\| + \|E\|\,\|Z_E\|^2) \\ &\leq \frac{3}{s}\|E\|\,\|Z_E\| \leq \frac{6}{s^2}\|E\|^2. \end{aligned}$$

This concludes the proof of Theorem 3.1. □

3.2 The 2×2 case

In this section we show via a 2×2 example that the bounds in Theorem 3.1 are sharp. Let

$$A = \begin{bmatrix} -s/2 & c \\ 0 & s/2 \end{bmatrix}, \qquad E = \begin{bmatrix} 0 & \varepsilon \\ -\varepsilon & 0 \end{bmatrix}, \qquad s > 0, \quad c, \varepsilon \geq 0.$$

Then s is the separation of the diagonal elements of A and $\|E\|_2 = \varepsilon$. Furthermore, let $\rho = \frac{1}{2}(\sqrt{s^2 + c^2} - c)$. Then $\rho(\rho + c) = s^2/4$. The ε-pseudospectrum of A can be calculated as

$$\begin{aligned} \Lambda_\varepsilon(A) &= \{z \in \mathbb{C} \mid \sigma_{\min}(zI - A_0) \leq \varepsilon\} \\ &= \left\{ z \in \mathbb{C} \mid \frac{1}{2}\left(\sqrt{\big(|z - \tfrac{s}{2}| + |z + \tfrac{s}{2}|\big)^2 + c^2} - \sqrt{\big(|z - \tfrac{s}{2}| - |z + \tfrac{s}{2}|\big)^2 + c^2} \right) \leq \varepsilon \right\}, \end{aligned}$$

see also [11]. By Theorem 3.1 (a),

$$\Lambda_\varepsilon(A) \subseteq \mathcal{D}_{\sqrt{\varepsilon(\varepsilon+c)}}(-s/2) \cup \mathcal{D}_{\sqrt{\varepsilon(\varepsilon+c)}}(s/2), \tag{3.4}$$

where $\mathcal{D}_r(z) \subset \mathbb{C}$ denotes the closed disk of radius $r \geq 0$ around $z \in \mathbb{C}$. If $\varepsilon < \rho$ then $\varepsilon(\varepsilon + c) < s/2$ and the disks in (3.4) are disjoint. Hence, the pseudospectrum $\Lambda_\varepsilon(A)$ has two connected components. For $\varepsilon = \rho$ the disks in (3.4) touch each other at $0 \in \mathbb{C}$. From (3.4) it follows that also $0 \in \sigma_\rho(A)$. Hence, $\sigma_\varepsilon(A)$ has only one connected component for $\varepsilon \geq \rho$. The eigenvalues of $A + E$ are

$$\lambda_\pm(\varepsilon) = \pm\frac{1}{2}\sqrt{s^2 - 4\varepsilon(\varepsilon + c)}.$$

These eigenvalues lie close to the boundary of $\Lambda_\varepsilon(A)$. The situation is illustrated in Figure 1, where the shaded regions represent pseudospectra, the circles represent the boundaries of the disks $\mathcal{D}_{\sqrt{\varepsilon(\varepsilon+c)}}(\pm s/2)$, and the dots mark the eigenvalues $\lambda_\pm(\varepsilon)$. A right eigenvector to the eigenvalue $\lambda_-(\varepsilon)$ is given by $[1 \;\; z_\varepsilon]^\top$, where $z_\varepsilon = 2\varepsilon/(s + \sqrt{s^2 - 4\varepsilon(\varepsilon + c)})$. If $\varepsilon < \rho$ then the eigenvalues are

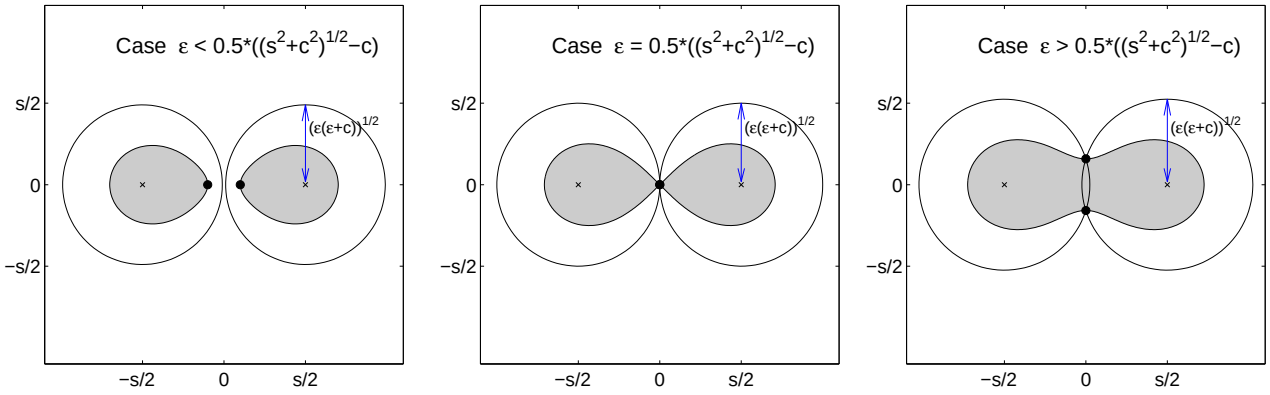

Figure 1: ε-pseudospectra of a 2×2 matrix for three values of ε

real and distinct. If $\varepsilon = \rho$ then $A + E$ is similar to a Jordan block. More specifically, in this case we have

$$A + E = \begin{bmatrix} 1 & -2/s \\ 2\varepsilon/s & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & -2/s \\ 2\varepsilon/s & 0 \end{bmatrix}^{-1}.$$

If $\varepsilon > \rho$ then the eigenvalues $\lambda_\pm(\varepsilon)$ are purely imaginary. Note that the function $\varepsilon \mapsto z_\varepsilon$ is not differentiable at $\varepsilon = \rho$.

3.3 Comparison with existing results

Let us first compare Theorem 3.1 with Stewart's result: Theorem 2.6 requires $s_E = s - \|E_{11}\| - \|E_{22}\| > 0$ and $\|E_{21}\| (\|A_{12}\| + \|E_{12}\|) < s_E^2/4$. Especially if the perturbations are due to roundoff error, it is unlikely that the individual norms of E_{ij} can be controlled or are even known. Therefore, the stronger requirement

$$\|E\| (\|A_{12}\| + \|E\|) < (s - 2\|E\|)^2/4$$

appears to be more realistic. In contrast, Theorem 3.1 requires $\|E\| (\|A_{12}\| + \|E\|) < s^2/4$, which is significantly less restrictive for small s. Theorem 2.6 concludes the bound

$$\|Z_E\| \leq \frac{2\|E_{21}\|}{s_E + \sqrt{s_E^2 - 4\|E_{21}\| (\|A_{12}\| + \|E_{12}\|)}} \leq \frac{2\|E\|}{s_E + \sqrt{s_E^2 - 4\|E\| (\|A_{12}\| + \|E\|)}}. \tag{3.5}$$

The second bound is nearly identical with our bound (3.1), except that we use s instead of the potentially much smaller s_E. The following example shows that this difference can have a large impact on the bound.

Example 3.5 *Consider*

$$A = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline 0 & A_{22} \end{array}\right], \quad E = \frac{st}{2}\left[\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right],$$

where $0 \le t < 1$ *and* $s = \operatorname{sep}(A_{11}, A_{22})$ *with respect to the matrix* 2*-norm. Then* $\|E\| = \|E_{11}\| = \|E_{22}\| = \|E_{21}\| = st/2$ *and* $s_E = s(1-t)$*. As* t *tends to 1 both bounds in (3.5) tend to infinity while our bound (3.1) tends to 1.*

However, it should be noted that our bound, in contrast to the first bound in (3.5), does not reflect the fact that $Z_E = 0$ if $E_{21} = 0$.

The restriction (2.14) on the perturbation E imposed by Demmel's result, Theorem 2.7, is nearly identical with the restriction imposed by Theorem 3.1. However, this assumes that $\|P\|_2$ is not much smaller than predicted by the upper bound (2.13). Otherwise, the condition (2.12) possibly allows for a significantly larger set of perturbation. There is an important exception to these statements for $A_{12} = 0$, e.g., when A is normal. In this case, Theorem 2.7 requires $\|E\|_F < \operatorname{sep}(A_{11}, A_{22})/4$ while our result only requires $\|E\|_F < \operatorname{sep}(A_{11}, A_{22})/2$. Also, the bound on $\|Z_E\|_F$ stated in Theorem 2.7 is by a factor of at least 2 larger than the bound of Theorem 3.1. Hence, Demmel's result is not optimal.

3.4 A bound in terms of the spectral decomposition

The purpose of this section is to derive a bound based on the block diagonalization of A. We assume the setting of Theorem 3.1. In particular, A is block triangular and $\Lambda(A_{11}) \cap \Lambda(A_{22}) = \emptyset$. Then there exists a unique $R \in \mathbb{C}^{k\times(n-k)}$ such that $RA_{22} - A_{11}R = A_{12}$. We have $A[R^\top \;\; I_{n-k}]^\top = [R^\top \;\; I_{n-k}]^\top A_{22}$. Hence, the columns of $[R^\top \;\; I_{n-k}]^\top$ span a right invariant subspace $\mathcal{X}^c$ of A which is complementary to $\mathcal{X} = \operatorname{range}([I_k \;\; 0]^\top)$. The projector onto $\mathcal{X}$ along $\mathcal{X}^c$ is the spectral projector $P = \begin{bmatrix} I_k & -R \\ 0 & 0 \end{bmatrix}$. Let

$$p = \sqrt{1 + \|R\|_2^2}, \qquad \kappa = p + \|R\|_2 = p + \sqrt{p^2 - 1}, \qquad G = \begin{bmatrix} I_k & R/p \\ 0 & I_{n-k}/p \end{bmatrix}. \tag{3.6}$$

Then $p = \|P\|_2$, $\kappa = \|G\|_2\|G^{-1}\|_2$ is the condition number of G [6] and

$$A = G \operatorname{diag}(A_{11}, A_{22})\, G^{-1}.$$

Note that

$$G^{-1} = \begin{bmatrix} I_k & -R \\ 0 & p\, I_{n-k} \end{bmatrix}.$$

With these preparations we are in a position to state and prove the theorem below.

Theorem 3.6 *Let A and s be defined as in Theorem 3.1, and let $\|\cdot\|$ be a family of unitarily invariant norms which dominates the spectral norm. Let R and κ be defined as above. If $\|E\| < s/(2\kappa)$ then there exists a unique $W_E \in \mathbb{C}^{(n-k)\times k}$, depending analytically on E with the following properties.*

(i) The columns of $\begin{bmatrix} I_k & R \\ 0 & I_{n-k} \end{bmatrix} \begin{bmatrix} I_k \\ W_E \end{bmatrix}$ span a right invariant subspace $\mathcal{X}_E$ of $A+E$.

(ii) The rows of $\begin{bmatrix} -W_E & I_{n-k} \end{bmatrix} \begin{bmatrix} I_k & -R \\ 0 & I_{n-k} \end{bmatrix}$ span a left invariant subspace $\mathcal{Y}_E$ of $A+E$.

(iii) The spectrum of the restriction of $A+E$ to $\mathcal{X}_E$ is contained in the pseudospectrum $\Lambda_{\kappa\,\|E\|_2}(A_{11})$. The spectrum of the restriction of $A+E$ to $\mathcal{Y}_E$ is contained in the pseudospectrum $\Lambda_{\kappa\,\|E\|_2}(A_{22})$.

(iv) The matrix W_E satisfies

$$\|W_E\| \le \frac{2\kappa}{s\,p}\,\|E\| \tag{3.7}$$

and

$$\|W_E - \mathbb{T}^{-1}(E_{21})\| \le \frac{6\,\kappa^2}{p\,s^2}\|E\|^2, \tag{3.8}$$

where $\mathbb{T} : Z \mapsto ZA_{11} - A_{22}Z$.

Proof. Let $\hat{A} = \operatorname{diag}(A_{11}, A_{22})$ and $\hat{E} = G^{-1}EG$. Then $A + E = G(\hat{A} + \hat{E})G^{-1}$. Furthermore, the equivalences

$$\begin{aligned} (\hat{A}+\hat{E})U = UL \quad &\Leftrightarrow \quad (A+E)(GU) = (GU)L, \\ V(\hat{A}+\hat{E}) = MV \quad &\Leftrightarrow \quad (VG^{-1})(A+E) = M(VG^{-1}) \end{aligned} \tag{3.9}$$

hold for any $U \in \mathbb{C}^{n\times k}$, $L \in \mathbb{C}^{k\times k}$, $V \in \mathbb{C}^{(n-k)\times n}$, $M \in \mathbb{C}^{(n-k)\times(n-k)}$. Thus, the columns of GU span a right invariant subspace of $A+E$ if and only if the columns of U span a right invariant subspace $\hat{A}+\hat{E}$. Furthermore, the rows of VG^{-1} span a left invariant subspace of $A+E$ if and only if the rows of V span a left invariant subspace $\hat{A}+\hat{E}$. Suppose $\|E\| < s/(2\kappa)$. Then

$$\|\hat{E}\| \le \kappa\|E\| < s/2. \tag{3.10}$$

Hence, according to Theorem 3.1 there exists a unique $Z_{\hat{E}} \in \mathbb{C}^{(n-k)\times k}$ depending analytically on $\hat{E}$ with the following properties.

(i') The columns of $[I_k \;\; Z_{\hat{E}}^\top]^\top$ span a right invariant subspace $\mathcal{X}_{\hat{E}}$ of $\hat{A}+\hat{E}$.

(ii') The rows of $[-Z_{\hat{E}} \;\; I_{n-k}]$ span a left invariant subspace $\mathcal{Y}_{\hat{E}}$ of $\hat{A}+\hat{E}$.

(iii') The spectrum of the restriction of $\hat{A}+\hat{E}$ to $\mathcal{X}_{\hat{E}}$ is contained in the pseudospectrum $\Lambda_{\|\hat{E}\|_2}(A_{11})$. The spectrum of the restriction of $A+E$ to $\mathcal{Y}_{\hat{E}}$ is contained in the pseudospectrum $\Lambda_{\|\hat{E}\|_2}(A_{22})$.

(iv') The matrix $Z_{\hat{E}}$ satisfies

$$\|Z_{\hat{E}}\| \le \frac{2}{s}\,\|\hat{E}\| \qquad \text{as well as} \qquad \|Z_{\hat{E}} - \mathbb{T}^{-1}(\hat{E}_{21})\| \le \frac{6}{s^2}\|\hat{E}\|^2.$$

Let $U = [I_k \;\; Z_{\hat{E}}^\top]^\top$, $V = [-Z_{\hat{E}} \;\; I_{n-k}]$ and $W_E = Z_{\hat{E}}/p$. Then

$$\begin{bmatrix} I_k & R \\ 0 & I_{n-k} \end{bmatrix} \begin{bmatrix} I_k \\ W_E \end{bmatrix} = GU, \qquad \begin{bmatrix} -W_E & I_{n-k} \end{bmatrix} \begin{bmatrix} I_k & -R \\ 0 & I_{n-k} \end{bmatrix} = VG^{-1}.$$

Hence, the claims (i)-(iv) follow from (i')-(iv'), the equivalences (3.9), the inequality (3.10) and the fact that $\hat{E}_{12} = p\, E_{12}$. □

4 The case of a simple eigenvalue

The theorem below gives the second order expansion of a simple eigenvalue as well as the first order expansion of the associated eigenvector. These expansions are well known and can be found, e.g., in [14]. The novel contribution is the existence regions and the bounds for the remainders in the expansion. Below, $M^\sharp$ denotes the group inverse of matrix $M \in \mathbb{C}^{n\times n}$, which equals the Drazin inverse if the zero eigenvalue of M is semisimple.

Theorem 4.1 *Let $x_0 \in \mathbb{C}^n$ be a normalized right eigenvector of $A \in \mathbb{C}^{n\times n}$ belonging to a simple eigenvalue $\lambda_0 \in \mathbb{C}$ (i.e $Ax_0 = \lambda_0 x_0$, $\|x_0\|_2 = 1$). Let $y_0 \in \mathbb{C}$ be a left eigenvector such that $y_0^* A = \lambda_0 y_0^*$ and $y_0^* x_0 = 1$. Let*

$$s_0 = \|(I - x_0 x_0^*)(A - \lambda_0 I)^\sharp\|_2^{-1}, \qquad \kappa_0 = \|y_0\|_2 + \sqrt{\|y_0\|_2^2 - 1}.$$

If $\|E\|_2 < s_0/(2\kappa_0)$ then there exists an eigenvector x_E of $A + E$ depending analytically on E such that $y_0^* x_E = 1$ and

$$x_E = x_0 + (A - \lambda_0 I)^\sharp E x_0 + \xi_E \tag{4.1}$$

for some $\xi_E \in \mathbb{C}^n$ with

$$\|\xi_E\|_2 \le \frac{6\kappa_0^2}{s_0^2}\|E\|_2^2.$$

The corresponding eigenvalue of $A + E$ satisfies

$$\lambda_E = \lambda_0 + y_0^* E x_E = \lambda_0 + y_0^* E x_0 + y_0^* E (A - \lambda_0 I)^\sharp E x_0 + \ell_E$$

with $\ell_E = y_0^* E \xi_E$ and

$$|\ell_E| \le \frac{6\kappa_0^2 \|y_0\|_2}{s_0^2}\|E\|_2^3.$$

Proof. After a unitary similarity transformation we may assume that

$$A = \begin{bmatrix} \lambda_0 & A_{12} \\ 0 & A_{22} \end{bmatrix}, \qquad A_{22} - \lambda_0\, I \text{ nonsingular}, \qquad x_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad y_0^* = [1 \;\; -r] \in \mathbb{C}^{1\times n},$$

where $r = A_{12}(A_{22} - \lambda_0\, I)^{-1}$. Furthermore,

$$(A - \lambda_0\, I)^\sharp = \begin{bmatrix} 0 & A_{12}(A_{22} - \lambda_0\, I)^{-2} \\ 0 & (A_{22} - \lambda_0\, I)^{-1} \end{bmatrix} = \begin{bmatrix} 0 & r(A_{22} - \lambda_0\, I)^{-1} \\ 0 & (A_{22} - \lambda_0\, I)^{-1} \end{bmatrix}.$$

We will derive the result by specializing Theorem 3.6 to the case $A_{11} = \lambda_0$, $\|\cdot\| = \|\cdot\|_2$. In this case, $\mathbb{T}(Z) = (\lambda_0 I - A_{22})Z$, which implies

$$\mathrm{sep}_2(\lambda_0, A_{22}) = \|\mathbb{T}^{-1}\|^{-1} = \|(\lambda_0 I - A_{22})^{-1}\|_2^{-1} = \|(I - x_0 x_0^*)(A - \lambda_0 I)^\sharp\|_2^{-1} = s_0,$$

see also [14], Moreover,

$$\kappa = \sqrt{1 + \|r\|_2^2} + \|r\|_2 = \|y_0\|_2 + \sqrt{\|y_0\|_2^2 - 1} = \kappa_0.$$

Let $E \in \mathbb{C}^{n\times n}$ with $\|E\|_2 < 2\kappa_0/s_0$ According to Theorem 3.6 there exists a vector $w_E \in \mathbb{C}^{n-1}$ depending analytically on E such that

$$\|w_E\|_2 \leq \frac{2\kappa_0}{s_0\|y_0\|_2}\|E\|_2$$

and $x_E = \begin{bmatrix} 1 & r \\ 0 & I \end{bmatrix} \begin{bmatrix} 1 \\ w_E \end{bmatrix}$ is an eigenvector of $A + E$. Clearly, $y_0^* x_E = 1$. Let

$$\xi_E = \begin{bmatrix} r \\ 1 \end{bmatrix} (w_E - \mathbb{T}^{-1}(E_{21})) = \begin{bmatrix} r \\ 1 \end{bmatrix} (w_E - (\lambda_0 I - A_{22})^{-1} E_{21}).$$

Then

$$x_E = \underbrace{\begin{bmatrix} 1 \\ 0 \end{bmatrix}}_{=x_0} + \underbrace{\begin{bmatrix} r \\ I \end{bmatrix} (\lambda_0 I - A_{22})^{-1} E_{21}}_{=(A-\lambda_0 I)^\sharp E x_0} + \xi_E, \tag{4.2}$$

and from (3.8) it follows that $\|\xi_E\|_2 \leq \frac{6\kappa_0^2}{s_0^2}\|E\|_2^2$. This completes the proof of the statements about the eigenvector x_E. The statements about the eigenvalue follow by multiplying the relation $(A + E)x_E = \lambda_E x_E$ from the left with y_0^*. □

4.1 Application to eigenvectors of normal matrices

As mentioned in the introduction, one motivation of this work was to derive a perturbation bound with less restrictive conditions on $\|E\|$. Indeed for a normal matrix A and the matrix 2-norm, the condition $\|E\|_2 < \mathrm{sep}(A_{11}, A_{22})/2$ of Theorem 3.1 appears to be optimal as a larger perturbation level would lead to coalescence of the pseudospectral components containing $\Lambda(A_{11})$ and $\Lambda(A_{22})$. We will now specialize Theorem 3.1 to the case of an individual eigenvalue of a normal matrix.

Theorem 4.2 *Let $A \in \mathbb{C}^{n\times n}$ be a normal matrix and consider an eigenvector $v_0 \in \mathbb{C}^n$ belonging to a simple eigenvalue λ_0 of A. Let s_0 denote the distance of λ_0 to the rest of the spectrum of A_0, that is, $s_0 = \min\{|\lambda_0 - \nu| : \nu \in \Lambda(A), \nu \neq \lambda_0\}$. Let $E \in \mathbb{C}^{n\times n}$ with $\|E\|_2 \leq s_0/2$. Then there an eigenvector v_E of $A_0 + E$ such that*

$$\angle(v_0, v_E) \leq \arctan(2\,\|E\|_2/s_0),$$

where $\angle(v_0, v_E)$ denotes the angle between v_0 and v_E.

Proof. Let $U \in \mathbb{C}^{n\times n}$ be a unitary matrix such that $Uv_0 = [\gamma\ 0 \dots 0]^\top$ with $\gamma = \|v_0\|_2$. Then $A = U^{-1}\mathrm{diag}(\lambda_0, A_{22})U$ for some normal matrix A_{22}, and $E = U^{-1}\tilde{E}U$ with $\|\tilde{E}\|_2 = \|E\|_2$. By Theorem 3.1, the matrix $\mathrm{diag}(\lambda_0, A_{22}) + \tilde{E}$ has an eigenvector $v_E = [1\ z_E^\top]^\top$ with $z_E \in \mathbb{C}^{n-1}$ and $\|z_E\|_2 \le 2\|E\|/s$, where $s = \mathrm{sep}(\lambda_0, A_{22})$. Since A_{22} is normal, s equals the distance of λ_0 to the spectrum of A_{22}. Together with (2.8), this completes the proof. □

5 Conclusions

By establishing a link to the coalescence of pseudospectral components, we have derived a new perturbation bound for invariant subspaces. As the bound turns out to be sharp for a 2×2 example, no further obvious improvement of the bound seems to be possible. Moreover, we establish a novel bound for the remainder term of a well-known perturbation expansion. Even the (modified) specialization of this remainder bound to the case of individual eigenvectors and eigenvalues appears to be new. We believe that such bounds for remainder terms are important; e.g., they can be used for quantifying the validity for condition numbers frequently used in practice, e.g., in MATLAB and LAPACK [2].

As a side result, we have shown that Stewart's classical result on the perturbation of invariant subspaces is a direct consequence of the Newton-Kantorovich theorem. We believe that there is some interest in this observation, as it may more easily allow for extensions of Stewart's result to different settings.

6 Acknowledgment

The authors thank Pete Stewart for helpful discussions. Parts of this work were prepared while the first author was visiting FIM (Institute for Mathematical Research) at ETH Zurich. The generous hospitality of FIM is gratefully acknowledged.

A Stewart's result via the Newton-Kantorovich theorem

In this section, we show that Theorem 2.6 is a special case of the Newton-Kantorovich theorem formulated as in [10, p. 536].

Theorem A.1 *Let $\mathcal{E}, \mathcal{Z}$ be Banach spaces and let $f : \mathcal{Z} \to \mathcal{E}$ be twice continuously differentiable in a sufficiently large neighborhood Ω of $Z \in \mathcal{Z}$. Suppose that there exists a linear operator $\mathbb{T} : \mathcal{Z} \to \mathcal{E}$ having a continuous inverse $\mathbb{T}^{-1}$ and satisfying the following conditions:*

$$\|\mathbb{T}^{-1} f(Z)\| \le \eta, \tag{A.1}$$

$$\|\mathbb{T}^{-1} f'(Z) - I\| \le \delta, \tag{A.2}$$

$$\|\mathbb{T}^{-1} f''(\tilde{Z})\| \le K, \qquad \forall \tilde{Z} \in \Omega. \tag{A.3}$$

If $\delta < 1$ and $h := \frac{\eta K}{(1-\delta)^2} < \frac{1}{2}$ then there exists a solution Z_E of $f(Z_E) = 0$ such that

$$\|Z_E - Z\| \le r_0, \quad \textit{with} \quad r_0 := \frac{2\eta}{(1-\delta)(1+\sqrt{1-2h})}$$

We apply Theorem A.1 to the setting of Section 2.4:

$$f(Z) \equiv f(A+E, Z) = E_{21} + (A_{22}+E_{22})Z - Z(A_{11}+E_{11}) - Z(A_{12}+E_{12})Z$$

with $\mathcal{E} = \mathcal{Z} = \mathbb{C}^{(n-k)\times k}$, $Z = 0$, and $\mathbb{T} : Z \mapsto A_{22}Z - ZA_{11}$.

Condition (A.1)**.** We have

$$\|\mathbb{T}^{-1} f(0)\| = \|\mathbb{T}^{-1}(E_{21})\| \leq \frac{\|E_{21}\|}{s} =: \eta,$$

where $s = \mathrm{sep}(A_{11}, A_{22})$.

Condition (A.2)**.** From

$$f'(Z) : \triangle Z \mapsto (A_{22}+E_{22})\triangle Z - \triangle Z(A_{11}+E_{11}) = \mathbb{T}(\triangle Z) + \triangle\mathbb{T}(\triangle Z),$$

with $\triangle\mathbb{T}(\triangle Z) := E_{22}\cdot\triangle Z - \triangle Z\cdot E_{11}$, it follows that

$$\|\mathbb{T}^{-1} f'(Z) - I\| = \|\mathbb{T}^{-1}\triangle\mathbb{T}\| \leq \frac{\|E_{11}\| + \|E_{22}\|}{s} =: \delta.$$

Condition (A.3)**.** Since the second derivative of f is constant, it immediately follows that

$$\|\mathbb{T}^{-1} f''(\tilde{Z})\| \leq 2\frac{\|A_{12}+E_{12}\|}{s} \leq 2\frac{\|A_{12}\| + \|E_{12}\|}{s} =: K.$$

Summary. Setting $s_E = s - \|E_{11}\| - \|E_{22}\|$, we finally obtain

$$\begin{aligned} h &= \frac{\eta K}{(1-\delta)^2} = 2\frac{\|E_{21}\|(\|A_{12}\| + \|E_{12}\|)}{s_E^2} \\ r_0 &= \frac{2\eta}{(1-\delta)(1+\sqrt{1-2h})} = \frac{2\|E_{21}\|}{s_E + \sqrt{s_E^2 - 4\|E_{21}\|(\|A_{12}\| + \|E_{12}\|)}} \end{aligned}$$

Theorem A.1 now states the existence of a solution Z_E to $f(A+E, Z_E) = 0$ with $\|Z_E\| \leq r_0$ if $\delta < 1$ and $h < \frac{1}{2}$. This coincides precisely with the statement of Theorem 2.6.

References

[1] R. Alam and S. Bora. On stable eigendecompositions of matrices. *SIAM J. Matrix Anal. Appl.*, 26(3):830–848, 2005.

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

[3] R. Byers. A LINPACK-style condition estimator for the equation $AX - XB^T = C$. *IEEE Trans. Automat. Control*, 29(10):926–928, 1984.

[4] F. Chatelin. *Spectral approximation of linear operators*. Academic Press Inc., New York, 1983.

[5] C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, 7:1–46, 1970.

[6] J. W. Demmel. Computing stable eigendecompositions of matrices. *Linear Algebra Appl.*, 79:163–193, 1986.

[7] I. Gohberg, P. Lancaster, and L. Rodman. *Invariant subspaces of matrices with applications*, volume 51 of *Classics in Applied Mathematics*. SIAM, Philadelphia, PA, 2006. Reprint of the 1986 original.

[8] L. Grammont and A. Largillier. On ϵ-spectra and stability radii. *J. Comput. Appl. Math.*, 147(2):453–469, 2002.

[9] M. Gu and M. L. Overton. An algorithm to compute Sep_λ. *SIAM J. Matrix Anal. Appl.*, 28(2):348–359, 2006.

[10] L. V. Kantorovich and G. P. Akilov. *Functional analysis*. Pergamon Press, Oxford, second edition, 1982.

[11] M. Karow. Inclusion theorems for pseudospectra of block triangular matrices. 2012. SIAM Conference on Applied Linear Algebra, Valencia. Manuscript in preparation.

[12] T. Kato. *Perturbation theory for linear operators*. Classics in Mathematics. Springer-Verlag, Berlin, 1995.

[13] D. Kressner. *Numerical Methods for General and Structured Eigenvalue Problems*, volume 46 of *Lecture Notes in Computational Science and Engineering*. Springer-Verlag, Berlin, 2005.

[14] C. D. Meyer and G. W. Stewart. Derivatives and perturbations of eigenvectors. *SIAM J. Numer. Anal.*, 25(3):679–691, 1988.

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

[16] G. W. Stewart. Error and perturbation bounds for subspaces associated with certain eigenvalue problems. *SIAM Rev.*, 15:727–764, 1973.

[17] G. W. Stewart and J.-G. Sun. *Matrix Perturbation Theory*. Academic Press, New York, 1990.

[18] J.-G. Sun. Perturbation expansions for invariant subspaces. *Linear Algebra Appl.*, 153:85–97, 1991.

[19] J.-G. Sun. Stability and accuracy: Perturbation analysis of algebraic eigenproblems. Technical report UMINF 98-07, Department of Computing Science, University of Umeå, Umeå, Sweden, 1998. Revised 2002.

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

[21] J. M. Varah. On the separation of two matrices. *SIAM J. Numer. Anal.*, 16(2):216–222, 1979.

Recent publications :

MATHEMATICS INSTITUTE OF COMPUTATIONAL SCIENCE AND ENGINEERING
Section of Mathematics
Ecole Polytechnique Fédérale
CH-1015 Lausanne

44.2012 D. KRESSNER, X. LIU:
Structured canonical forms for products of (skew-)symmetric matrices and the matrix equation XAX = B

45.2012 B. BECKERMANN, D. KRESSNER, C. TOBLER:
An error analysis of Galerkin projection methods for linear systems with tensor product structure

46.2012 J. BECK, F. NOBILE, L. TAMELLINI, R. TEMPONE:
A quasi-optimal sparse grids procedure for groundwater flows

47.2012 A. ABDULLE, Y. BAI:
Fully discrete analysis of the heterogeneous multiscale method for elliptic problems with multiple scales

48.2012 G. MIGLORATI, F. NOBILE, E. VON SCHWERIN, R. TEMPONE:
Approximation of quantities of interest in stochastic PDES by the random discrete L^2 projection on polynomial spaces

01.2013 A. ABDULLE, A. BLUMENTHAL:
Stabilized multilevel Monte Carlo method for stiff stochastic differential equations

02.2013 D. N. ARNOLD, D. BOFFI, F. BONIZZONI:
Tensor product finite element differential forms and their approximation properties

03.2013 N. GUGLIELMI, D. KRESSNER, C. LUBICH:
Low-rank differential equations for Hamiltonian matrix nearness problems

04.2013 P. CHEN, A. QUARTERONI, G. ROZZA:
A weighted reduced basis method for elliptic partial differential equations with random input data

05.2013 P. CHEN, A. QUARTERONI, G. ROZZA:
A weighted empirical interpolation method: a priori convergence analysis and applications

06.2013 R. SCHNEIDER, A. USCHMAJEW:
Approximation rates for the hierarchical tensor format in periodic Sobolev spaces

07.2013 C. BAYER, H. HOEL, E. VON SCHWERIN, R. TEMPONE:
On non-asymptotic optimal stopping criteria in Monte Carlo simulation.

08.2013 L. GRASEDYCK, D. KRESSNER, C. TOBLER:
A literature survey of low-rank tensor approximation techniques

09.2012 M. KAROW, D. KRESSNER:
On a perturbation bound for invariant subspaces of matrices