Michael Karow

Geometry of Spectral Value Sets

Dissertation, Juli 2003

Geometry of Spectral Value Sets

von Michael Karow

Dissertation
zur Erlangung des Grades eines Doktors der Naturwissenschaften
– Dr. rer. nat. –

Vorgelegt im Fachbereich 3 (Mathematik & Informatik)
der Universität Bremen
im Juni 2003

Datum des Promotionskolloquiums: 4. Juli, 2003

Gutachter: Prof. Dr. Diederich Hinrichsen (Bremen)
Prof. Dr. Anthony J. Pritchard (Warwick, England)

Contents

Introduction

Perturbation theory of eigenvalues is a classical and important branch of linear algebra. Its applications are multifaceted in areas such as numerical linear algebra, linear systems theory, theoretical physics and engineering. During the last 20 years the concept of spectral value sets, also known under the name pseudospectra, emerged as a new and powerful tool in eigenvalue perturbation theory.

In the broadest sense a spectral value set consists of all eigenvalues of a family of matrices which depend continuously on a parameter in a metric space. Usually the metric space is a subset of a real or complex vector space containing 0. The matrix corresponding to the parameter 0 is said to be the nominal matrix and the whole family is regarded as the set of possible perturbations of this matrix. The size of the parameter, measured in some norm, reflects the level of perturbation of the nominal matrix. By visualizing spectral value sets via the computer one gets insight into the mobility of the eigenvalues under the perturbations in question. This is in particular useful for the stability analysis of uncertain linear systems. A linear system is said to be stable with respect to a given stability region in the complex plane if all eigenvalues of the system matrix lie in that region. In this context the nominal matrix is regarded as an approximation to a system matrix whose exact value is unknown, and the perturbation level is the level of uncertainty. If the associated spectral value set is contained in the stability region then one knows that the exact system matrix is stable.

In this work we concentrate on spectral value sets for additive matrix perturbations of the form $A \rightsquigarrow A(\Delta) = A + B\Delta C$, where A, B, C are fixed matrices and the parameter Δ is an element of a given set of matrices $\mathbf{\Delta}$. The set $\mathbf{\Delta}$ is assumed to be closed and star-shaped with respect to zero. We will see that this framework is flexible enough to cover all kinds of linear perturbation structures. The idea to consider perturbations of the form $A \rightsquigarrow A + B\Delta C$ and not merely those of the form $A \rightsquigarrow A + \Delta$ is due to Hinrichsen and Pritchard [54, 42, 43, 54, 52]. They were motivated by the following basic system theoretic fact. Suppose we are given a time invariant linear system of the form

$$\dot{x}(t) = Ax(t) + Bu(t), \qquad y(t) = Cx(t).$$

Then the feedback law $u(t) = \Delta y(t)$ yields a new linear system whose matrix is $A + B\Delta C$. The transfer function $G(s) = C(sI - A)^{-1}B$, $s \in \mathbb{C}$, associated with the linear system (A, B, C) plays an important role in the theory of spectral value sets. If both B and C equal the identity matrix and the set $\mathbf{\Delta}$ is the set of all complex $n \times n$ matrices then the associated spectral value sets are called pseudospectra in the narrower sense. These sets are known to contain information about the transient behaviour of a single linear system [45, 46, 47, 77, 92, 93]. An introduction to the basic theory and history of pseudospectra can be found on the web site [26] maintained by M. Embree and N. Trefethen. A collection of examples and a comprehensive bibliography is included.

This work is devoted to the study of geometric properties of spectral value sets for *structured* perturbations, but we also treat pseudospectra in some detail. A good half of the thesis is about spectral value sets with respect to real perturbations. Despite their importance such sets have been investigated only rarely in the literature, see however [43, 51]. One reason is that the theory of real spectral value sets is difficult. The associated computational problem has been solved (in principle) only for the case that the underlying norm is the spectral norm.

The present thesis benefits much from the work of other researchers in the field, especially from the Ph.D. theses of Gallestey [30] and Harrabi [37] and the papers of Bernhardsson, Qiu and Rantzer [6, 7, 8, 9].

Outline of the thesis

Chapter 1 contains preliminary material and establishes some of our notation. Most of the statements of this chapter are from elementary linear algebra and are well known. However, there are some results which we could not find in the literature, for instance Proposition 1.6.1 which is about real and complex operator norms of a real matrix.

In Chapter 2 we give the precise definition of spectral value sets and study their basic properties. We first consider topological properties of a single spectral value set. Then we prove that spectral value sets depend continuously on all parameters. Next, we show that each spectral value set is the superlevel set of an upper semi-continuous function f. The function f is the composition of a rational matrix function and the μ-function associated with the underlying perturbation structure. By means of this characterization spectral value sets can be calculated provided that the μ-function in question is computable. In Section 5 we discuss cases in which the μ-function equals an operator norm. The last section of the chapter contains our first main result. It is about the relationship between the behaviour of spectral value sets for small perturbations, μ-functions, and condition numbers of eigenvalues with respect to structured matrix perturbations.

In Chapter 3 we show that spectral value sets and μ-functions with respect to a large class of perturbation structures are semi-algebraic. Semi-algebraic spectral value sets have piecewise analytic boundary, and semi-algebraic μ-functions are real analytic on an open and dense subset of $\mathbb{C}^{q\times l}$.

Spectral value sets for structued perturbations of block diagonal matrices are discussed in Chapter 4. The perturbations are not necessarily block diagonal but they are assumed to have zero blocks at some precribed positions. Such perturbation structures occur in the stability analysis of composite linear systems. We give formulas for the associated μ-function with respect to two classes of norms. By specializing our results to the diagonal case we obtain a new proof for the classical eigenvalue inclusion theorem of Brualdi [18]. The latter implies the inclusion theorems of Brauer and Gershgorin.

Chapter 5 is about complex pseudospectra. These are spectral value sets with respect to unstructured perturbations. First, we give several upper and lower bounds for pseudospectra. Then we discuss the relationship between pseudospectra and the nonnonormality of the perturbed matrix. In addition, we provide a reformulation of the eigenvalue inclusion theorem of Henrici [38] in terms of pseudospectra and extend it to block triangular matrices. Finally, we prove a generalization of a result of Ruhe [80], which is a refinement of Henrici's theorem.

In Chapters 6 and 7 we study spectral value sets with respect to *real* perturbations. If the underlying norm is the spectral norm, then the associated μ-functions are called real perturbation values. Chapter 6 gives an introduction to the theory of real perturbation values as developed by Bernhardsson, Qiu and Rantzer [7, 8, 9]. We introduce higher μ-functions and study their basic properties, as we believe that real perturbation values are best understood in this framework. The main goal of the first part of Chapter 6 is to establish a theorem which characterizes the real perturbation values by hermitian-symmetric inequalities. This characterization will be used frequently in the sequel. In the second part of Chapter 6 we study special properties of real μ-functions. In particular we show that for a large class of norms the associated real μ-function, $M \mapsto \mu_{\mathbb{R}}(M)$, is continuous at all nonreal matrices M.

Chapter 7 deals with real spectral value sets for special matrices. In the first section we give a formula for the largest real perturbation value of block diagonal matrices. The proof of this formula is based on the characterization of real perturbation values by hermitian-symmetric inequalities. As an application real spectral value sets of diagonal matrices and real pseudospectra of normal matrices are determined in Sections 7.2 and 7.3. Real pseudospectra of 2×2 matrices with respect to arbitrary unitarily invariant norms are treated in Section 7.4. We close the chapter with the discussion of root sets of uncertain polynomials.

The appendix deals with hermitian-symmetric inequalities. It is mainly a review of the theory which has been developed by Bernhardsson, Qiu and Rantzer. First, we discuss the canonical form of hermitian symmetric matrix pairs with respect to simultaneous congruence transformations. This canonical form is used to prove a criterion for the existence of a solution of a hermitian-symmetric inequality. This in turn leads to a formula for the computation of real perturbation values.

Acknowledgements

First and foremost, I thank my advisor Prof. Dr. D. Hinrichsen for introducing me to the topic of spectral value sets, for his support, guidance and valuable comments, but also for the freedom he has given me in my research. I really enjoyed to work in his research group “Regelungssysteme” at the University of Bremen. Furthermore, I would like to thank Prof. Dr. A.J. Pritchard from Warwick University for fruitful discussions and for his readiness to review this thesis. I am also indebted to Prof. Dr. Volker Mehrmann who gave me the opportunity to continue my research as a member of his group “Modellierung, Numerik, Differentialgleichungen” at the Berlin University of Technology. I would like to thank all members and visitors of both research groups for creating a very pleasant working atmosphere. Special thanks go to Fabian Wirth and Daniel Kressner for reading and critizising parts of the manuscript, and to Niloufer Mackey for her help in the last minute. I also thank Tobias Damm for many stimulating discussions on mathematics and philosophy.

I would like to express my deep gratitude to my parents for their encouragement and support over the years.

Last but not least I thank the Deutsche Forschungsgemeinschaft for the financial support of this work through the Graduiertenkolleg “Komplexe Dynamische Systeme”.

Basic notation

Below we list our basic notation. More notation will be introduced as needed.

$\mathbb{Z}$	the set of integers
$\mathbb{N}$	the set positive integers $\{1, 2, \ldots\}$
$\mathbb{N}_0$	nonnegative integers $\{0, 1, 2, \ldots\}$
$\underline{n}$	set of positive integers less or equal than n, $\underline{n} = \{1, \ldots, n\}$
$\#S$	cardinality of the set S.
$\mathbb{R}, \mathbb{C}$	field of real, complex numbers
$\mathbb{F}$	$\mathbb{R}$ or $\mathbb{C}$
$\mathbb{F}^{m\times n}$	vector space of matrices with m columns and n rows
$\mathbb{F}^m$	equals $\mathbb{F}^{m\times 1}$
$A = [a_{jk}]$	matrix $A \in \mathbb{F}^{m\times n}$ with entries $a_{jk} \in \mathbb{F}$
$\overline{A}$	complex conjugate of the matrix A, $\overline{A} = [\,\overline{a_{jk}}\,]$
$\Re A$	the real part of A, $\Re A = [\Re a_{jk}]$
$\Im A$	the imaginary part of A, $\Im A = [\Im a_{jk}]$
$\lvert A\rvert$	matrix of absolute values, $\lvert A\rvert = [\ \lvert a_{jk}\rvert\ ]$.
$A^T,\ A^*$	the transpose and the conjugate transpose of A, $A^* = \overline{A}^T$
$\ker A$	kernel of a matrix
$\operatorname{range} A$	range of a matrix
$\operatorname{rank} A$	rank of a matrix
$\sigma(A)$	spectrum of the matrix A
$\varrho(A)$	spectral radius of A, equals $\max\{\ \lvert\lambda\rvert \mid \lambda \in \sigma(A)\ \}$
$\sigma_j(A)$	the jth singular value of A
$\operatorname{tr}(A)$	trace
$\det(A)$	determinant
I_n	$n \times n$ unit matrix
I	unit matrix, whose dimension is clear by context
i	imaginary unit or index, depending on context
$\operatorname{span}_{\mathbb{F}}(S)$	vector space over $\mathbb{F}$ spaned by the set S
$U^\perp$	the orthogonal complement of the subspace $U \subseteq \mathbb{F}^m$ with respect to the standard inner product.
$\operatorname{cl}(S),\ \operatorname{int}(S)$	the closure of the set S
∂S	the boundary of S

For $A = [a_{jk}] \in \mathbb{R}^{m\times n}$, $B = [b_{jk}] \in \mathbb{R}^{m\times n}$ we write

$0 \le A$ if A is nonnegative, i.e. $a_{jk} \ge 0$ for all $(j,k) \in \underline{m} \times \underline{n}$,

$A \le B$ if $a_{jk} \le b_{jk}$ for all $(j,k) \in \underline{m} \times \underline{n}$.

By $\mathcal{D}(\lambda, \rho)$ we denote the closed disk of radius $\rho \ge 0$ about $\lambda \in \mathbb{C}$,

$$\mathcal{D}(\lambda, \rho) = \{\ s \in \mathbb{C} \mid \ \lvert s - \lambda\rvert \le \rho\ \}.$$

Chapter 1

Preliminaries

This chapter contains some basic facts which will be used frequently in this work. Here we also introduce some of our notation. Most of the material is elementary and well known. We give proofs for all statements which can not be found in the text books [61, 62, 86].

1.1 Vector norms

Let X be a vector space over $\mathbb{F}\in\{\mathbb{R},\mathbb{C}\}$. A function $\|\cdot\|: X \to [0,\infty)$ is said to be a norm on X if the following conditions hold for all $x, y\in X$ and all $\lambda\in\mathbb{F}$.

1. $\|x\| = 0$ if and only if $x = 0$.
2. $\|\lambda x\| = |\lambda|\,\|x\|$.
3. $\|x + y\| \le \|x\| + \|y\|$.

A norm $\|\cdot\| : \mathbb{F}^n \to [0,\infty)$ is said to be

- *symmetric* if $\|x\| = \|Px\|$ for all permutation matrices $P\in\{0,1\}^{n\times n}$ and all $x\in\mathbb{F}^n$.
- *absolute* if $\|x\| = \|\,|x|\,\|$ for all $x\in\mathbb{F}^n$.

An absolute norm has the following monotonicity property: if $|x| \le |y|$ then $\|x\| \le \|y\|$. The most important absolute and symmetric norms on $\mathbb{C}^n$ are the Hölder p-norms defined by

$$\|x\|_p := \left(\sum_{k\in\underline{n}} |x_k|^p\right)^{1/p} \quad \text{for } p\in[1,\infty), \qquad \|x\|_\infty := \lim_{p\to\infty} \|x\|_p = \max_{k\in\underline{n}} |x_k|.$$

The norms $\|\cdot\|_\infty$, $\|\cdot\|_1$ and $\|\cdot\|_2$ are called the maximum, sum and Euclidean norm, respectively. The Hölder norms satisfy the following inequalities for all $p_1, p_2\in[0,\infty]$ and all $x\in\mathbb{C}^n$.

$$\text{If } p_1 \le p_2 \text{ then } \|x\|_{p_2} \le \|x\|_{p_1} \le n^{\frac{1}{p_1}-\frac{1}{p_2}}\|x\|_{p_2}. \qquad (\text{where } 1/\infty := 0)$$

These inequalities become equalities for $x = \begin{bmatrix}1 & 0 & \dots & 0\end{bmatrix}^T$ resp. $x = \begin{bmatrix}1 & 1 & \dots & 1\end{bmatrix}^T$. The dual norm to a given norm $\|\cdot\|$ on $\mathbb{F}^n$ is defined by

$$\|\cdot\|^D : \mathbb{F}^n \to [0,\infty) \qquad \|x\|^D := \max_{\|y\|=1} |x^T y|.$$

For every norm on $\mathbb{F}^n$ one has that $\left(\|\cdot\|^D\right)^D = \|\cdot\|$. The dual to a Hölder norm is again a Hölder norm. Precisely,

$$\|\cdot\|_p^D = \|\cdot\|_q, \qquad \text{where } \frac{1}{p} + \frac{1}{q} = 1.$$

1.2 The singular value decomposition

Let $A \in \mathbb{C}^{m\times n} \setminus \{0\}$ and $r = \operatorname{rank} A$. Let $\sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq \sigma_r(A) > 0$ be the square roots of the nonzero eigenvalues of A^*A counting multiplicities. Then there exists an orthonormal basis $v_1, \ldots, v_r$ of $(\ker A)^\perp$ such that $A^*A\, v_j = \sigma_j(A)^2 v_j$ for all $j \in \underline{r}$. Set $u_j := Av_j/\|Av_j\|_2$. Then the vectors $u_1, \ldots, u_r$ form an orthonormal basis of $\operatorname{range} A$ and we have $AA^*\, u_j = \sigma_j(A)^2\, u_j$, $Av_j = \sigma_j(A) u_j$, $A^* u_j = \sigma_j(A) v_j$ for $j \in \underline{r}$. Moreover, A has the representation

$$A = \sum_{j=1}^{r} \sigma_j(A)\, u_j v_j^*. \tag{1.1}$$

The numbers $\sigma_j(A)$ are called the singular values of A. We set $\sigma_j(A) := 0$ for all $j > \operatorname{rank} A$, $j \in \mathbb{N}$. The singular values of zero matrices are defined to be zero. We sometimes use the notation

$$\sigma_{\max}(A) := \sigma_1(A), \qquad \sigma_{\min}(A) := \sigma_{\min\{m,n\}}(A).$$

Then $\sigma_{\min}(A) \neq 0$ if and only A has full rank. The representation (1.1) is said to be a singular value decomposition of A. The vectors v_j and u_j are called left resp. right singular vectors of A. Let $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ be unitary matrices whose first r columns are $u_1, \ldots, u_r$ resp. $v_1, \ldots, v_r$. Then

$$A = U\Sigma V^*, \qquad \text{where } \Sigma = \begin{bmatrix} \Sigma_+ & 0 \\ 0 & 0 \end{bmatrix}, \ \Sigma_+ = \operatorname{diag}(\sigma_1(A), \ldots, \sigma_r(A)). \tag{1.2}$$

This factorization is also called a singular value decomposition. It implies that

$$\begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix} = \begin{bmatrix} U & -U \\ V & V \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{bmatrix} \begin{bmatrix} U & -U \\ V & V \end{bmatrix}^*.$$

Thus, the nonzero singular values of A are the positive eigenvalues of the hermitian matrix $\begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix}$. Finally, we note that if $A \in \mathbb{R}^{m\times n}$ is a real matrix then the singular vectors u_j, v_j can be chosen to be real.

1.3 The Moore-Penrose generalized inverse

For any matrix $A \in \mathbb{C}^{m\times n}$ the linear map $f_A : (\ker A)^\perp \to \operatorname{range} A$, $f_A(x) = Ax$, is bijective. Thus, there is a unique matrix $A^\dagger \in \mathbb{C}^{n\times n}$ such that

$$A^\dagger x = \begin{cases} f_A^{-1}(x) & \text{for } x \in \operatorname{range} A \\ 0 & \text{for } x \in (\operatorname{range} A)^\perp. \end{cases}$$

The matrix $A^\dagger$ is called the Moore-Penrose generalized inverse of A. If A has the singular value decompositions (1.1) and (1.2) then

$$A^\dagger = \sum_{j=1}^{r} \sigma_j(A)^{-1}\, v_j u_j^* = V \begin{bmatrix} \Sigma_+^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^*.$$

Moreover, $A^\dagger$ equals the unique matrix $X \in \mathbb{C}^{n\times n}$ which satisfies the following so called Penrose conditions.

$$\begin{aligned} AXA &= A, \\ XAX &= X, \\ (AX)^* &= AX, \\ (XA)^* &= XA. \end{aligned}$$

The matrices $A^\dagger A$ and $AA^\dagger$ are the orthogonal projectors onto $(\ker A)^\perp$ and range A respectively. If A has full column rank then $A^\dagger = (A^*A)^{-1}A^*$. If $A \in \mathbb{C}^{n\times n}$ is nonsingular then $A^\dagger = A^{-1}$. If $A \in \mathbb{R}^{m\times n}$ then $A^\dagger \in \mathbb{R}^{n\times m}$.

1.4 Unitarily invariant norms

A norm $\|\cdot\| : \mathbb{C}^{m\times n} \to [0,\infty)$ is called unitarily invariant if $\|UAV\| = \|A\|$ for all $A \in \mathbb{C}^{m\times n}$ and all unitary matrices $U \in \mathbb{C}^{m\times m}, V \in \mathbb{C}^{n\times n}$. Each unitarily invariant norm is a function of the singular values of the matrix. More precisely, the following statement holds [62, Theorem 3.5.18].

Theorem 1.4.1 *Let $|||\cdot|||$ be an absolute and symmetric norm on $\mathbb{R}^{\min\{m,n\}}$. Then the function*

$$\|\cdot\| : \mathbb{C}^{m\times n} \to [0,\infty) \qquad \|A\| := |||\ [\,\sigma_1(A), \ldots, \sigma_{\min\{m,n\}}(A)\,]^T\ ||| \tag{1.3}$$

is a unitarily invariant norm. All unitarily invariant norms on $\mathbb{C}^{m\times n}$ have such a representation.

A unitarily invariant norm is called normalized if $\|A\| = \sigma_1(A)$ for all matrices $A \in \mathbb{C}^{m\times n}$ of rank 1. Examples for normalized unitarily invariant norms are the Schatten p-norms:

$$\|A\|_{(p)} := \|\ [\,\sigma_1(A), \ldots, \sigma_{\min\{m,n\}}(A)\,]^T\ \|_p, \qquad p \in [1,\infty].$$

The norm $\|A\|_{(\infty)} = \sigma_1(A)$ is the spectral norm, and the norm

$$\|A\|_{(2)} = \left(\textstyle\sum_{k=1}^{\min\{m,n\}} \sigma_k(A)^2\right)^{\frac{1}{2}} = (\mathrm{tr}(A^*A))^{\frac{1}{2}} =: \|A\|_F$$

is the Frobenius norm.

1.5 Operator norms

Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be arbitrary norms on $\mathbb{F}^n$ and $\mathbb{F}^m$ respectively. Then the function

$$\|\cdot\|_{\alpha,\beta} : \mathbb{F}^{m\times n} \to [0,\infty) \qquad \|A\|_{\alpha,\beta} := \max_{x\in\mathbb{F}^n\setminus\{0\}} \frac{\|Ax\|_\beta}{\|x\|_\alpha} = \max_{\|x\|_\alpha=1} \|Ax\|_\beta.$$

is a norm on $\mathbb{F}^{m\times n}$. This norm is said to be the operator norm induced by $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ It is also called the operator norm subordinate to the norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$. Note that in this notation the indices α and β are used just to distinguish different norms. They don't have any specific meaning except for the case that $\alpha,\beta\in[1,\infty]$. In this case $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ denote Hölder norms. We list the most important operator norms. Let $A=[a_{jk}]\in\mathbb{C}^{m\times n}$. Then

- $\|A\|_{\infty,\infty}=\max_{j\in\underline{m}}\sum_{k\in\underline{n}}|a_{jk}|$ row sum norm
- $\|A\|_{1,1}=\max_{k\in\underline{n}}\sum_{j\in\underline{m}}|a_{jk}|$ column sum norm
- $\|A\|_{1,\infty}=\max_{j,k\in\underline{n}}|a_{jk}| =: \|A\|_{\max}$ componentwise maximum norm
- $\|A\|_{2,2}=(\varrho(A^*A))^{\frac{1}{2}}=\sigma_{\max}(A)$ spectral norm

In the sequel we simplify the notation and write $\|\cdot\|_\alpha$ instead of $\|\cdot\|_{\alpha,\alpha}$. It is easily seen that for every norm $\|\cdot\|_\alpha$ on $\mathbb{C}^n$,

$$\varrho(A)\le\|A\|_\alpha,\qquad A\in\mathbb{C}^{n\times n}.$$

Operator norms are submultiplicative, i.e.

$$\|AB\|_{\alpha,\gamma}\le\|A\|_{\beta,\gamma}\|B\|_{\alpha,\beta}$$

for all $A\in\mathbb{F}^{m\times n}$, $B\in\mathbb{F}^{n\times q}$ and all norms $\|\cdot\|_\alpha$ on $\mathbb{F}^q$, $\|\cdot\|_\beta$ on $\mathbb{F}^n$ and $\|\cdot\|_\gamma$ on $\mathbb{F}^m$. In particular one has for all nonsingular $A\in\mathbb{F}^{n\times n}$,

$$1=\|I_n\|_\alpha\le\|A^{-1}\|_{\beta,\alpha}\|A\|_{\alpha,\beta}.$$

An important consequence of the Hahn-Banach theorem is the following proposition.

Proposition 1.5.1 *Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be arbitrary norms on $\mathbb{F}^n$ resp. $\mathbb{F}^m$. Then for all $A\in\mathbb{F}^{m\times n}$,*

$$\|A\|_{\alpha,\beta}=\max_{\substack{x\in\mathbb{C}^n\setminus\{0\}\\ y\in\mathbb{C}^m\setminus\{0\}}}\frac{|y^TAx|}{\|x\|_\alpha\|y\|_\beta^D}=\max_{\|x\|_\alpha=\|y\|_\beta^D=1}|y^TAx|.$$

A matrix $A\in\mathbb{F}^{m\times n}$ of rank 1 can be written in the form $A=xy^T$ with $x\in\mathbb{F}^m$, $y\in\mathbb{F}^n$, and one has

$$\|xy^T\|_{\alpha,\beta}=\|x\|_\beta\,\|y\|_\alpha^D. \tag{1.4}$$

A quantity closely related to operator norms is the following.

$$\inf\nolimits_{\alpha,\beta}(A):=\min_{x\in\mathbb{F}^n\setminus\{0\}}\frac{\|Ax\|_\beta}{\|x\|_\alpha}=\min_{\|x\|_\alpha=1}\|Ax\|_\beta\qquad A\in\mathbb{F}^{m\times n}.$$

If $m<n$ then $\inf_{\alpha,\beta}(A)=0$. If $m=n$ then

$$\inf\nolimits_{\alpha,\beta}(A)=\begin{cases}0 & \text{if } A \text{ is singular,}\\ \|A^{-1}\|_{\beta,\alpha}^{-1} & \text{otherwise.}\end{cases} \tag{1.5}$$

For the 2-norm one has,

$$\inf\nolimits_{2,2}(A)=\sigma_{\min}(A),\qquad A\in\mathbb{C}^{n\times n}.$$

For diagonal matrices the following statement holds.

Proposition 1.5.2 *Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be absolute norms on $\mathbb{F}^n$ such that $\|I_n\|_{\alpha,\beta} = 1$ and $\|e_k\|_\alpha = \|e_k\|_\beta$ for all standard basis vectors $e_1, \ldots, e_n$ of $\mathbb{F}^n$. Let $A = \operatorname{diag}(a_1, \ldots, a_n) \in \mathbb{F}^{n\times n}$ be a diagonal matrix. Then*

$$\|A\|_{\alpha,\beta} = \max_{k\in\underline{n}} |a_k| \qquad \text{and} \qquad \inf_{\beta,\alpha}(A) = \min_{k\in\underline{n}} |a_k|.$$

Proof: Let $k_0 \in \underline{n}$ be an index such that $|a_{k_0}| = \max_{k\in\underline{n}} |a_k|$. Then we have for all $x\in\mathbb{F}^n$ that $|Ax| \le |a_{k_0}|\,|x|$, and therefore

$$\|Ax\|_\beta = \|\,|Ax|\,\|_\beta \le \|\,|a_{k_0}|\,|x|\,\|_\beta = |a_{k_0}|\,\|\,|x|\,\|_\beta \le |a_{k_0}|\,\|\,|x|\,\|_\alpha = |a_{k_0}|\,\|x\|_\alpha.$$

(The left inequality holds by the monotonicity of absolute norms and the right inequality follows from the assumption that $\|I_n\|_{\alpha,\beta} = 1$.) Hence $\|A\|_{\alpha,\beta} \le |a_{k_0}|$. On the other hand, $\|Ae_{k_0}\|_\beta = |a_{k_0}|\,\|e_{k_o}\|_\beta = |a_{k_0}|\,\|e_{k_o}\|_\alpha$. Thus $\|A\|_{\alpha,\beta} = |a_{k_0}| = \max_{k\in\underline{n}} |a_k|$. Now, the relation $\inf_{\beta,\alpha}(A) = \min_{k\in\underline{n}} |a_k|$ follows from (1.5). $\square$

Note that the assumptions in Proposition 1.5.2 are satisfied for the Hölder norms $\|\cdot\|_\alpha, \|\cdot\|_\beta$ if $1 \le \alpha \le \beta \le \infty$. For these norms the following generalization of Proposition 1.5.2 holds.

Proposition 1.5.3 *Let $1 \le p_1 \le p_2 \le \infty$. Let $A = \operatorname{diag}(A_1, \ldots, A_m)$, where $A_k \in \mathbb{C}^{n_k\times n_k}$, $k\in\underline{m}$. Then*

$$\|A\|_{p_1,p_2} = \max_{k\in\underline{m}} \|A_k\|_{p_1,p_2} \qquad \text{and} \qquad \inf_{p_2,p_1}(A) = \min_{k\in\underline{m}} \inf_{p_2,p_1}(A_k).$$

Proof: Let $c := \max_{k\in\underline{m}} \|A_k\|_{p_1,p_2}$ and let $x = [x_1^T, \ldots, x_m^T]^T$, where $x_k \in \mathbb{C}^{n_k}$, $k \in \underline{m}$. For $\xi = [\|x_1\|_{p_1}, \ldots, \|x_m\|_{p_1}]^T \in \mathbb{R}^m$ it is easily seen that $\|\xi\|_{p_1} = \|x\|_{p_1}$. Since $p_1 \le p_2$ we have that $\|\xi\|_{p_2} \le \|\xi\|_{p_1}$. Therefore

$$\|Ax\|_{p_2}^{p_2} = \sum_{k\in\underline{m}} \|A_k x_k\|_{p_2}^{p_2} \le c^{p_2} \sum_{k\in\underline{m}} \|x_k\|_{p_1}^{p_2} = c^{p_2}\|\xi\|_{p_2}^{p_2} \le c^{p_2}\|\xi\|_{p_1}^{p_2} = c^{p_2}\|x\|_{p_1}^{p_2}.$$

Thus $\|A\|_{p_1,p_2} \le c$. Without loss of generality we may assume that $c = \|A_1\|_{p_1,p_2}$. Let $x_0 \in \mathbb{C}^{n_1}$ such that $\|A_1 x_0\|_{p_2} = c\|x_0\|_{p_1}$, and let $\widetilde{x}_0 = [x_0^T\ 0\ \ldots 0]^T \in \mathbb{C}^n$, $n = \sum_{k\in\underline{m}} n_k$. Then $\|A\widetilde{x}_0\|_{p_2} = c\|\widetilde{x}_0\|_{p_2}$. Thus $\|A\| = c$. Now, the relation $\inf_{p_2,p_1}(A) = \min_{k\in\underline{m}} \inf_{p_2,p_1}(A_k)$ follows from (1.5). $\square$

1.6 Operator norms of a real matrix

Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be norms on $\mathbb{C}^n$ and $\mathbb{C}^m$ respectively. Moreover, let $A\in\mathbb{R}^{m\times n}$ be a matrix with real elements. Then there are two operator norms of A with respect to the norms $\|\cdot\|_\alpha$, $\|\cdot\|_\beta$ depending on whether the maximum of the quotient $\|Ax\|_\beta/\|x\|_\alpha$ is taken over all $x\in\mathbb{C}^n\setminus\{0\}$ or over all $x\in\mathbb{R}^n\setminus\{0\}$. We distinguish these norms by a superscript $\mathbb{F}$, i.e. we set

$$\|A\|_{\alpha,\beta}^{\mathbb{F}} := \max_{x\in\mathbb{F}^n\setminus\{0\}} \frac{\|Ax\|_\beta}{\|x\|_\alpha}, \qquad \mathbb{F} = \mathbb{C} \text{ or } \mathbb{R}.$$

Obviously, $\|A\|_{\alpha,\beta}^{\mathbb{C}} \ge \|A\|_{\alpha,\beta}^{\mathbb{R}}$. In this section we address the question for which types of vector norms $\|\cdot\|_\alpha$, $\|\cdot\|_\beta$ the corresponding operator norms $\|A\|_{\alpha,\beta}^{\mathbb{C}}$ and $\|A\|_{\alpha,\beta}^{\mathbb{R}}$ coincide. This is not always the case as the following counterexample (from [53]) shows. Let $A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$. Then, for every $x = [x_1\ x_2]^T \in \mathbb{C}^2 \setminus \{0\}$,

$$\frac{\|Ax\|_1}{\|x\|_\infty} = \frac{|x_1 - x_2| + |x_1 + x_2|}{\max\{|x_1|, |x_2|\}}.$$

Thus, $\|A\|_{\infty,1}^{\mathbb{R}} = 2$ but $\|A\|_{\infty,1}^{\mathbb{C}} \geq |1-i| + |1+i| = 2\sqrt{2}$.

The result below will be used in Section 6.8.

Proposition 1.6.1 *Let $\|\cdot\|_\alpha$, $\|\cdot\|_\beta$ be norms on $\mathbb{C}^n$ and $\mathbb{C}^m$ respectively, and suppose that one of the following conditions holds.*

(*a*) $n = 1$.

(*b*) *The norm $\|\cdot\|_\alpha$ is absolute and $m = 1$.*

(*c*) *The norm $\|\cdot\|_\alpha$ is absolute and $\|\cdot\|_\beta$ is a weighted maximum norm, i.e. there are weights $r_1, \ldots, r_m > 0$ such that $\|y\|_\beta = \max_{j \in \underline{m}} r_j |y_j|$ for all $y = [y_1 \ldots y_m]^T \in \mathbb{C}^m$.*

(*d*) $\alpha = \beta =: p \in [1, \infty]$.

Then $\|A\|_{\alpha,\beta}^{\mathbb{R}} = \|A\|_{\alpha,\beta}^{\mathbb{C}}$ for every $A = [a_{jk}] \in \mathbb{R}^{m \times n}$.

Proof: If $n = 1$ then A is a column vector, and we have for $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$,

$$\|A\|_{\alpha,\beta}^{\mathbb{F}} = \max_{x \in \mathbb{F} \setminus \{0\}} \frac{\|Ax\|_\beta}{\|x\|_\alpha} = \max_{x \in \mathbb{F} \setminus \{0\}} \frac{|x|\,\|A\,1\|_\beta}{|x|\,\|1\|_\alpha} = \frac{\|A\|_\beta}{\|1\|_\alpha}.$$

Condition (*b*) is a special case of condition (*c*) since every norm $\|\cdot\|_\beta$ on $\mathbb{C} = \mathbb{C}^1$ is of the form $\|y\|_\beta = r\,|y|$ with $r = \|1\|_\beta$.

We are now going to show that, if condition (*c*) or (*d*) holds, then to each $x = [x_1\ \ldots\ x_n]^T \in \mathbb{C}^n$ there is an $x' \in \mathbb{R}^n$ such that $\|Ax'\|_\beta / \|x'\|_\alpha \geq \|Ax\|_\beta / \|x\|_\alpha$. This implies $\|A\|_{\alpha,\beta}^{\mathbb{R}} = \|A\|_{\alpha,\beta}^{\mathbb{C}}$.

Assume that condition (*c*) is satisfied. Let $j \in \underline{m}$ be such that $\|Ax\|_\beta = r_j \left|\sum_{k \in \underline{n}} a_{jk} x_k\right|$. Set $x'_k := |x_k|$ if $a_{jk} \geq 0$ and $x'_k := -|x_k|$ otherwise. Then we have for the vector $x' = [x'_1\ \ldots\ x'_n]^T \in \mathbb{R}^n$ that $\|x'\|_\alpha = \|\,|x|\,\|_\alpha = \|x\|_\alpha$ and $\|Ax'\|_\beta \geq r_j \left|\sum_{j \in \underline{m}} a_{jk} x'_k\right| = r_j \sum_{k \in \underline{n}} |a_{jk} x_k| \geq \|Ax\|_\beta$. Thus $\|Ax'\|_\beta / \|x'\|_\alpha \geq \|Ax\|_\beta / \|x\|_\alpha$.

For the case (*d*) we need a lemma.

Lemma 1.6.2 *Let $p \in [1, \infty)$, $x \in \mathbb{C}^n$, $y \in \mathbb{C}^m$. Then*

(*i*) *We have*

$$\int_0^{2\pi} \|\Re(e^{i\phi} x)\|_p^p \, d\phi = \|x\|_p^p \int_0^{2\pi} |\cos \phi|^p \, d\phi.$$

(*ii*) *Suppose that $e^{i\phi} x \notin \mathbb{R}^n$ for all $\phi \in [0, 2\pi]$. Then*

$$\max_{\phi \in [0,2\pi]} \frac{\|\Re(e^{i\phi} y)\|_p}{\|\Re(e^{i\phi} x)\|_p} \geq \frac{\|y\|_p}{\|x\|_p}. \tag{1.6}$$

Proof: (i). Write the components of the vector x in the form $x_k = r_k\, e^{i\alpha_k}$, where $r_k \geq 0$, $\alpha_k \in [0, 2\pi]$. Then

$$\begin{aligned} \int_0^{2\pi} \|\Re(e^{i\phi}x)\|_p^p \, d\phi &= \int_0^{2\pi} \sum_{k=1}^n r_k^p \left|\Re(e^{i(\alpha_k+\phi)})\right|^p d\phi \\ &= \sum_{k=1}^n r_k^p \int_0^{2\pi} |\cos(\alpha_k+\phi)|^p \, d\phi \\ &= \sum_{k=1}^n r_k^p \int_0^{2\pi} |\cos\phi|^p \, d\phi \\ &= \|x\|_p^p \int_0^{2\pi} |\cos\phi|^p \, d\phi. \end{aligned}$$

(ii). If $e^{i\phi}x \notin \mathbb{R}^n$ for all $\phi \in [0, 2\pi]$ then $\Re(e^{i\phi}x) = \Im(e^{i(\phi+(\pi/2))}x) \neq 0$ for all $\phi \in [0, 2\pi]$. Hence, the quotients in (1.6) are well defined. Set $g(\phi) := \|x\|_p^p \, \|\Re(e^{i\phi}y)\|_p^p - \|y\|_p^p \, \|\Re(e^{i\phi}x)\|_p^p$. Then (i) yields that $\int_0^{2\pi} g(\phi)\, d\phi = 0$. The latter equation together with the continuity of g implies that there exists a $\phi_0 \in [0, 2\pi]$ such that $g(\phi_0) \geq 0$, whence $\|\Re(e^{i\phi_0}y)\|_p / \|\Re(e^{i\phi_0}x)\|_p \geq \|y\|_p/\|x\|_p$. □

In order to finish the proof of the proposition it remains to treat case (d) for $p \neq \infty$. Let $x \in \mathbb{C}^n \setminus \{0\}$. According to the lemma there is a $\phi_0 \in [0, 2\pi]$ such that $e^{i\phi_0}x \in \mathbb{R}^n$ or there is a $\phi_0 \in [0, 2\pi]$ such that $\Re(e^{i\phi_0}x) \neq 0$ and $\|\Re(e^{i\phi_0}Ax)\|_p / \|\Re(e^{i\phi_0}x)\|_p \geq \|Ax\|_p/\|x\|_p$. In both cases we have $\|Ax'\|_p/\|x'\|_p \geq \|Ax\|_p/\|x\|_p$, where $x' = \Re(e^{i\phi_0}x) \in \mathbb{R}^n \setminus \{0\}$. □

1.7 Semi-continuity

Let Y be a topolgical space. A function $\phi : Y \to [-\infty, \infty]$ is said to be lower (upper) semi-continuous if for all $(y_0, \rho) \in Y \times (-\infty, \infty)$ the inequality $\phi(y_0) > \rho$ ($\phi(y_0) < \rho$) implies that $\phi(y) > \rho$ ($\phi(y) < \rho$) for all y in a neighborhood V of y_0. We will frequently use the following Lemma in order to show that certain functions are well defined and semi-continuous. Throughout this work we use the conventions

$$\min \emptyset := \infty, \qquad \infty^{-1} := 0, \qquad 0^{-1} := \infty. \tag{1.7}$$

Lemma 1.7.1 *Let X be a nonempty closed subset of a finite dimensional real or complex vector space X_0, endowed with a norm $\|\cdot\|$. Let Y and Z be topological spaces and K be a closed subset of Z. Let $f : X \times Y \to Z$ be a continuous function. Then*

(i) For each $y \in Y$: either $\{\, x \in X \mid f(x,y) \in K \,\} = \emptyset$ or the minimum $\min\{\, \|x\| \mid x \in X,\ f(x,y) \in K \,\}$ is attained.

(ii) The function $\Phi_f : Y \to [0, \infty]$ defined by

$$\Phi_f(y) := \min\{\, \|x\| \mid x \in X,\ f(x,y) \in K \,\}$$

is lower semi-continuous. The function $\Psi_f : Y \to [0, \infty]$ defined by

$$\Psi_f(y) := (\, \min\{\, \|x\| \mid x \in X,\ f(x,y) \in K \,\}\,)^{-1}$$

is upper semi-continuous.

Proof: By continuity the set $f^{-1}(K)$ is closed and the balls $B(\rho) := \{\, x \in X_0 \mid \|x\| \leq \rho \,\}$ are compact. Thus the sets

$$\begin{aligned} H(\rho, y) &:= ((B(\rho) \cap X) \times \{y\}) \cap f^{-1}(K) \\ &= \{\ (x, y) \mid x \in B(\rho) \cap X,\ f(x, y) \in K\ \} \end{aligned}$$

are compact subsets of $X_0 \times Y$ for all $\rho \in [0, \infty)$, $y \in Y$.

(i). Suppose there is an $x_* \in X$ such that $f(x_*, y) \in K$. Then $H(\|x_*\|, y) \neq \emptyset$. The continuous function $g : H(\|x_*\|, y) \to \mathbb{R}$, $g(x, y) := \|x\|$, attains its minimum. This minimum coincides with the mimimum in claim (i).

(ii). It is easily seen that the following equivalence holds for $\rho \geq 0$ and $y \in Y$.

$$\Phi_f(y) > \rho \ \Leftrightarrow\ H(\rho, y) = \emptyset.$$

Let $y_0 \in Y$ and suppose that $\Phi_f(y_0) > \rho$. Then the set $B(\rho) \times \{y_0\}$ does not intersect the closed set $f^{-1}(K)$. Thus to each $x \in B(\rho) \cap X$ there are neighborhoods $U_x \subseteq X$ of x and $V_x \subseteq Y$ of y_0 such that $f^{-1}(K) \cap (U_x \times V_x) = \emptyset$. By compactness of $B(\rho) \cap X$ there are $x_1, \ldots, x_r \in B(\rho)$ such that $B(\rho) \subseteq \bigcup_{j \in \underline{r}} U_{x_j}$. Let $V := \bigcap_{j \in \underline{r}} V_{x_j}$. Then $H(\rho, y) = \emptyset$ for all $y \in V$. Thus $\Phi_f(y) > \rho$ for all $y \in V$. Thus, we have shown that the function Φ_f is lower semi-continuous. The upper semi-continuity of Ψ_f is then obvious. $\square$

Chapter 2

Definition and basic properties of spectral value sets

In the first section of this chapter we define spectral value sets for structured matrix perturbations. Sections 2 and 3 deal with the topological properties of these sets. Section 4 is about the characterization of spectral value sets via μ-functions. This characterization yields the basis for the visualization of spectral value sets. In Section 5 we discuss cases in which the μ-value of a matrix M equals its norm, $\|M\|$. Section 6 contains the first main result of this work, a theorem on the relationsship between spectral value sets for small perturbations, μ-functions and condition numbers of eigenvalues for structured perturbations.

2.1 Definition

We now introduce our main object of study, spectral value sets for structured matrix perturbations. A spectral value set of all complex numbers to which at least one eigenvalue of the nominal matrix $A \in \mathbb{C}^{n\times n}$ can be shifted by a perturbation of the form $A \rightsquigarrow A(\Delta) = A + B\Delta C$, where B, C are fixed matrices and the perturbation matrix Δ is an element of a perturbation class $\mathbf{\Delta}$. Moreover, the norm of the perturbation Δ is assumed to be bounded by a perturbation level $\rho > 0$. A more precise definition of spectral value sets will be given below. We first have to specify the types of perturbation classes under consideration.

Definition 2.1.1 *Let $l, q \in \mathbb{N}$. By $\mathcal{P}_{l,q}$ we denote the set of pairs $(\mathbf{\Delta}, \|\cdot\|)$, where*

- *$\mathbf{\Delta}$ is a nonempty closed subset of $\mathbb{C}^{l\times q}$ which is star-shaped with respect to 0, i.e. $\Delta \in \mathbf{\Delta}$ implies that $t\Delta \in \mathbf{\Delta}$ for every $t \in [0,1]$.*
- *$\|\cdot\|$ is a norm on the real vector space* $\operatorname{span}_{\mathbb{R}} \mathbf{\Delta} \subseteq \mathbb{C}^{l\times q}$.

By $\mathcal{P}^{\mathbb{C}}_{l,q}$ we denote the set of pairs $(\mathbf{\Delta}, \|\cdot\|)$, where

- *$\mathbf{\Delta}$ is a nonempty closed subset of $\mathbb{C}^{l\times q}$ which satisfies $\mathbb{C}\mathbf{\Delta} = \mathbf{\Delta}$, i.e. $\Delta \in \mathbf{\Delta}$ implies that $t\Delta \in \mathbf{\Delta}$ for every $t \in \mathbb{C}$.*
- *$\|\cdot\|$ is a norm on the complex vector space* $\operatorname{span}_{\mathbb{C}} \mathbf{\Delta}$.

The pairs $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}$ will be called perturbation structures and the pairs $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}^{\mathbb{C}}_{l,q}$ will be called complex *perturbation structures. We have that $\mathcal{P}^{\mathbb{C}}_{l,q} \subset \mathcal{P}_{l,q}$. Note that if $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}^{\mathbb{C}}_{l,q}$ then the norm satisfies $\|t\Delta\| = |t|\|\Delta\|$ for all* complex *numbers t. This is not necessarily the case if $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}$.*

There are many ways to define spectral value sets for matrices and linear operators , see e. g. [21, 51, 42, 43, 53, 91, 92]. In this work we consider spectral value sets for perturbation structures $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}$ and matrix triples $(A,B,C) \in L_{n,l,q}$, where

$$L_{n,l,q} := \mathbb{C}^{n\times n} \times \mathbb{C}^{n\times l} \times \mathbb{C}^{q\times n}.$$

Definition 2.1.2 *Let $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}$ be a perturbation structure. Then the spectral value set of the triple $(A,B,C) \in L_{n,l,q}$ with respect to the perturbation class $\mathbf{\Delta}$, the underlying norm $\|\cdot\|$ and the perturbation level $\rho \geq 0$ is the following subset of the complex plane.*

$$\begin{aligned} \sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho) &:= \bigcup_{\Delta\in\mathbf{\Delta},\ \|\Delta\|\leq\rho} \sigma(A+B\Delta C) \qquad (2.1) \\ &= \{\, s\in\mathbb{C} \mid \exists\, \Delta\in\mathbf{\Delta}:\ \|\Delta\|\leq\rho,\ \ \det(sI_n-(A+B\Delta C))=0\}. \end{aligned}$$

In words, $\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$ is the union of all spectra of matrices of the form $A+B\Delta C$, where $\Delta\in\mathbf{\Delta}$, $\|\Delta\|\leq\rho$.

For the perturbation level $\rho = 0$ we have $\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,0) = \sigma(A)$. The assumption that the perturbation class $\mathbf{\Delta}$ is star-shaped guaranties that each connected component of $\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$ contains an eigenvalue of A. This will be shown in the next section. Note that the seemingly more complicated perturbations of the form

$$A \rightsquigarrow A(\Delta_1,\ldots,\Delta_m,\delta_1,\ldots,\delta_r) := A + \sum_{j\in\underline{m}} B_j\Delta_jC_j + \sum_{j\in\underline{r}} \delta_jE_j, \qquad \delta_j\in\mathbb{F},$$

can be modelled in our framework since we have $A(\Delta_1,\ldots,\Delta_m,\delta_1,\ldots,\delta_r) = A+B\Delta C$ with

$$\begin{aligned} \Delta &= \operatorname{diag}(\Delta_1,\ldots,\Delta_m,\delta_1I_n,\ldots,\delta_rI_n), \\ B &= [B_1,\ldots,B_m,E_1,\ldots,E_r], \\ C &= [C_1^T,\ldots,C_m^T,\underbrace{I_n,\ldots,I_n}_{r \text{ times}}]^T. \end{aligned}$$

There are some important special cases of spectral value sets for which we introduce simplified notations.

- $\sigma_{\mathbb{F}}^{\|\cdot\|}(A,B,C,\rho) := \sigma_{\mathbb{F}^{l\times q}}^{\|\cdot\|}(A,B,C,\rho)$.
- $\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,\rho) := \sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,I_n,I_n,\rho)$ and $\sigma_{\mathbb{F}}^{\|\cdot\|}(A,\rho) := \sigma_{\mathbb{F}^{l\times q}}^{\|\cdot\|}(A,I_n,I_n,\rho)$.
- If $\|\cdot\| = \|\cdot\|_{\alpha,\beta}$ is an operator norm subordinate to the vector norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ then $\sigma_{\mathbf{\Delta}}^{(\alpha,\beta)}(A,B,C,\rho) := \sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$ and $\sigma_{\mathbb{F}}^{(\alpha,\beta)}(A,\rho) := \sigma_{\mathbb{F}}^{\|\cdot\|}(A,\rho)$.

  Moreover, $\sigma_{\mathbf{\Delta}}^{(\alpha)}(A,B,C,\rho) := \sigma_{\mathbf{\Delta}}^{(\alpha,\alpha)}(A,B,C,\rho)$ and $\sigma_{\mathbb{F}}^{(\alpha)}(A,\rho) := \sigma_{\mathbb{F}}^{(\alpha,\alpha)}(A,\rho)$.

Spectral value sets of the form $\sigma_{\mathbb{C}}^{(\alpha)}(A,\rho)$ are known under the name pseudospectra and have been investigated by many authors. In our terminology any set of the form $\sigma_{\mathbb{F}}^{\|\cdot\|}(A,\rho)$, $\mathbb{F}\in\{\mathbb{R},\mathbb{C}\}$ is said to be a pseudospectrum. Some examples of spectral value sets and pseudospectra are given in the Figures 2.1, 2.2 and 2.3. We will comment these figures in the next section.

On replacing in (2.1) the condition $\|\Delta\| \le \rho$ by the stronger condition $\|\Delta\| < \rho$ one obtains the sets

$$\begin{aligned}\overset{\circ}{\sigma}{}^{\|\cdot\|}_{\mathbf{\Delta}}(A,B,C,\rho) &:= \bigcup_{\Delta\in\mathbf{\Delta},\ \|\Delta\|<\rho} \sigma(A+B\Delta C) \\ &= \{\, s\in\mathbb{C} \mid \exists\,\Delta\in\mathbf{\Delta}:\ \|\Delta\|<\rho,\ \det(sI_n-(A+B\Delta C))=0\}.\end{aligned}$$

The sets $\overset{\circ}{\sigma}{}^{\|\cdot\|}_{\mathbb{F}^{l\times q}}(A,B,C,\rho)$, where $\mathbb{F}=\mathbb{R}$ or $\mathbb{C}$ and the underlying norm is an operator norm, have been studied in [30, 42, 43, 51, 53]. It will be shown in the next section that $\sigma^{\|\cdot\|}_{\mathbf{\Delta}}(A,B,C,\rho)$ is the closure of $\overset{\circ}{\sigma}{}^{\|\cdot\|}_{\mathbf{\Delta}}(A,B,C,\rho)$. Moreover, we will see that sets $\overset{\circ}{\sigma}{}^{\|\cdot\|}_{\mathbf{\Delta}}(A,B,C,\rho)$ are open if the underlying perturbation structure is complex.

A concept closely related to spectral value sets is the stability radius.

Definition 2.1.3 *Let $\mathbb{C}_g$ be a nonempty open subset of $\mathbb{C}$. A matrix $A\in\mathbb{C}^{n\times n}$ is said to be $\mathbb{C}_g$-stable if $\sigma(A)\subset\mathbb{C}_g$. The $\mathbb{C}_g$-stability radius of $(A,B,C)\in L_{n,l,q}$ with respect to $(\mathbf{\Delta},\|\cdot\|)\in\mathcal{P}_{l,q}$ is defined as follows.*

$$\begin{aligned}r^{\|\cdot\|}_{\mathbf{\Delta}}(A,B,C,\mathbb{C}_g) &= \min\{\,\|\Delta\| \mid \Delta\in\mathbf{\Delta}, A+B\Delta C \textit{ is not } \mathbb{C}_g\textit{-stable}\ \} \\ &= \min\{\,\|\Delta\| \mid \Delta\in\mathbf{\Delta},\ \sigma(A+B\Delta C)\not\subset\mathbb{C}_g\ \}\end{aligned} \tag{2.2}$$

Using Lemma (1.7) it is easily seen that, with the convention $\min\emptyset=\infty$, the minimum in (2.2) always exists. Obviously,

$$r^{\|\cdot\|}_{\mathbf{\Delta}}(A,B,C,\mathbb{C}_g) = \min\{\ \rho\ge 0 \mid\ \sigma^{\|\cdot\|}_{\mathbf{\Delta}}(A,B,C,\rho)\not\subseteq\mathbb{C}_g\ \}.$$

We will briefly mention those results on stability radii that follow as immediate consequences from our considerations concerning spectral value sets.

2.2 Topological properties

In this section we discuss basic topological properties of spectral value sets. We use the notation $D(\lambda,\epsilon) := \{\,s\in\mathbb{C} \mid\ |s-\lambda|<\epsilon\,\}$ for the open disks in $\mathbb{C}$. The proposition below describes the structure of spectral value sets $\sigma^{\|\cdot\|}_{\mathbf{\Delta}}(A,\rho)$, where $\mathbf{\Delta}$ is a 1-dimensional real or complex subspace of $\mathbb{C}^{n\times n}$.

Proposition 2.2.1 *Let $A,\Delta\in\mathbb{C}^{n\times n}$. Then the following statements hold.*

(i) There exists an integer $n'\le n$ and a finite set $S\subset\mathbb{C}$ (which may be empty) such that $\#\sigma(A+t\Delta)=n'$ for all $t\in\mathbb{C}\setminus S$ and $\#\sigma(A+t\Delta)<n'$ for all $t\in S$.

(ii) For each simply connected open set $U\subset\mathbb{C}\setminus S$ there exist analytic functions $h_1,\ldots,h_{n'}: U\to\mathbb{C}$ such that $\sigma(A+t\Delta)=\{h_1(t),\ldots,h_{n'}(t)\}$ for all $t\in U$.

(iii) Let $t_0\in S$ and let $\lambda_0\in\mathbb{C}$ be an eigenvalue of $A+t_0\Delta$ of algebraic multiplicity m. Then there are integers $m_1,\ldots,m_r\in\underline{m}$ with $\sum_{j=1}^r m_j=m$, an open neighborhood $U\subset\mathbb{C}$ of λ_0, an $\epsilon>0$, and analytic functions $h_j: D(0,\epsilon^{1/m_j})\to\mathbb{C}$ with $h_j(0)=\lambda_0$ for all $j\in\underline{r}$, such that for every $t\in D(t_0,\epsilon)$,

$$U\cap\sigma(A+t\Delta) = \bigcup_{j=1}^{r}\{\,h_j(\tau) \mid\ \tau\in\mathbb{C},\ \tau^{m_j}=t-t_0\}\,.$$

Consequently,

$$U \cap \bigcup_{t \in D(t_0,\epsilon)} \sigma(A+t\Delta) = \bigcup_{j=1}^{r} h_j\left(D(0,\epsilon^{1/m_j})\right).$$

(iv) *There exist continuous functions* $f_1,\ldots,f_{n'}:\mathbb{R}\to\mathbb{C}$ *which are analytic on* $\mathbb{R}\setminus S$*, such that* $\{f_1(t),\ldots,f_{n'}(t)\}=\sigma(A+t\Delta)$ *for all* $t\in\mathbb{R}$.

(v) *Let* V *be an open subset of* $\mathbb{C}$*. Then* $\left(\bigcup_{t\in V}\sigma(A+t\Delta)\right)\setminus\sigma(A)$ *is also an open subset of* $\mathbb{C}$.

(vi) *Either* $\sigma(A+t\Delta)=\sigma(A)$ *for all* $t\in\mathbb{C}$*, or* $\#\left(\mathbb{C}\setminus\bigcup_{t\in\mathbb{C}}\sigma(A+t\Delta)\right)\le n-1$*. If* Δ *is nonsingular then* $\bigcup_{t\in\mathbb{C}}\sigma(A+t\Delta)=\mathbb{C}$.

Proof: The statements $(i),(ii),(iii)$ are special cases of basic results in analytic perturbation theory of eigenvalues. They are obtained by applying the theory of complex algebraic curves to the polynomial

$$p(s,t) := \det(sI_n-(A+t\Delta)), \qquad s,t\in\mathbb{C}. \tag{2.3}$$

Let $q_1(s,t),\ldots,q_k(s,t)$ be the irreducible factors of $p(s,t)$. The set S in (i) is the finite set of parameters $t\in\mathbb{C}$ for which at least one of the polynomials $s\mapsto q_j(s,t), j\in\underline{k}$ has a multiple root. See Sections 3.2, 3.3 in [3] or [28] for details.

(iv). The existence of continuous functions $f_1,\ldots,f_{n'}:\mathbb{R}\to\mathbb{C}$ such that $\{f_1(t),\ldots,f_{n'}(t)\}=\sigma(A+t\Delta)$ for all $t\in\mathbb{R}$ is shown in [67, Subsection 2.5.2]. Claim (ii) implies that to each f_j and each $t_0\in\mathbb{R}\setminus S$ there is an open neighborhood $U\subset\mathbb{C}$ of t_0 and an analytic function $h_j:U\to\mathbb{C}$ such that $f_j(t)=h_j(t)$ for all $t\in U\cap\mathbb{R}$. Thus, the functions f_j are analytic on $\mathbb{R}\setminus S$.

(v). Let $t_0\in V$ and let $\lambda_0\in\sigma(A+t_0\Delta)\setminus\sigma(A)$. By (ii) and (iii) there are, an open neighborhood U of λ_0, an open neighborhood V_0 of t_0, open neighborhoods $W_1,\ldots W_r$ of $0\in\mathbb{C}$ and analytic functions $h_j:W_j\to\mathbb{C}$ with $h(0)=\lambda_0$ such that $U\cap\bigcup_{t\in V_0}\sigma(A+t\Delta)=\bigcup_{j=1}^r h_j(W_j)$. If $t_0\notin S$ then we have $r=1$. We may assume that $V_0\subset V$. By (iv) there exists a continuous function $\phi:[0,1]\to\mathbb{C}$ with $\phi(\theta)\in\sigma(A+\theta t_0\Delta)$ for all $\theta\in[0,1]$ and $\phi(1)=\lambda_0$. Since $\lambda_0\notin\sigma(A)$ there is a $\theta'\in(0,1)$ such that $\phi(\theta')\in U\setminus\{\lambda_0\}$. Hence at least one of the functions h_j is nonconstant. Since a nonconstant analytic function is an open mapping the set $U\cap\bigcup_{t\in V_0}\sigma(A+t\Delta)$ is open.

(vi). The polynomial p defined in (2.3) can be written in the form $p(s,t)=\det(\Delta)t^n+\sum_{j=1}^{n-1}a_j(s)t^j+\chi_A(s)$, where χ_A denotes the characteristic polynomial of A, and the functions $a_j(\cdot)$ are polynomials of degree $\le n-1$. Now, (vi) follows from the fact that for each fixed $s\in\mathbb{C}\setminus\sigma(A)$ the equation $p(s,t)=0$ has a solution $t\in\mathbb{C}$ if and only if $\det(\Delta)\ne0$ or $a_j(s)\ne0$ for at least one j. □

Now, we address the topological properties of general spectral value sets. To this end we need the following well known continuity property of eigenvalues [86, Theorem 1.1].

Proposition 2.2.2 *Let* $\lambda_0\in\mathbb{C}$ *be an eigenvalue of* $A_0\in\mathbb{C}^{n\times n}$ *with algebraic multiplicity* m*. Let* $\epsilon>0$ *be such that* $\sigma(A_0)\cap D(\lambda_0,\epsilon)=\{\lambda_0\}$*. Then there is a neighborhood* $U\subset\mathbb{C}^{n\times n}$ *of* A_0 *such that for all* $A\in U$ *precisely* m *eigenvalues of* A *(counting multiplicities) are contained in the open disk* $D(\lambda_0,\epsilon)$.

The proofs of all statements in the remainder of this section are based on Propositions 2.2.2 and 2.2.1.

Proposition 2.2.3 *Let* $(A,B,C)\in L_{n,l,q}$ *and* $(\mathbf{\Delta},\|\cdot\|)\in\mathcal{P}_{l,q}$. *Then*

(*i*) *The connected components of* $\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$ *are path connected and each of the connected components contains at least one eigenvalue of* A.

(*ii*) *The set of isolated points of* $\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$ *is contained in* $\sigma(A)$.

(*iii*) *Let* $\mathcal{K}$ *be a connected component of* $\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$ *which contains* m *eigenvalues of* A *counting (algebraic) multiplicities. Let* $\Delta\in\mathbf{\Delta}$ *and* $\|\Delta\|\le\rho$. *Then* $\mathcal{K}$ *contains exactly* m *eigenvalues of* $A+B\Delta C$ *counting multiplicities.*

(*iv*) *The number of connected components of* $\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$ *is a decreasing function of* ρ. *Let* $k:=\#\sigma(A)$. *Then there is a* $\rho_0>0$ *such that for all* $\rho\in[0,\rho_0]$ *the set* $\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$ *has* k *connected components, and each of the connected components contains only one eigenvalue of* A.

Proof: Let $\Delta\in\mathbf{\Delta}$ with $\|\Delta\|\le\rho$. By part (iv) of Proposition 2.2.1 there are continuous functions $f_1,\ldots,f_{n'}:\mathbb{R}\to\mathbb{C}$ such that $\{f_1(t),\ldots,f_{n'}(t)\}=\sigma(A+tB\Delta C)$ for all $t\in\mathbb{R}$. By our assumption on the perturbation class $\mathbf{\Delta}$ we have that $t\Delta\in\mathbf{\Delta}$ for all $t\in[0,1]$. Thus $f_j(t)\in\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$ for all $t\in[0,1]$ and all $j\le n'$. Therefore $(i),(ii)$ and (iii) hold. The first claim in (iv) is a consequence of the obvious implication

$$0\le\rho_1\le\rho_2 \quad\Rightarrow\quad \sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho_1)\subseteq\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho_2).$$

It remains to show the second claim in (iv). Let $\lambda_1,\ldots,\lambda_k$ be the eigenvalues of A. Let $\epsilon>0$ be such that the disks $D(\lambda_j,\epsilon)$, $j\in\underline{k}$, are pairwise disjoint. By Proposition 2.2.2 there is a neighborhood $U\subset\mathbb{C}^{n\times n}$ of A such that $\sigma(M)\subset\bigcup_{j\in\underline{k}}D(\lambda_j,\epsilon)$ for all $M\in U$. Choose $\rho_0>0$ such that $A+B\Delta C\in U$ for all $\Delta\in\mathbb{C}^{l\times q}$ with $\|\Delta\|\le\rho_0$. Then for all $\rho\in[0,\rho_0]$, $\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)=\bigcup_{j\in\underline{k}}\mathcal{K}_j(\rho)$, where $\mathcal{K}_j(\rho):=\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)\cap D(\lambda_j,\epsilon)$. Each $\mathcal{K}_j(\rho)$ contains only one eigenvalue of $\sigma(A)$ and is separated from the sets $\mathcal{K}_i(\rho)$, $i\ne j$. From (i) it follows that $\mathcal{K}_j(\rho)$ is a connected component of $\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$. □

Proposition 2.2.4 *Let* $(A,B,C)\in L_{n,l,q}$, $(\mathbf{\Delta},\|\cdot\|)\in\mathcal{P}_{l,q}$ *and* $\rho\ge0$.

(*i*) *The set* $\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$ *is compact.*

(*ii*) *The set* $\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$ *is the closure of* $\mathring{\sigma}_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$.

Proof: (i). Let $(s_k)_{k\in\mathbb{N}}$ be a sequence in $\mathbb{C}$ with $s_k\in\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$ for all $k\in\mathbb{N}$. Then $s_k\in\sigma(A+B\Delta_kC)$ with $\Delta_k\in\mathcal{B}(\rho):=\{\,\Delta\in\mathbf{\Delta}\mid\|\Delta\|\le\rho\,\}$. By compactness of $\mathcal{B}(\rho)$ there is a subsequence $(\Delta_{k_j})_{j\in\mathbb{N}}$ and a $\Delta_0\in\mathcal{B}(\rho)$ such that $\lim_{j\to\infty}\Delta_{k_j}=\Delta_0$. From the continuity of eigenvalues (Proposition 2.2.2) it follows that to each $\epsilon>0$ there is a j_0 such that $s_{k_j}\in\bigcup_{\lambda\in\sigma(A+B\Delta_0C)}D(\lambda,\epsilon)$ for all $j\ge j_0$. Thus, there is a converging subsequence $(s_{k_{j_i}})_{i\in\mathbb{N}}$ such that $\lim_{i\to\infty}s_{k_{j_i}}\in\sigma(A+B\Delta_0C)\subseteq\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$.

(ii). Let $s\in\sigma(A+B\Delta C)$ with $\Delta\in\mathbf{\Delta}$, $\|\Delta\|=\rho$. By claim (iv) of Proposition 2.2.1 there exists a continuous path $f:[0,1]\to\mathbb{C}$ such that $f(t)\in\sigma(A+B(t\Delta)C)$ for all $t\in[0,1]$ and $f(1)=s$. Since $t\Delta\in\mathbf{\Delta}$ and $\|t\Delta\|<\rho$ for all $t\in[0,1)$ we have $f(t)\in\mathring{\sigma}_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$ for all $t\in[0,1)$. Thus

s is an accumulation point of $\overset{\circ}{\sigma}{}^{\|\cdot\|}_{\mathbf{\Delta}}(A,B,C,\rho)$. Hence, the closure of $\overset{\circ}{\sigma}{}^{\|\cdot\|}_{\mathbf{\Delta}}(A,B,C,\rho)$ contains the set $\sigma^{\|\cdot\|}_{\mathbf{\Delta}}(A,B,C,\rho)$. Since $\sigma^{\|\cdot\|}_{\mathbf{\Delta}}(A,B,C,\rho)$ is closed the proof is complete. □

The spectral value sets with respect to complex perturbation structures $(\mathbf{\Delta},\|\cdot\|)\in\mathcal{P}^{\mathbb{C}}_{l,q}$ have some nice additional properties.

Proposition 2.2.5 *Let $(A,B,C)\in L_{n,l,q}$ and $(\mathbf{\Delta},\|\cdot\|)\in\mathcal{P}^{\mathbb{C}}_{l,q}$. Then*

(i) The set $\overset{\circ}{\sigma}{}^{\|\cdot\|}_{\mathbf{\Delta}}(A,B,C,\rho)\setminus\sigma(A)$ is an open subset of $\mathbb{C}$ for all $\rho\geq 0$.

(ii) Let $s\in\mathbb{C}\setminus\sigma(A)$ be a point on the boundary of $\sigma^{\|\cdot\|}_{\mathbf{\Delta}}(A,B,C,\rho)$, $\rho>0$. Then

$$\min\{\,\|\Delta\|\mid \Delta\in\mathbf{\Delta},\ s\in\sigma(A+B\Delta C)\,\}=\rho. \tag{2.4}$$

(iii) Either $\sigma^{\|\cdot\|}_{\mathbf{\Delta}}(A,B,C,\rho)=\sigma(A)$ for all $\rho\geq 0$ or $\#\left(\mathbb{C}\setminus\bigcup_{\rho\geq 0}\sigma^{\|\cdot\|}_{\mathbf{\Delta}}(A,B,C,\rho)\right)\leq n-1$. If the set $\{B\Delta C\mid \Delta\in\mathbf{\Delta}\,\}$ contains a nonsingular matrix then $\bigcup_{\rho\geq 0}\sigma^{\|\cdot\|}_{\mathbf{\Delta}}(A,B,C,\rho)=\mathbb{C}$.

Proof: (i). Let $s\in\overset{\circ}{\sigma}{}^{\|\cdot\|}_{\mathbf{\Delta}}(A,B,C,\rho)\setminus\sigma(A)$. Then there is a $\Delta\in\mathbf{\Delta}$ such that $0<\|\Delta\|<\rho$ and $s\in\sigma(A+B\Delta C)$. Let $V:=\{\,t\in\mathbb{C}\mid 0<|t|<\rho\|\Delta\|^{-1}\,\}$ and $W:=\bigcup_{t\in V}\sigma(A+tB\Delta C)\setminus\sigma(A)$. Then $t\Delta\in\mathbf{\Delta}$ and $\|t\Delta\|<\rho$ for all $t\in V$. Moreover, $1\in V$. Therefore, $s\in W\subseteq\overset{\circ}{\sigma}{}^{\|\cdot\|}_{\mathbf{\Delta}}(A,B,C,\rho)$. By claim (v) of Proposition 2.2.1 the set W is open. Thus $\overset{\circ}{\sigma}{}^{\|\cdot\|}_{\mathbf{\Delta}}(A,B,C,\rho)\setminus\sigma(A)$ is open. (ii) is an immediate consequence of (i). Claim (iii) follows from (i) and and claim (vi) of Proposition 2.2.1. □

Some of the statements of this section are illustrated in Figures 2.1, 2.2 and 2.3. Figure 2.1 shows spectral value sets to the nominal matrix

$$A=\begin{bmatrix}2&-3&1&1\\3&2&1&1\\0&0&3&1\\0&0&0&4\end{bmatrix}. \tag{2.5}$$

The crosses mark the eigenvalues of A. In the first row of the figure one sees the real spectral spectral value sets $\sigma^{(2)}_{\mathbb{R}}(A,1)$, $\sigma^{(2)}_{\mathbb{R}}(A,B_1,C_1,1)$ and $\sigma^{(2)}_{\mathbb{R}}(A,B_2,C_2,1)$, where

$$B_1=\begin{bmatrix}0&0&0&0\\0&0&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix},\qquad C_1=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix},\qquad B_2=\begin{bmatrix}2\\0\\0\\0\end{bmatrix},\qquad C_2=\begin{bmatrix}1&0&0&0\end{bmatrix}.$$

Note that these sets are not closures of open sets. The set $\sigma^{(2)}_{\mathbb{R}}(A,B_2,C_2,1)$ has no interior points at all. The second row of the figure shows the sets $\sigma^{(2)}_{\mathbb{C}}(A,1)$, $\sigma^{(2)}_{\mathbb{C}}(A,B_1,C_1,1)$ and $\sigma^{(2)}_{\mathbb{C}}(A,B_2,C_2,1)$. In the third row one sees the sets $\sigma^{(1)}_{\mathbb{C}}(A,1)$, $\sigma^{(1)}_{\mathbb{C}}(A,B_1,C_1,1)$ and $\sigma^{(1)}_{\mathbb{C}}(A,B_2,C_2,1)$. The sets in the third column have isolated points.

Figure 2.2 shows the pseudospectra $\sigma_{\mathbb{F}}^{(2)}(A_0, \rho)$, where

$$A_0 \quad = \quad \begin{bmatrix} -7 & 5 & 0 & 0 \\ 0 & 0 & 8 & 2 \\ 0 & 0 & 6 & -2 \\ 0 & 0 & 8 & 6 \end{bmatrix}. \tag{2.6}$$

The crosses in the figure mark the eigenvalues of A_0. Note that the sets $\overset{\circ}{\sigma}{}^{(2)}_{\mathbb{R}}(A_0, \rho) = \overset{\circ}{\sigma}{}^{(2)}_{\mathbb{R}^{4\times 4}}(A_0, I_4, I_4, \rho)$ are not open. They consist of an open set and one ore two open intervals on the real axis. Figure 2.2 also shows that the connected components of spectral value sets are not nessarily simply connected. Another example for the latter fact is given in Figure 2.3 which displays pseudospectra of the nilpotent matrix

$$N = \begin{bmatrix} 0 & 1 & -6 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}. \tag{2.7}$$

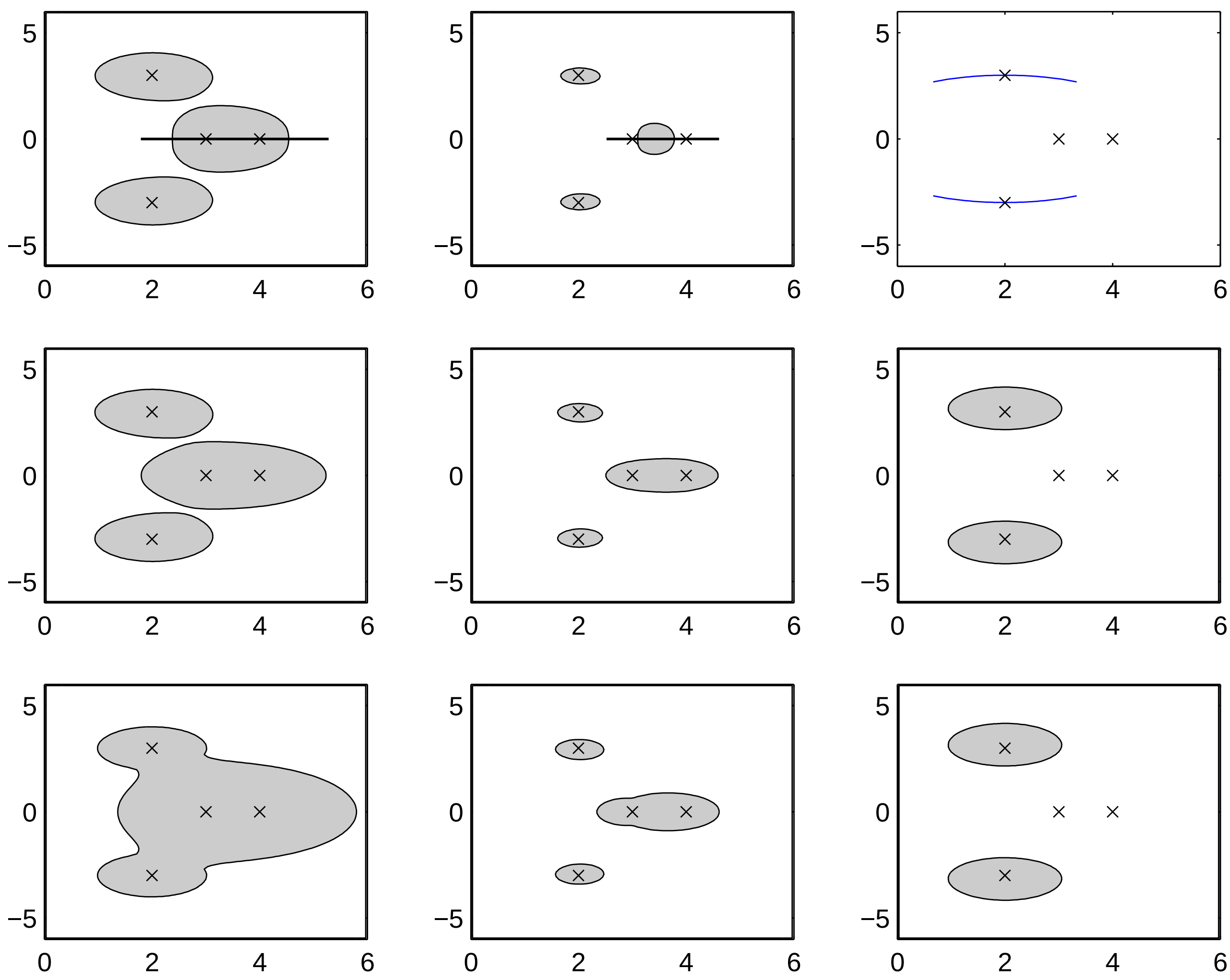


Figure 2.1: Various spectral value sets of the matrix A defined in (2.5).

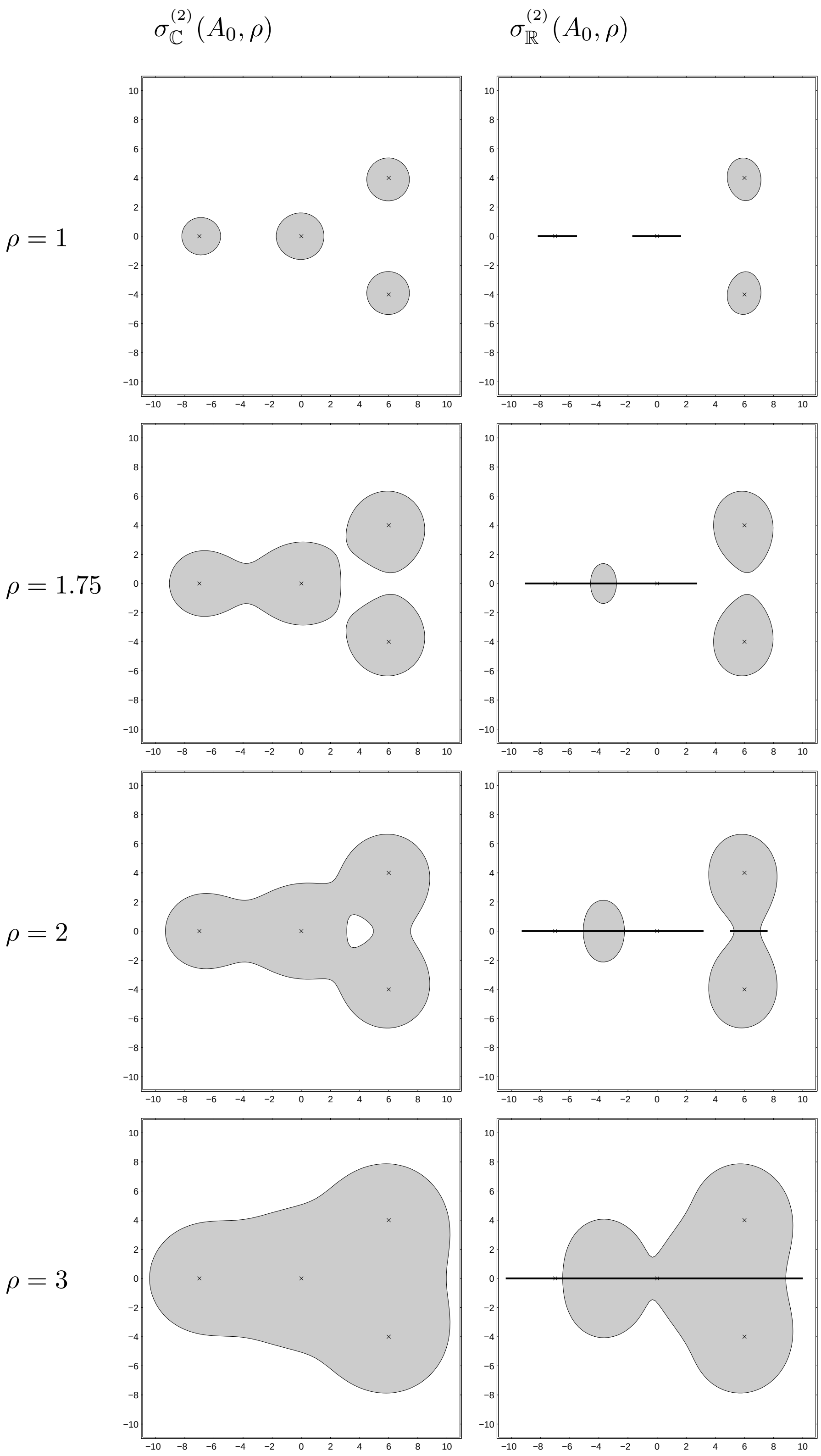


Figure 2.2: Pseudospectra of the matrix $A \in \mathbb{R}^{4\times 4}$ defined in (2.6).

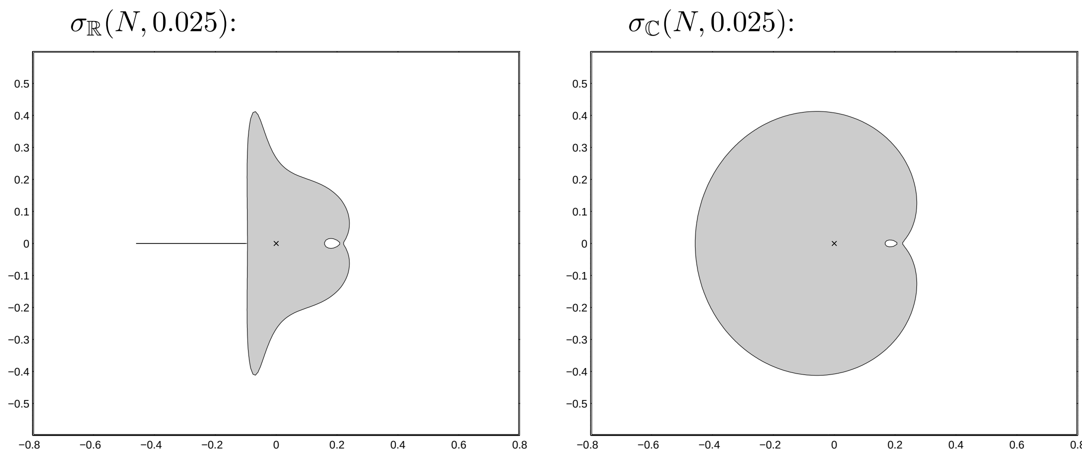


Figure 2.3: Pseudospectra of the nilpotent matrix $N \in \mathbb{R}^{3\times 3}$ defined in (2.7).

2.3 Continuous dependence on the parameters

The aim of this section is to show that spectral value sets depend continuously on the parameters A, B, C, ρ, on the perturbation class $\mathbf{\Delta}$ and on the norm $\|\cdot\|$. Though the proof is not difficult, it is a bit technical. We need some preliminary topological results.

For a metric space (Z, d) we denote by $\mathcal{C}(Z)$ the set of all nonempty compact subsets of Z. The map

$$d_H : \mathcal{C}(Z) \times \mathcal{C}(Z) \to [0, \infty) \qquad d_H(M, N) := \max \left\{ \max_{z \in M} \min_{y \in N} d(z, y)\,,\ \max_{z \in N} \min_{y \in M} d(z, y) \right\}$$

is a metric on $\mathcal{C}(Z)$. This metric is called the Hausdorff metric induced by the metric d. The lemma below follows immediately from the definition of d_H.

Lemma 2.3.1 *Le $\epsilon > 0$ and $C_1, C_2 \in \mathcal{C}(Z)$. Then the statements below are equivalent.*

(i) $d_H(C_1, C_2) < \epsilon$.

(ii) To each $z_1 \in C_1$ there is a $z_2 \in C_2$ such that $d(z_1, z_2) < \epsilon$; and to each $z_2 \in C_2$ there is a $z_1 \in C_1$ such that $d(z_1, z_2) < \epsilon$.

By combining Lemma 2.3.1 and Proposition 2.2.2 one obtains the following well known result.

Proposition 2.3.2 *Let d be a metric on $\mathbb{C}$ which induces the standard topology. Then the map*

$$\sigma : \mathbb{C}^{n\times n} \to \mathcal{C}(\mathbb{C}), \qquad A \mapsto \sigma(A),$$

which sends the matrix A to its spectrum, is continuous.

We are now going to derive some general results about continuous maps between metric spaces.

Proposition 2.3.3 *Let X be a topological space and let Y, Z be metric spaces. Let $f : X \times Y \to Z$ be a continuous map. Then the map*

$$\Phi_f : X \times \mathcal{C}(Y) \to \mathcal{C}(Z), \qquad \Phi_f(x, C) := \{\ f(x, y) \ |\ y \in C\ \}$$

is also continuous.

Proof: We denote both the metric on Y and the metric on Z by he same letter d. For $y \in Y$ and $\rho > 0$ we set $B(y,\rho) := \{ \tilde{y} \in Y \mid d(\tilde{y}, y) < \rho \}$.

Let $\epsilon > 0$, $x_0 \in X$, $C_0 \in \mathcal{C}(Y)$. By continuity of f there exist to each $y \in C_0$ a neighborhood U_y of x_0 and a $\delta_y > 0$ such that

$$d(f(x,\eta), f(x_0, y)) < \epsilon \quad \text{for all } (x,\eta) \in U_y \times B(y, \delta_y). \tag{2.8}$$

Since C_0 is compact there are $y_1, \ldots, y_r \in C_0$ such that

$$C_0 \subseteq \bigcup_{k=1}^{r} B\left(y_k, \frac{1}{2}\delta_{y_k}\right). \tag{2.9}$$

Set $\delta := \min_{k \in \underline{r}} \delta_{y_k}$ and $U := \bigcap_{k \in \underline{r}} U_{y_k}$, and consider a pair $(x, C) \in X \times \mathcal{C}(Y)$ with $x \in U$ and $d_H(C, C_0) < \frac{1}{2}\delta$. We show

(a) to each $\eta \in C$ there is a $k \le r$ such that $d(f(x,\eta), f(x_0, y_k)) < \epsilon$;

(b) to each $y \in C_0$ there is a $\eta \in C$ such that $d(f(x,\eta), f(x_0, y)) < 2\epsilon$.

By Lemma 2.3.1 the statements (a) and (b) together imply that $d_H(\Phi_f(x,C), \Phi_f(x_0, C_0)) < 2\epsilon$. Then the proof is complete.

(a). Since $d_H(C, C_0) < \frac{1}{2}\delta$ there exists to each $\eta \in C$ a $y \in C_0$ with $d(\eta, y) < \frac{1}{2}\delta$, and by (2.9) there is a y_k with $d(y, y_k) < \frac{1}{2}\delta_{y_k}$. Hence $d(\eta, y_k) < \delta_{y_k}$ for some $k \le r$. Now (2.8) yields that $d(f(x,\eta), f(x_0, y_k)) < \epsilon$.

(b). By (2.9) there exists to each $y \in C_0$ a y_k whith $d(y, y_k) < \frac{1}{2}\delta_{y_k}$ and hence by (2.8),

$$d(f(x,y), f(x_0, y_k)) < \epsilon. \tag{2.10}$$

Since $d_H(C, C_0) < \frac{1}{2}\delta$ we have $d(y_k, \eta) < \frac{1}{2}\delta$ for some $\eta \in C$. Again, (2.8) yields

$$d(f(x,\eta), f(x_0, y_k)) < \epsilon. \tag{2.11}$$

(2.10) and (2.11) imply $d(f(x,\eta), f(x_0, y)) < 2\epsilon$. □

In the next proposition we consider the metric space $(\mathcal{C}(\mathcal{C}(Z)), (d_H)_H)$. The elements of $\mathcal{C}(\mathcal{C}(Z))$ are compact sets of compact subsets of the metric space Z. The metric $(d_H)_H$ is the Hausdorff metric induced by the Hausdorff metric d_H on $\mathcal{C}(Z)$. We use the notation

$$\cup\mathcal{K} := \bigcup_{C \in \mathcal{K}} C, \quad \text{where} \quad \mathcal{K} \in \mathcal{C}(\mathcal{C}(Z)).$$

Proposition 2.3.4 *Let (Z,d) be a metric space. Then the following holds.*

(i) *If $\mathcal{K} \in \mathcal{C}(\mathcal{C}(Z))$ then $\cup\mathcal{K} \in \mathcal{C}(Z)$.*

(ii) *For all $\mathcal{K}_1, \mathcal{K}_2 \in \mathcal{C}(\mathcal{C}(Z))$, $d_H(\cup\mathcal{K}_1, \cup\mathcal{K}_2) \le (d_H)_H(\mathcal{K}_1, \mathcal{K}_2)$. In other words, the map*

$$\cup : \mathcal{C}(\mathcal{C}(Z)) \to \mathcal{C}(Z) \qquad \mathcal{K} \mapsto \cup\mathcal{K}$$

is a contraction. In particular, this map is continuous.

Proof: (i). We show that each sequence in $\cup\mathcal{K}$ has a convergent subsequence with limit in $\cup\mathcal{K}$. If $(z_n)_{n\in\mathbb{N}}$ is a sequence in $\cup\mathcal{K} = \bigcup_{C\in\mathcal{K}} C$ then there is a sequence $(C_n)_{n\in\mathbb{N}}$ in $\mathcal{K}$ with $z_n \in C_n$ for all $n\in\mathbb{N}$. Since $\mathcal{K}$ is a compact metric space there is a subsequence $(C_{n_j})_{j\in\mathbb{N}}$ with limit $\tilde{C}\in\mathcal{K}$ and by definition of the Hausdorff-metric there is a sequence $(\tilde{z}_{n_j})_{j\in\mathbb{N}}$ in $\tilde{C}$ with $\lim_{j\to\infty} d(z_{n_j}, \tilde{z}_{n_j}) = 0$. The sequence $(\tilde{z}_{n_j})_{j\in\mathbb{N}}$ has a convergent subsequence $(\tilde{z}_{n_{j_k}})_{k\in\mathbb{N}}$ with limit $\tilde{z}\in\tilde{C}$. It follows that $\lim_{k\to\infty} z_{n_{j_k}} = \tilde{z}\in\cup\mathcal{K}$.

(ii). Let $\mathcal{K}_1, \mathcal{K}_2 \in \mathcal{C}(\mathcal{C}(Z))$ and let $z_1\in\cup\mathcal{K}_1$. Then there is a $C_1\in\mathcal{C}(Z)$ such that $z_1\in C_1\in\mathcal{K}_1$. By definition of the metric $(d_H)_H$ there is a $C_2\in\mathcal{K}_2$ with $d_H(C_1, C_2) \le (d_H)_H(\mathcal{K}_1,\mathcal{K}_2)$. The latter implies that $d(z_1, z_2) \le (d_H)_H(\mathcal{K}_1,\mathcal{K}_2)$ for some $z_2\in C_2 \subseteq \cup\mathcal{K}_2$. By symmetry we also have that to each $z_2\in\cup\mathcal{K}_2$ there is a $z_1\in\cup\mathcal{K}_1$ such that $d(z_1,z_2) \le (d_H)_H(\mathcal{K}_1,\mathcal{K}_2)$. Thus $d_H(\cup\mathcal{K}_1, \cup\mathcal{K}_2) \le (d_H)_H(\mathcal{K}_1,\mathcal{K}_2)$. □

Corollary 2.3.5 *Let X be a topological space and let Y,Z be metric spaces. Let $F : X\times Y \to \mathcal{C}(Z)$ be a continuous map. Then the map*

$$\Psi_F : X\times\mathcal{C}(Y)\to\mathcal{C}(Z) \qquad \Psi_F(x, C) := \bigcup_{y\in C} F(x,y)$$

is also continuous.

Proof: According to Proposition 2.3.3 the continuity of $F : X\times Y \to \mathcal{C}(Z)$ implies the continuity of the map

$$\Phi_F : X\times\mathcal{C}(Y)\to\mathcal{C}(\mathcal{C}(Z)), \qquad \Phi_F(x,C) := \{\ F(x,y)\ \mid\ y\in C\ \}.$$

We have that $\Psi_F = \cup\circ\Phi_F$, and by Proposition 2.3.4 the map $\cup : \mathcal{C}(\mathcal{C}(Z)) \to \mathcal{C}(Z)$ is continuous. □

We are now in the position to prove the main result of this section.

Theorem 2.3.6 *Let X be a topological space and let Y be a metric space. Let $M : X\times Y\to\mathbb{C}^{n\times n}$ be a continuous matrix valued function. Then the function*

$$\psi : X\times\mathcal{C}(Y)\to\mathcal{C}(\mathbb{C}) \qquad \psi(x, C) := \bigcup_{y\in C}\sigma(M(x,y))$$

is continuous.

Proof: By Proposition 2.3.2 the continuity of the function $(x,y)\mapsto M(x,y)$ implies the continuity of the map

$$F : X\times Y\to\mathcal{C}(\mathbb{C}), \qquad F(x,y) := \sigma(M(x,y)).$$

Thus Corollary 2.3.5 yields the result. □

We now apply Theorem 2.3.6 to show the continuous dependence of spectral value sets on the parameters: Let $\boldsymbol{B}_{l,q} := \{\, \mathcal{B}^{\|\cdot\|}_{\boldsymbol{\Delta}}(\rho) \mid \rho\ge 0,\ (\boldsymbol{\Delta}, \|\cdot\|)\in\mathcal{P}_{l,q}\,\}$, where $\mathcal{B}^{\|\cdot\|}_{\boldsymbol{\Delta}}(\rho) := \{\ \Delta\in\boldsymbol{\Delta}\ \mid\ \|\Delta\|\le\rho\ \}$. Then $\boldsymbol{B}_{l,q}\subset\mathcal{C}(\mathbb{C}^{l\times q})$, and the spectral value sets can be written in the form

$$\sigma^{\|\cdot\|}_{\boldsymbol{\Delta}}(A,B,C,\rho) = \bigcup_{\Delta\in\mathcal{B}^{\|\cdot\|}_{\boldsymbol{\Delta}}(\rho)} \sigma(A + B\Delta C).$$

Furthermore, the map $((A,B,C),\Delta) \mapsto A + B\Delta C$ is continuous. Hence, we have the following corollary to Theorem 2.3.6.

Theorem 2.3.7 *The map*

$$\psi_{n,l,q} : L_{n,l,q} \times \boldsymbol{B}_{l,q} \to \mathcal{C}(\mathbb{C}), \qquad (A,B,C,\mathcal{B}_{\boldsymbol{\Delta}}^{\|\cdot\|}(\rho)) \longmapsto \sigma_{\boldsymbol{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$$

is continuous.

Using the fact that $[0,1]\boldsymbol{\Delta} = \boldsymbol{\Delta}$ it is easily verified that for each $(\boldsymbol{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}$ the function $\rho \mapsto \mathcal{B}_{\boldsymbol{\Delta}}^{\|\cdot\|}(\rho) \in \mathcal{C}(\mathbb{C})$ is continuous. Thus Theorem 2.3.7 yields

Corollary 2.3.8 *For each* $(\boldsymbol{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}$ *the map*

$$\psi_{\boldsymbol{\Delta},n,l,q}^{\|\cdot\|} : L_{n,l,q} \times [0,\infty) \to \mathcal{C}(\mathbb{C}) \qquad (A,B,C,\rho) \longmapsto \sigma_{\boldsymbol{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$$

is continuous

If the perturbation class $\boldsymbol{\Delta}$ satisfies $[0,\infty)\boldsymbol{\Delta} = \boldsymbol{\Delta}$ then a direct proof of Corollary 2.3.8 can be given by using the following well known fact.

Proposition 2.3.9 *Let X and Y be topological spaces where Y is compact. If $f : X \times Y \to \mathbb{R}$ is continuous then the map*

$$\Phi_f : X \to \mathbb{R}, \qquad \Phi_f(x) := \max_{y \in Y} f(x,y)$$

is also continuous.

Proof of Corollary 2.3.8 for the case that $[0,\infty)\boldsymbol{\Delta} = \boldsymbol{\Delta}$:
First note that the relation $[0,\infty)\boldsymbol{\Delta} = \boldsymbol{\Delta}$ implies that $\mathcal{B}_{\boldsymbol{\Delta}}^{\|\cdot\|}(\rho) = \rho\, \mathcal{B}_{\boldsymbol{\Delta}}^{\|\cdot\|}(1)$ for all $\rho \geq 0$. Therefore,

$$\sigma_{\boldsymbol{\Delta}}^{\|\cdot\|}(A,B,C,\rho) = \bigcup_{\Delta \in \mathcal{B}_{\boldsymbol{\Delta}}^{\|\cdot\|}(1)} \sigma(A + \rho B \Delta C) \qquad \text{for all } (A,B,C,\rho) \in L_{n,l,q} \times [0,\infty).$$

Now choose $(A_0, B_0, C_0,\ \rho_0) \in L_{n,l,q} \times [0,\infty)$ and define

$$f : L_{n,l,q} \times [0,\infty) \times \mathcal{B}_{\boldsymbol{\Delta}}^{\|\cdot\|}(1) \to \mathbb{R} \qquad f(A,B,C,\ \rho;\Delta) := d_H(\,\sigma(A + \rho B\Delta C)\ ,\ \sigma(A_0 + \rho_0 B_0 \Delta C_0)\,).$$

Since f is continuous, the map

$$\Phi_f : L_{n,l,q} \times [0,\infty) \to \mathbb{R} \qquad \Phi_f(A,B,C;\rho) = \max_{\Delta \in \mathcal{B}_{\boldsymbol{\Delta}}^{\|\cdot\|}(1)} f(A,D,E,\ \rho;\Delta)$$

is also continuous by Proposition 2.3.9. We show that $\Phi_f(A,B,C;\rho)$ is an upper bound for the Hausdorff-distance of $\sigma_{\boldsymbol{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$ and $\sigma_{\boldsymbol{\Delta}}^{\|\cdot\|}(A_0,B_0,C_0,\rho_0)$:

$$\begin{aligned}
&d_H(\ \sigma_{\boldsymbol{\Delta}}^{\|\cdot\|}(A,B,C,\rho)\ ,\ \sigma_{\boldsymbol{\Delta}}^{\|\cdot\|}(A_0,B_0,C_0,\rho_0)\ ) \\
&\qquad = d_H(\ \bigcup_{\Delta \in \mathcal{B}_{\boldsymbol{\Delta}}^{\|\cdot\|}(1)} \sigma(A + \rho\, B\Delta C)\ ,\ \bigcup_{\Delta \in \mathcal{B}_{\boldsymbol{\Delta}}^{\|\cdot\|}(1)} \sigma(A_0 + \rho_0\, B_0 \Delta C_0)\ ) \\
&\qquad \leq \max_{\Delta \in \mathcal{B}_{\boldsymbol{\Delta}}^{\|\cdot\|}(1)} d_H(\ \sigma(A + \rho\, B\Delta C)\ ,\ \sigma(A_0 + \rho_0\, B_0 \Delta C_0)\ ) \\
&\qquad = \Phi_f(A,B,C;\rho)
\end{aligned}$$

By continuity we have $\Phi_f(A,B,C,\rho) \to 0$ as $(A,B,C,\rho) \to (A_0,D_0,E_0,\rho_0)$. $\square$

2.4 Characterization via μ-functions

In this section we show that the spectral value sets $\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$, $\rho>0$, are superlevel-sets of an upper semi-continuous function. This fact can be used to visualize spectral value sets provided that the function in question can be effectively calculated. For the computation of spectral value sets in the case $\mathbf{\Delta}=\mathbb{C}^{l\times q}$ see [5, 30, 31, 91, 26] and the references therein. The function just mentioned is the composition of the transfer function $G(s)=C(sI_n-A)^{-1}B$ and the μ-function associated with the perturbation structure $(\mathbf{\Delta},\|\cdot\|)\in\mathcal{P}_{l,q}$. The μ-functions are defined as follows [55, 53, 74, 75].

Definition 2.4.1 *Let $(\mathbf{\Delta},\|\cdot\|)\in\mathcal{P}_{l,q}$. Then we define for $M\in\mathbb{C}^{q\times l}$,*

$$\begin{aligned}\mu_{\mathbf{\Delta}}^{\|\cdot\|}(M) &:= (\min\{\|\Delta\| \mid \Delta\in\mathbf{\Delta},\ \det(I_l-\Delta M)=0\})^{-1} \\ &= (\min\{\|\Delta\| \mid \Delta\in\mathbf{\Delta},\ 1\in\sigma(\Delta M)\})^{-1}.\end{aligned} \tag{2.12}$$

Lemma 1.7.1 yields that, with the convention $\min\emptyset=\infty$, the minimum in (2.12) always exists, and the function $\mu_{\mathbf{\Delta}}^{\|\cdot\|}:\mathbb{C}^{q\times l}\to[0,\infty)$ is upper semi-continuous. Observe that $\mu_{\mathbf{\Delta}}^{\|\cdot\|}(M)=0$ if and only if there is no matrix $\Delta\in\mathbf{\Delta}$ such that $1\in\sigma(\Delta M)$. We give now a characterization of $\mu_{\mathbf{\Delta}}^{\|\cdot\|}(\cdot)$ by the spectral radius (if $\mathbb{C}\mathbf{\Delta}=\mathbf{\Delta}$) and the real spectral radius (if $\mathbb{R}\mathbf{\Delta}=\mathbf{\Delta}$). The real spectral radius of $A\in\mathbb{C}^{n\times n}$ is

$$\varrho_{\mathbb{R}}(A):=\begin{cases}0 & \text{if } \sigma(A)\cap\mathbb{R}=\emptyset,\\ \max\{\,|\lambda| \mid \lambda\in\sigma(A)\cap\mathbb{R}\,\} & \text{otherwise.}\end{cases}$$

Lemma 2.4.2 *The function $\varrho_{\mathbb{R}}:\mathbb{C}^{n\times n}\to[0,\infty)$ is upper semi-continuous.*

Proof: Let $A_0\in\mathbb{C}^{n\times n}$ and let $\rho>0$ such that $\varrho_{\mathbb{R}}(A_0)<\rho$. Then all real eigenvalues of A_0 are contained in the interval $(-\rho,\rho)$. By continuity of the spectrum (Proposition 2.2.2) the same statement holds for all matrices A in a neighborhood $U\subset\mathbb{C}^{n\times n}$ of A_0. □

Proposition 2.4.3 *Let $M\in\mathbb{C}^{q\times l}$.*

(i) Let $(\mathbf{\Delta},\|\cdot\|)\in\mathcal{P}_{l,q}$ and assume that $\mathbb{R}\mathbf{\Delta}=\mathbf{\Delta}$. Then

$$\mu_{\mathbf{\Delta}}^{\|\cdot\|}(M)=\max\{\,\varrho_{\mathbb{R}}(\Delta M) \mid \Delta\in\mathbf{\Delta},\|\Delta\|=1\,\}.$$

(ii) If $(\mathbf{\Delta},\|\cdot\|)\in\mathcal{P}_{l,q}^{\mathbb{C}}$ then $\mu_{\mathbf{\Delta}}^{\|\cdot\|}(M)=\max\{\,\varrho(\Delta M) \mid \Delta\in\mathbf{\Delta},\|\Delta\|=1\,\}$.

Proof: We only show (i). The proof of (ii) is analogous. It can be found in [74].

Since $\varrho_{\mathbb{R}}(\cdot)$ is upper semi-continuous and the set $\{\,\Delta\in\mathbf{\Delta} \mid \|\Delta\|=1\}$ is compact, the maximum $c:=\max_{\Delta\in\mathbf{\Delta},\|\Delta\|=1}\varrho_{\mathbb{R}}(\Delta M)$ exists. Let

$$\hat{\mathbf{\Delta}}:=\{\,\Delta\in\mathbf{\Delta} \mid \sigma(\Delta M)\cap\mathbb{R}=\{0\} \text{ or } \sigma(\Delta M)\cap\mathbb{R}=\emptyset\}.$$

If $\hat{\mathbf{\Delta}}=\mathbf{\Delta}$ then $c=0=\mu_{\mathbf{\Delta}}^{\|\cdot\|}(M)$. Suppose that $\hat{\mathbf{\Delta}}\neq\mathbf{\Delta}$. Then $c>0$. Thus there is a $\Delta\in\mathbf{\Delta}$ and an $s\in\mathbb{R}$ such that $\|\Delta\|=1$, $s\in\sigma(\Delta M)$ and $|s|=\varrho_{\mathbb{R}}(\Delta M)=c$. Let $\Delta_0:=s^{-1}\Delta$. Then $\Delta_0\in\mathbf{\Delta}$, $1\in\sigma(\Delta_0 M)$ and $\|\Delta_0\|=|s|^{-1}$. Thus $\mu_{\mathbf{\Delta}}^{\|\cdot\|}(M)\geq|s|=c$. In particular, we have $\mu_{\mathbf{\Delta}}^{\|\cdot\|}(M)>0$. Hence there is a $\Delta\in\mathbf{\Delta}$ such that $\|\Delta\|=\mu_{\mathbf{\Delta}}^{\|\cdot\|}(M)^{-1}$ and $1\in\sigma(\Delta M)$. Let $\Delta_1:=\|\Delta\|^{-1}\Delta$. Then

$\Delta_1 \in \mathbf{\Delta}$, $\|\Delta_1\| = 1$ and $\|\Delta\|^{-1} \in \sigma(\Delta_1 M)$. The latter implies that $\|\Delta\|^{-1} \leq \varrho_{\mathbb{R}}(\Delta_1 M)$. Hence, $\mu_{\mathbf{\Delta}}^{\|\cdot\|}(M) \leq c$. $\square$

The set $\{\Delta \in \mathbf{\Delta} \mid \|\Delta\| \leq 1\}$ is compact and the spectral radius $\varrho : \mathbb{C}^{l\times l} \to [0,\infty)$ is a continuous function. Therefore, the mapping $M \mapsto \max_{\Delta\in\mathbf{\Delta},\|\Delta\|=1} \varrho(\Delta M)$, $M \in \mathbb{C}^{q\times l}$, is continuous as well. Hence we have the following corollary to Proposition 2.4.3.

Corollary 2.4.4 *If $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}^{\mathbb{C}}$ then the function $\mu_{\mathbf{\Delta}}^{\|\cdot\|} : \mathbb{C}^{q\times l} \to [0,\infty)$ is continuous.*

The next theorem is basic. It establishes the connection between μ-functions and spectral value sets.

Theorem 2.4.5 *Let $(A, B, C) \in L_{n,l,q}$ and $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}$. Define*

$$f : \mathbb{C} \setminus \sigma(A) \to [0,\infty) \qquad f(s) := \mu_{\mathbf{\Delta}}^{\|\cdot\|}(G(s)), \quad \text{where} \quad G(s) = C(sI_n - A)^{-1}B.$$

Then, for $s \in \mathbb{C} \setminus \sigma(A)$,

$$f(s) = (\, \min\{\|\Delta\| \mid \ \Delta \in \mathbf{\Delta},\ s \in \sigma(A + B\Delta C)\}\,)^{-1}. \tag{2.13}$$

We have for every $\rho > 0$,

$$\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A, B, C, \rho) = \sigma(A) \cup \{\, s \in \mathbb{C} \setminus \sigma(A) \mid f(s) \geq \rho^{-1}\,\}, \tag{2.14}$$

$$\overset{\circ}{\sigma}{}_{\mathbf{\Delta}}^{\|\cdot\|}(A, B, C, \rho) = \sigma(A) \cup \{\, s \in \mathbb{C} \setminus \sigma(A) \mid f(s) > \rho^{-1}\,\}. \tag{2.15}$$

Proof: The relation (2.13) follows from the definition of $\mu_{\mathbf{\Delta}}^{\|\cdot\|}(\cdot)$ and the equivalence

$$s \in \sigma(A + B\Delta C) \ \Leftrightarrow\ 1 \in \sigma(\Delta G(s)), \tag{2.16}$$

which holds for all $s \in \mathbb{C} \setminus \sigma(A)$ and all $\Delta \in \mathbb{C}^{l\times q}$. We prove the equivalence:

Let $s \in \sigma(A + B\Delta C) \setminus \sigma(A)$, and let $x \neq 0$ be a corresponding eigenvector. Then

$$0 = (sI_n - A - B\Delta C)x = (sI_n - A)(I_n - (sI_n - A)^{-1}B\Delta C)x.$$

Thus

$$x = (sI_n - A)^{-1}B\Delta Cx. \tag{2.17}$$

It follows that $\Delta Cx \neq 0$. By multiplying equation (2.17) with ΔC one obtains

$$\Delta Cx = \Delta C(sI_n - A)^{-1}B\Delta Cx = \Delta G(s)(\Delta Cx).$$

Hence, $1 \in \sigma(\Delta G(s))$.

The proof of the converse direction is similar: Suppose that $1 \in \sigma(\Delta G(s))$. Then there is a $y \in \mathbb{C}^l \setminus \{0\}$ such that

$$y = \Delta G(s)y = \Delta C(sI_n - A)^{-1}By.$$

It follows that $(sI_n - A)^{-1}By \neq 0$, and

$$(sI_n - A)((sI_n - A)^{-1}By) = B\Delta C((sI_n - A)^{-1}By).$$

Thus the vector $(sI_n - A)^{-1}By$ is an eigenvector of $A + B\Delta C$ to the eigenvalue s. This concludes the proof of the equivalence (2.16) and the relation (2.13). The relations (2.14) and (2.15) are immediate consequences of (2.13). $\square$

Below we list some further properties of the function f and relate them to the properties of spectral value sets.

Proposition 2.4.6 *For the function f in Theorem 2.4.5 we define*

$$\begin{aligned} C_f(\rho) &:= \{\, s\in\mathbb{C}\setminus\sigma(A) \mid f(s)=\rho^{-1} \,\}, \quad \rho>0,\\ Z_f &:= \{\, s\in\mathbb{C}\setminus\sigma(A) \mid f(s)=0 \,\},\\ M_f &:= \{\, s\in\mathbb{C}\setminus\sigma(A) \mid f \text{ has a local minimum at } s \,\},\\ D_f &:= \{\, s\in\mathbb{C}\setminus\sigma(A) \mid f \text{ is discontinuous at } s \,\}, \end{aligned}$$

Then the following holds.

(i) *The function $f:\mathbb{C}\setminus\sigma(A)\to[0,\infty)$ is upper semi-continuous.*

(ii) *The function f satisfies $\lim_{s\to\infty} f(s)=0$.*

(iii) *We have $\bigcup_{\rho>0}\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)=\sigma(A)\cup(\mathbb{C}\setminus Z_f)$.*

(iv) *Let $s\in\mathbb{C}\setminus\sigma(A)$ and $\rho>0$. Assume that $f(s)=\rho^{-1}$. Then in each neighborhood $U\subseteq\mathbb{C}$ of s there is an $s'\in U$ such that $f(s')>\rho^{-1}$.*

(v) *The restriction of f to the set $\mathbb{C}\setminus(\sigma(A)\cup Z_f)$ has no local maxima.*

(vi) *Let $\rho>0$. Then $\operatorname{int}(C_f(\rho))=\operatorname{int}(M_f\setminus Z_f)=\emptyset$.*

(vii) *Assume that $s\notin D_f$ and $f(s)>\rho^{-1}$. Then $s\in\operatorname{int}(\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho))$.*

(viii) *Assume that $s\in D_f$. Then there exists a $\rho>0$ such that $f(s)>\rho^{-1}$ and $s\in\partial\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$.*

(ix) *For all $\rho>0$,*

$$C_f(\rho)\setminus M_f \;\subseteq\; \partial\sigma(A,B,C,\rho) \tag{2.18}$$

and

$$C_f(\rho)\setminus(M_f\cup D_f) \;=\; \partial\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)\setminus(\sigma(A)\cup D_f). \tag{2.19}$$

(x) *Let K be a connected component of $\mathbb{C}\setminus(C_f(\rho)\cup D_f)$. Then either $K\cap\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)=\emptyset$ or $K\subset\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$.*

(xi) *Let K be a bounded connected component of $\mathbb{C}\setminus\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$ and suppose that $\operatorname{cl}(K)\cap(D_f\cup\sigma(A))=\emptyset$. Then $K\cap M_f\neq\emptyset$.*

(xii) *Let $(\mathbf{\Delta},\|\cdot\|)\in\mathcal{P}_{l,q}^{\mathbb{C}}$. Then*

(a) *The function f is continuous.*

(b) *Either $Z_f=\mathbb{C}\setminus\sigma(A)$ or $\#Z_f\leq n-1$. If the set $\{B\Delta C \mid \Delta\in\mathbf{\Delta}\}$ contains a nonsingular matrix then $Z_f=\emptyset$.*

Remark 2.4.7 *The claims (i) and (v) imply that for each compact set $K\subset\mathbb{C}\setminus\sigma(A)$ the restriction $f|_K$ of the upper semi-continuous function f attains its maximum on the boundary of K. This fact has been used by Gallestey [31] to develop an algorithm for the computation of the spectral value sets $\sigma_{\mathbb{C}}^{(2)}(A,B,C,\rho)$.*

Proof: (i) follows from the upper semi-continuity of $\mu_{\mathbf{\Delta}}^{\|\cdot\|}(\cdot)$.

(ii). Suppose that the claim fails. Then there is a $\rho > 0$ and an unbounded sequence (s_k) in $\mathbb{C}$ such that $f(s_k) \geq \rho^{-1}$ for all $k \in \mathbb{N}$. By Equation (2.14) this implies that the spectral value set $\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A, B, C, \rho)$ is unbounded. The latter contradicts the first claim of Proposition 2.2.4.

(iii) is immediate from Equation (2.14). (iv) follows from Equations (2.14), (2.15) and claim (ii) of Proposition 2.2.4.

(v) and (vi) are consequences of (iv).

(vii). If $s \notin D_f$ and $f(s) > \rho^{-1}$ then there is an open neighborhood $U \subset \mathbb{C}$ of s such that $f(s') > \rho^{-1}$ for all $s' \in U$. Thus, by (2.14), $U \subseteq \sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A, B, C, \rho)$. Thus s is an interior point of $\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A, B, C, \rho)$.

$(viii)$. Suppose that f is not continuous at s. Then there is an $\epsilon \in (0, f(s))$ and a sequence (s_k) converging to s such that $|f(s_k) - f(s)| > \epsilon$. Since f is upper semi-continuous, we may assume that $f(s_k) < f(s) - \epsilon$ for all $k \in \mathbb{N}$. The latter implies that $s \in \partial\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A, B, C, \rho)$, where $\rho = (f(s) - \epsilon)^{-1}$.

(ix). Suppose that $s \in C_f(\rho) \setminus M_f$. Then $f(s) = \rho^{-1}$, and in every neighborhood of s there is an s' such that $f(s') < \rho^{-1}$. Thus (2.18) holds. Suppose now, that $s \in \partial\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A, B, C, \rho) \setminus (\sigma(A) \cup D_f)$. Then, by (vii), we have that $s \in C_f(\rho)$, i.e. $f(s) = \rho^{-1}$. We also have that $s \notin M_f$ since otherwise there is a neighborhood U of s such that $f(s') \geq \rho^{-1}$ for every $s' \in U$, which implies that $s \in \operatorname{int}(\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A, B, C, \rho))$. Thus (2.19) holds.

(x). Let K be a connected component of $\mathbb{C} \setminus (C_f(\rho) \cup D_f)$. Then the restriction $f|_K$ is continuous and $f(s) \neq \rho^{-1}$ for all $s \in K$. Thus, either $f(s) > \rho^{-1}$ for all $s \in K$ or $f(s) < \rho^{-1}$ for all $s \in K$.

(xi). Let K be a bounded connected component of $\mathbb{C} \setminus \sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A, B, C, \rho)$ and suppose that $\operatorname{cl}(K) \cap (D_f \cup \sigma(A)) = \emptyset$. Then f is continuous on the compact set $\operatorname{cl}(K)$. Thus there is an $s_0 \in \operatorname{cl}(K)$ such that $f(s_0) = \min_{s \in \operatorname{cl}(K)} f(s)$. For $s \in K$ we have $f(s) < \rho^{-1}$, and for $s \in \partial K$ we have $f(s) = \rho^{-1}$. Thus s_0 is contained in the open set K, whence $K \cap M_f \neq \emptyset$.

(xii). Claim (a) is a consequence of Corollary 2.4.4. (b) follows from (iii) and the second claim of Proposition 2.2.5. $\square$

In view of Theorem 2.4.5 it is of importance to have computable formulas or efficient algorithms to calculate or approximate the function $f(\cdot)$. Such formulas (algorithms) are known for the following cases.

(a) $\mathbf{\Delta} = \mathbb{C}^{l \times q}$ and $\|\cdot\|$ is rank-1-compatible with a computable operator norm. (The definition of rank-1-compatiblity will be given in the next section).

(b) $\mathbf{\Delta} = \mathbb{R}^{l \times q}$ and $\|\cdot\|$ is the spectral norm.

(c) $l = 1$ or $q = 1$ and $\|\cdot\|$ is an arbitrary norm on $\mathbf{\Delta} = \mathbb{R}^{l \times q}$.

(d) $\mathbf{\Delta} = \left\{ \operatorname{diag}(\delta_1 I_{r_1}, \ldots, \delta_p I_{r_p}, \Delta_1, \ldots, \Delta_m) \;\middle|\; \delta_j \in \mathbb{C},\ \Delta_k \in \mathbb{C}^{l_k \times l_k} \right\}$, where $r_j, l_k \in \mathbb{N}$ are fixed numbers. The underlying norm $\|\cdot\|$ is the spectral norm. In this case $\mu_{\mathbf{\Delta}}^{\|\cdot\|}(M)$ is called the structured singular value of M. It plays an important role in control theory, see [75, 84, 98].

We will treat the case (a) in the next section and the cases $(b), (c)$ in the chapters 6 and 7. However, we will not deal with case (d) in this work. The reader who is interested in the computation of the structured singular value is refered to [1, 69, 75] and the references therein. In Chapter 4 we will study μ-functions for the case (*e*) $M \in \mathbb{C}^{l\times q}$ is block diagonal and $\mathbf{\Delta} = \mathbf{\Delta}_{\mathcal{I}}$ is the set of complex block matrices $\Delta = [\Delta_{jk}]_{j,k\leq m}$ of fixed block sizes whose blocks satisfy $\Delta_{jk} = 0$ for all pairs $(j,k) \in \underline{m} \times \underline{m}$ which are not contained in the index set $\mathcal{I} \subseteq \underline{m} \times \underline{m}$.

All figures of spectral value sets in this work have been obtained by evaluating the function $s \mapsto f(s) = \mu_{\mathbf{\Delta}}^{\|\cdot\|}(G(s))$ on a grid. For the calculation of $G(s) = C(sI_n - A)^{-1}B$ from the data A, B, C and s the application of the inverse of $sI_n - A$ is needed. This is computationally expensive and may cause numerical problems. However, in the case that $B = C = I_n$ the application of the inverse can be avoided if one uses the following functions.

Definition 2.4.8 *Let* $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{n,n}$. *Then we define for* $M \in \mathbb{C}^{n\times n}$,

$$\widetilde{\mu}_{\mathbf{\Delta}}^{\|\cdot\|}(M) \quad := \quad \min\{\ \|\Delta\| \ \mid \Delta \in \mathbf{\Delta},\ M - \Delta \text{ is singular }\}. \tag{2.20}$$

Lemma 1.7.1 yields that the functions $\widetilde{\mu}_{\mathbf{\Delta}}^{\|\cdot\|} : \mathbb{C}^{n\times n} \to [0,\infty]$ are well defined and lower semi-continuous. Note that $\widetilde{\mu}_{\mathbf{\Delta}}^{\|\cdot\|}(M) = \infty$ if and only if there is no $\Delta \in \mathbf{\Delta}$ such that $M - \Delta$ is singular.

Theorem 2.4.9 *Let* $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{n,n}$. *Then for every* $A \in \mathbb{C}^{n\times n}$ *and every* $\rho \geq 0$,

$$\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,\rho) = \{\ s \in \mathbb{C}\ \mid\ \widetilde{\mu}_{\mathbf{\Delta}}^{\|\cdot\|}(sI_n - A) \leq \rho\ \}.$$

Proof: This follows from the definitions of $\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,\rho)$ and $\widetilde{\mu}_{\mathbf{\Delta}}^{\|\cdot\|}(\cdot)$ and the equivalence

$$s \in \sigma(A + \Delta) \quad \Leftrightarrow \quad (sI_n - A) - \Delta \text{ is singular }.$$

. □

The functions $\mu_{\mathbf{\Delta}}^{\|\cdot\|}(\cdot)$ and $\widetilde{\mu}_{\mathbf{\Delta}}^{\|\cdot\|}(\cdot)$ are related in the following way.

Proposition 2.4.10 *Let* $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{n,n}$ *and let* $M \in \mathbb{C}^{n\times n}$. *Then*

$$\widetilde{\mu}_{\mathbf{\Delta}}^{\|\cdot\|}(M) = \begin{cases} 0 & \text{if } M \text{ is singular,} \\ \mu_{\mathbf{\Delta}}^{\|\cdot\|}(M^{-1})^{-1} & \text{otherwise.} \end{cases} \tag{2.21}$$

Proof: This follows from the definitions of $\widetilde{\mu}_{\mathbf{\Delta}}^{\|\cdot\|}(\cdot)$ and $\mu_{\mathbf{\Delta}}^{\|\cdot\|}(\cdot)$ and the equivalence

$$M - \Delta \text{ is singular} \quad \Leftrightarrow \quad 1 \in \sigma(\Delta M^{-1}),$$

which holds for all $\Delta \in \mathbb{C}^{n\times n}$ and all nonsingular $M \in \mathbb{C}^{n\times n}$. The equivalence is an immediate consequence of the relation $M - \Delta = (I_n - \Delta M^{-1})M$. □

In the sequel we will make use of some simplified notation.

- $\mu_{\mathbb{F}}^{\|\cdot\|}(\cdot) := \mu_{\mathbb{F}^{l\times q}}^{\|\cdot\|}(\cdot)$ and $\widetilde{\mu}_{\mathbb{F}}^{\|\cdot\|}(\cdot) := \widetilde{\mu}_{\mathbb{F}^{n\times n}}^{\|\cdot\|}(\cdot)$.
- If $\|\cdot\| = \|\cdot\|_{\alpha,\beta}$ is an operator norm subordinate to the vector norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ then $\mu_{\mathbf{\Delta}}^{(\alpha,\beta)}(\cdot) := \mu_{\mathbf{\Delta}}^{\|\cdot\|}(\cdot)$ and $\widetilde{\mu}_{\mathbf{\Delta}}^{(\alpha,\beta)}(\cdot) := \widetilde{\mu}_{\mathbf{\Delta}}^{\|\cdot\|}(\cdot)$. Furthermore, $\mu_{\mathbf{\Delta}}^{(\alpha)}(\cdot) := \mu_{\mathbf{\Delta}}^{(\alpha,\alpha)}(\cdot)$ and $\widetilde{\mu}_{\mathbf{\Delta}}^{(\alpha)}(\cdot) := \widetilde{\mu}_{\mathbf{\Delta}}^{(\alpha,\alpha)}(\cdot)$.

We close this section with a proposition on the stability radius. Its easy proof via Theorem 2.4.5 is left to the reader.

Proposition 2.4.11 *Let $\mathbb{C}_g$ be a nonempty open subset of $\mathbb{C}$. Let $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}$ and $(A,B,C) \in L_{n,l,q}$. Let, as in Theorem 2.4.5,*

$$f : \mathbb{C} \setminus \sigma(A) \to [0,\infty) \qquad f(s) := \mu_{\mathbf{\Delta}}^{\|\cdot\|}(G(s)), \quad \textit{where} \quad G(s) := C(sI_n - A)^{-1}B.$$

Suppose that A is $\mathbb{C}_g$-stable. Then

(i) *The following assertions are equivalent.*

   (a) $r_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\mathbb{C}_g) = \infty$.

   (b) *$A + B\Delta C$ is $\mathbb{C}_g$-stable for all $\Delta \in \mathbf{\Delta}$.*

   (c) $\bigcup_{\rho \geq 0} \sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho) \subseteq \mathbb{C}_g$.

   (d) *$\mathbb{C}_g \subseteq Z_f$, where $Z_f = \{\, s \in \mathbb{C} \setminus \sigma(A) \mid f(s) = 0 \,\}$.*

(ii) *We have*

$$\begin{aligned} r_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\mathbb{C}_g) &= \left(\max_{s \in \partial \mathbb{C}_g} f(s) \right)^{-1} && (2.22) \\ &= \min\{\, \rho > 0 \mid \sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho) \cap \partial \mathbb{C}_g \neq \emptyset \,\}. && (2.23) \end{aligned}$$

2.5 μ-functions and operator norms

In this section we consider μ-functions for perturbation structures of the form $(\mathbb{F}^{l \times q}, \|\cdot\|)$ where $\|\cdot\|$ is rank-1-compatible with an operator norm.

Definition 2.5.1 *Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be a norms on $\mathbb{F}^q$ and $\mathbb{F}^l$ respectively, $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$. A norm $\|\cdot\|$ on $\mathbb{F}^{l \times q}$ is said to be rank-1-compatible with the operator norm $\|\cdot\|_{\alpha,\beta}$ if*

(a) *$\|\Delta\| \geq \|\Delta\|_{\alpha,\beta}$ for all $\Delta \in \mathbb{F}^{l \times q}$.*

(b) *$\|\Delta\| = \|\Delta\|_{\alpha,\beta}$ for all $\Delta \in \mathbb{F}^{l \times q}$ with $\operatorname{rank}(\Delta) = 1$.*

The theorem below (from [74]) is the main result of this section. It extends a result of Hinrichsen and Pritchard (see e.g. [54]).

Theorem 2.5.2 *Let $\|\cdot\|_\alpha$ be a norm on $\mathbb{F}^q$ and let $\|\cdot\|_\beta$ be a norm on $\mathbb{F}^l$, $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$. Let $\|\cdot\|$ be a norm on $\mathbb{F}^{l \times q}$ which is rank-1-compatible with the operator norm $\|\cdot\|_{\alpha,\beta}$. Then we have for $M \in \mathbb{F}^{q \times l}$,*

$$\mu_{\mathbb{F}}^{\|\cdot\|}(M) = \|M\|_{\beta,\alpha}.$$

Suppose that $M \neq 0$ and let $x \in \mathbb{F}^l$ and $y \in \mathbb{F}^q$ be vectors such that $\|M\|_{\beta,\alpha} = y^T M x$ and $\|x\|_\beta = \|y\|_\alpha^D = 1$. Then the rank-1-matrix $\Delta_0 := \|M\|_{\beta,\alpha}^{-1} x y^T$ satisfies

$$1 \in \sigma(\Delta_0 M) \quad \textit{and} \quad \|\Delta_0\| = \|\Delta_0\|_{\alpha,\beta} = \|M\|_{\beta,\alpha}^{-1} = \mu_{\mathbb{F}}^{\|\cdot\|}(M)^{-1}.$$

Proof: The case $M = 0$ is obvious. Let $M \neq 0$ and suppose that $1 \in \sigma(\Delta M)$. Then there is a $\xi \in \mathbb{F}^l \setminus \{0\}$ such that $\xi = \Delta M \xi$. Hence,

$$\|\xi\|_\alpha \leq \|\Delta\|_{\alpha,\beta} \|M\|_{\beta,\alpha} \|\xi\|_\alpha \leq \|\Delta\| \|M\|_{\beta,\alpha} \|\xi\|_\alpha.$$

Thus $\|M\|_{\beta,\alpha} \geq \|\Delta\|^{-1}$. Thus, $\|M\|_{\beta,\alpha} \geq \sup\{\, \|\Delta\|^{-1} \,|\, \Delta \in \mathbb{F}^{l\times q},\, 1 \in \sigma(\Delta M)\,\} = \mu_{\mathbb{F}}^{\|\cdot\|}(M)$. On the other hand we have for the matrix Δ_0 that $\Delta_0 M x = x$ and $\|\Delta_0\| = \|\Delta_0\|_{\alpha,\beta} = \|M\|_{\beta,\alpha}^{-1} \|x\|_\beta \|y\|_\alpha^D = \|M\|_{\beta,\alpha}^{-1}$. Thus $\|M\|_{\beta,\alpha} = \mu_{\mathbb{F}}^{\|\cdot\|}(M)$. □

Observe that Theorem 2.5.2 does not cover the case $\mathbb{F} = \mathbb{R}$ and $M \notin \mathbb{C}^{q\times l} \setminus \mathbb{R}^{q\times l}$. Recall from the preliminaries the definition $\inf_{\alpha,\beta}(M) := \min_{\|x\|_\alpha = 1} \|Mx\|_\beta$, $M \in \mathbb{F}^{q\times l}$. If $q = l$ and M is nonsingular then $\inf_{\alpha,\beta}(M) = \|M^{-1}\|_{\beta,\alpha}^{-1}$.

Corollary 2.5.3 *Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be norms on $\mathbb{F}^n$, and let $\|\cdot\|$ be a norm on $\mathbb{F}^{n\times n}$ which is rank-1-compatible with $\|\cdot\|_{\alpha,\beta}$. Then, for all $M \in \mathbb{F}^{n\times n}$,*

$$\widetilde{\mu}_{\mathbb{F}}^{\|\cdot\|}(M) = \inf_{\alpha,\beta}(M).$$

Proof: If M is singular then $\widetilde{\mu}_{\mathbb{F}}^{\|\cdot\|}(M) = 0 = \inf_{\alpha,\beta}(M)$. If M is nonsingular then $\mu_{\mathbb{F}}^{\|\cdot\|}(M^{-1}) = \|M^{-1}\|_{\beta,\alpha}$ by Theorem 2.5.2. Thus $\inf_{\alpha,\beta}(M) = \|M^{-1}\|_{\beta,\alpha}^{-1} = \mu_{\mathbb{F}}^{\|\cdot\|}(M^{-1})^{-1} = \widetilde{\mu}_{\mathbb{F}}^{\|\cdot\|}(M)$. □

By applying Theorem 2.5.2 and Corollary 2.5.3 to the norm $\|\cdot\| = \|\cdot\|_{\alpha,\beta}$ we obtain

Corollary 2.5.4 *Let $\|\cdot\|_\alpha$ be a norm on $\mathbb{F}^q$ and let $\|\cdot\|_\beta$ be a norm on $\mathbb{F}^l$, $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$. Then we have for $M \in \mathbb{F}^{q\times l}$,*

$$\mu_{\mathbb{F}}^{(\alpha,\beta)}(M) = \|M\|_{\beta,\alpha}.$$

If, in addition, $l = q =: n$ then for $M \in \mathbb{F}^{n\times n}$,

$$\widetilde{\mu}_{\mathbb{F}}^{(\alpha,\beta)}(M) = \inf_{\alpha,\beta}(M).$$

Next, we treat unitarily invariant norms. Recall from the preliminaries that a unitarily invariant norm is said to be normalized if $\|\Delta\| = \sigma_1(\Delta)$ for all all rank-1 matrices Δ. Now, let $\|\cdot\|$ be an arbitrary unitarily invariant norm on $\mathbb{C}^{l\times q}$, and let $|\!|\!|\cdot|\!|\!|$ be the absolute and symmetric norm on $\mathbb{R}^{\min\{l,q\}}$ such that

$$\|\Delta\| = |\!|\!|\; [\sigma_1(\Delta), \ldots, \sigma_{\min\{m,n\}}(\Delta)]^T \;|\!|\!|$$

Set $c := |\!|\!|\; [1, 0, \ldots, 0]^T \;|\!|\!|$. Then it follows from the monotonicity of absolute norms that for all $\Delta \in \mathbb{C}^{l\times q}$,

$$\|\Delta\| \geq c\, \sigma_1(\Delta).$$

Equality holds if $\operatorname{rank} \Delta = 1$. Thus $\|\cdot\|$ is normalized if and only if $c = 1$. Moreover, any normalized unitarily invariant norm is rank-1-compatible with the spectral norm. Thus Theorem 2.5.2 and Corollary 2.5.3 yield

Corollary 2.5.5 *Let $\|\cdot\|$ be a normalized unitarily invariant norm on $\mathbb{C}^{l\times q}$ and let $M \in \mathbb{F}^{q\times l}$ $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$. Then*

$$\mu_{\mathbb{F}}^{\|\cdot\|}(M) = \mu_{\mathbb{F}}^{(2)}(M) = \|M\|_2 = \sigma_{\max}(M).$$

If, in addition, $l = q =: n$ then for $M \in \mathbb{F}^{n\times n}$,

$$\widetilde{\mu}_{\mathbb{F}}^{\|\cdot\|}(M) = \widetilde{\mu}_{\mathbb{F}}^{(2)}(M) = \inf_{2,2}(M) = \sigma_{\min}(M),$$

where $\sigma_{\max}(\cdot)$ and $\sigma_{\min}(\cdot)$ denote the maximal and the minimal singular value.

We note that if $M \in \mathbb{C}^{l\times q} \setminus \mathbb{R}^{l\times q}$ then in general we have $\mu_{\mathbb{R}}^{(2)}(M) \neq \|M\|_2$. The function $\mu_{\mathbb{R}}^{(2)}(\cdot)$ will be discussed in Chapter 6. There is another class of norms for which Theorem 2.5.2 is applicable:

Proposition 2.5.6 (see [74, Corollary 2.2.11]) *Let $\|\cdot\|_\alpha$ be a norm on $\mathbb{F}^q$ and let $\|\cdot\|_\beta$ be an* absolute *norm on $\mathbb{F}^l$, $\mathbb{F}\in\{\mathbb{R},\mathbb{C}\}$. Let Δ_j denote the jth row of $\Delta\in\mathbb{C}^{l\times q}$. Then the norm*

$$\|\cdot\| : \mathbb{C}^{l\times q} \to [0,\infty), \qquad \|\Delta\| := \|\ [\ \|\Delta_1^T\|_\alpha^D, \ldots, \|\Delta_l^T\|_\alpha^D\ ]^T\ \|_\beta,$$

Then $\|\cdot\|$ is rank-1-compatible with $\|\cdot\|_{\alpha,\beta}$. Thus, $\mu_{\mathbb{F}}^{\|\cdot\|}(M) = \|M\|_{\beta,\alpha}$ for all $M\in\mathbb{F}^{q\times l}$, and if $q = l$, $\widetilde{\mu}_{\mathbb{F}}^{\|\cdot\|}(M) = \inf_{\alpha,\beta}(M)$.

Proof: First note that the function $\|\cdot\|$ is in fact a norm. The triangle inequality follows from the monotonicity of the absolute norm $\|\cdot\|_\beta$.

We have for $x\in\mathbb{F}^q$,

$$\begin{aligned} |\Delta x| &= [\ |\Delta_1 x|, \ldots, |\Delta_l x|\ ]^T \\ &\leq [\ \|\Delta_1^T\|_\alpha^D\, \|x\|_\alpha,\ \ldots,\ \|\Delta_l^T\|_\alpha^D\, \|x\|_\alpha\ ]^T \\ &= [\ \|\Delta_1^T\|_\alpha^D, \ldots, \|\Delta_l^T\|_\alpha^D\ ]^T\ \|x\|_\alpha. \end{aligned}$$

Hence,

$$\|\Delta x\|_\beta = \|\ |\Delta x|\ \|_\beta \leq \|\ [\ \|\Delta_1^T\|_\alpha^D, \ldots, \|\Delta_l^T\|_\alpha^D\ ]^T\ \|_\beta\ \|x\|_\alpha = \|\Delta\|\ \|x\|_\alpha.$$

Thus $\|\Delta\|_{\alpha,\beta} \leq \|\Delta\|$. Now suppose that $\mathrm{rank}(\Delta) = 1$. Then there are $v\in\mathbb{F}^l, w\in\mathbb{F}^q$ such that $\Delta = xy^T$. The jth row of Δ is $\Delta_j = x_j y^T$, where x_j denotes the jth entry of x. We have

$$\begin{aligned} \|\Delta\| &= \|\ [\ \|x_1 y\|_\alpha^D, \ldots, \|x_l y\|_\alpha^D\ ]^T\ \|_\beta \\ &= \|\ [\ |x_1|\ \|y\|_\alpha^D, \ldots, |x_l|\ \|y\|_\alpha^D\ ]^T\ \|_\beta \\ &= \|\ |x|\ \|_\beta\ \|y\|_\alpha^D \\ &= \|x\|_\beta\ \|y\|_\alpha^D \\ &= \|\Delta\|_{\alpha,\beta}. \end{aligned}$$

□

Remark 2.5.7 *The identification of a μ-function with an operator norm does not imply that the computational problem for μ-function in question is solved. Clearly, it is well known how to compute the norms $\|M\|_p$ for $p = 1, 2, \infty$ even for large matrices. However, to our knowledge there is no reliable algorithm to compute Hölder norms for $p \neq 1, 2, \infty$. There are algorithms which estimate these norms, see [39] and the references therein. Another interesting norm is $\|\cdot\|_{\infty,1}$ for the following reason. By Theorem 2.5.2,*

$$\mu_{\mathbb{F}}^{\|\cdot\|_{1,\infty}}(M) = \|M\|_{\infty,1}.$$

The norm $\|\cdot\|_{1,\infty}$ is the componentwise maximum norm. Thus $\|\cdot\|_{\infty,1}$ is the μ-function which corresponds to componentwise measured perturbations. It is known that the computation of this norm is NP-hard, see [79].

2.6 Small perturbations and condition numbers

In this section we study the asymptotic behavior of spectral values sets for small perturbations. In doing so we obtain formulas for the Hölder condition numbers of eigenvalues with respect to structured perturbations. These formulas generalize known results for the unstructured case [19, 37, 73] which will be discussed in Chapter 5.

In order to define Hölder condition numbers of eigenvalues we need some preparations. First, recall the following definition from [86, p.167].

Definition 2.6.1 *Let $A \in \mathbb{C}^{n\times n}$ have eigenvalues $\lambda_1, \dots, \lambda_n \in \mathbb{C}$ and let $\widetilde{A} \in \mathbb{C}^{n\times n}$ have eigenvalues $\widetilde{\lambda}_1, \dots, \widetilde{\lambda}_n \in \mathbb{C}$. Then the spectral variation of $\widetilde{A}$ with respect to A is*

$$\mathrm{sv}_A(\widetilde{A}) := \max_{i\in\underline{n}} \min_{j\in\underline{n}} |\widetilde{\lambda}_i - \lambda_j|.$$

Note that

$$\mathrm{sv}_A(\widetilde{A}) = \min\{\, \rho \geq 0 \mid \ \sigma(\widetilde{A}) \subseteq \textstyle\bigcup_{j\in\underline{n}} \mathcal{D}(\lambda_j, \rho) \,\}.$$

Inspired by the definition above we introduce the spectral variation of $\widetilde{A}$ with respect to an eigenvalue λ_j of A. We use the following notation. Let $a_1, \dots, a_n \in \mathbb{R}$, and let $\pi : \underline{n} \to \underline{n}$ be a permutation such that $a_{\pi(1)} \leq a_{\pi(2)} \leq \ldots \leq a_{\pi(n)}$. Then we set for $m \in \underline{n}$,

$$\min\nolimits_{i\in\underline{n}}^{(m)} a_i := a_{\pi(m)}.$$

Definition 2.6.2 *Let $\lambda_j \in \mathbb{C}$ be an eigenvalue of algebraic multiplicity m of $A \in \mathbb{C}^{n\times n}$. Let $\widetilde{A} \in \mathbb{C}^{n\times n}$ have eigenvalues $\widetilde{\lambda}_1, \dots, \widetilde{\lambda}_n \in \mathbb{C}$. Then the spectral variation of $\widetilde{A}$ with respect to the eigenvalue λ_j of A is*

$$\mathrm{sv}_{(A,\lambda_j)}(\widetilde{A}) := \min\nolimits_{i\in\underline{n}}^{(m)} |\widetilde{\lambda}_i - \lambda_j|.$$

The definition implies that for an eigenvalue λ_j of algebraic multiplicity m,

$$\mathrm{sv}_{(A,\lambda_j)}(\widetilde{A}) = \min\{\, \rho \geq 0 \mid \ \text{the disk } \mathcal{D}(\lambda_j, \rho) \text{ contains at least } m \text{ eigenvalues of } \widetilde{A} \,\}.$$

We now give the definition of condition numbers.

Definition 2.6.3 *Let $(X, \|\cdot\|_\alpha)$ and $(Y, \|\cdot\|_\beta)$ be Banach spaces. Let U be a subset of X and let $x_0 \in U$ be an accumulation point of U. The Hölder condition number of order $\omega > 0$ of a function $g : U \to Y$ at the point x_0 is defined by*

$$\mathcal{C}_\omega(g, x_0) := \lim_{\rho \searrow 0} \sup \left\{ \frac{\|g(x) - g(x_0)\|_\beta}{\|x - x_0\|_\alpha^\omega} \;\middle|\; 0 < \|x - x_0\|_\alpha \leq \rho,\ x \in U \right\}.$$

Note that $\mathcal{C}_\omega(g, x_0) \in [0, \infty]$ is well defined for all $\omega > 0$. However, there is at most one order $\omega > 0$ such that $0 \neq \mathcal{C}_\omega(g, x_0) \neq \infty$, since these inequalities imply that $\mathcal{C}_{\widetilde{\omega}}(g, x_0) = 0$ for $\widetilde{\omega} < \omega$ and $\mathcal{C}_{\widetilde{\omega}}(g, x_0) = \infty$ for $\widetilde{\omega} > \omega$. If g is discontinuous at x_0 then $\mathcal{C}_\omega(g, x_0) = \infty$ for all $\omega > 0$. We remark that our terminology here differs slightly from that in [21, Definiton 4.1] where the quantity $\mathcal{C}_\omega(g, x_0)$ is called *asymptotic* Hölder condition number.

Definition 2.6.4 *Let $(A, B, C) \in L_{n,l,q}$ and $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}$. For $\Delta \in \mathbf{\Delta}$ let $A_\Delta = A + B\Delta C$. Then the Hölder condition number of the eigenvalue $\lambda_j \in \sigma(A)$ of order $\omega > 0$ is*

$$\mathrm{cond}_{\mathbf{\Delta},\omega}^{\|\cdot\|}(A, B, C, \lambda_j) \;:=\; \lim_{\rho \searrow 0} \max_{\substack{0 \neq \Delta \in \mathbf{\Delta} \\ \|\Delta\| \leq \rho}} \frac{\mathrm{sv}_{(A,\lambda_j)}(A_\Delta)}{\|\Delta\|^\omega}. \tag{2.24}$$

We have,

$$\mathrm{cond}_{\mathbf{\Delta},\omega}^{\|\cdot\|}(A,B,C,\lambda_j) = \mathcal{C}_\omega(g,0),$$

where

$$g : (\mathbf{\Delta},\,\|\cdot\|\,) \to (\mathbb{R},\,|\cdot|\,), \qquad g(\Delta) = \mathrm{sv}_{(A,\lambda_j)}(A_\Delta).$$

Moreover, we define the spectral variation with respect to λ_j and the perturbation level ρ by

$$\mathrm{sv}_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\lambda_j,\rho) \quad := \quad \max_{\substack{\Delta\in\mathbf{\Delta}\\ \|\Delta\|\le\rho}} \mathrm{sv}_{(A,\lambda_j)}(A_\Delta).$$

In the sequel we consider only the case of complex perturbation structures, i.e. we assume that $(\mathbf{\Delta},\|\cdot\|)\in\mathcal{P}_{l,q}^{\mathbb{C}}$. For these perturbation structures the associated μ-functions satisfy

$$\mu_{\mathbf{\Delta}}^{\|\cdot\|}(sM) = |s|\,\mu_{\mathbf{\Delta}}^{\|\cdot\|}(M)$$

for all $M\in\mathbb{C}^{q\times l}$ and all $s\in\mathbb{C}$. This fact will be needed in the proof of our main result.

The proposition below gives the relationship between condition numbers, spectral variations and spectral value sets.

Proposition 2.6.5 *Let $(A,B,C)\in L_{n,l,q}$ and $(\mathbf{\Delta},\|\cdot\|)\in\mathcal{P}_{l,q}^{\mathbb{C}}$. Let $\mathcal{K}_j(\rho)$ denote the connected component of $\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\rho)$ which contains the eigenvalue $\lambda_j\in\sigma(A)$. Then the following holds.*

(i) Let $\rho>0$ be such that $\mathcal{K}_j(\rho)\cap\sigma(A)=\{\lambda_j\}$. Then

$$\mathrm{sv}_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\lambda_j,\rho) \quad = \quad \max_{\lambda\in\partial\mathcal{K}_j(\rho)} |\lambda-\lambda_j| \;=\; \max_{\substack{\Delta\in\mathbf{\Delta}\\ \|\Delta\|=\rho}} \mathrm{sv}_{(A,\lambda_j)}(A_\Delta).$$

(ii) Let $c,\omega>0$, and suppose that the spectral variation of λ_j satisfies [1]

$$\mathrm{sv}_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\lambda_j,\rho) \quad = \quad c\,\rho^\omega + o(\rho^\omega).$$

Then we have

$$\mathrm{cond}_{\mathbf{\Delta},\omega}^{\|\cdot\|}(A,B,C,\lambda_j) = c.$$

Proof: (i). Let $\Delta\in\mathbf{\Delta}$, $\|\Delta\|\le\rho$. By Proposition 2.2.3 the assumption $\mathcal{K}_j(\rho)\cap\sigma(A)=\{\lambda_j\}$ implies that precisely m eigenvalues of A_Δ are contained in $\mathcal{K}_j(\rho)$, where m is the algebraic multiplicity of λ_j. Thus

$$\mathrm{sv}_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\lambda_j,\rho) \le \max_{\lambda\in\mathcal{K}_j(\rho)} |\lambda-\lambda_j| = \max_{\lambda\in\partial\mathcal{K}_j(\rho)} |\lambda-\lambda_j|.$$

Let $s_0\in\partial\mathcal{K}_j(\rho)$ be such that $|s_0-\lambda_j| = \max_{\lambda\in\partial\mathcal{K}_j(\rho)}|\lambda-\lambda_j|$. Then by claim (ii) of Proposition 2.2.5 there exists a $\Delta_\rho\in\mathbf{\Delta}$ with $\|\Delta_\rho\|=\rho$ and $s_0\in\sigma(A_{\Delta_\rho})$. This concludes the proof of (i).

(ii). Let $\epsilon>0$ and $\rho_0>0$ be such that for $\rho\in(0,\rho_0)$,

$$|\,\rho^{-\omega}\,\mathrm{sv}_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\lambda_j,\rho) - c\,| < \epsilon$$

[1]The Landau-symbol $o(\rho^\omega)$ stands for a function satisfying $\lim_{\rho\searrow 0}\rho^{-\omega}o(\rho^\omega)=0$.

and $\mathcal{K}_j(\rho) \cap \sigma(A) = \{\lambda_j\}$. Then for every $\rho \in (0, \rho_0)$ the matrix Δ_ρ defined in the proof of (i) satisfies

$$\frac{\mathrm{sv}_{(A,\lambda_j)}(A_{\Delta_\rho})}{\|\Delta_\rho\|^\omega} = \frac{\mathrm{sv}_{\mathbf{\Delta}}^{\|\cdot\|}(A, B, C, \lambda_j, \rho)}{\rho^\omega} \geq c - \epsilon.$$

Thus

$$\max_{\substack{0 \neq \Delta \in \mathbf{\Delta} \\ \|\Delta\| \leq \rho}} \frac{\mathrm{sv}_{(A,\lambda_j)}(A_\Delta)}{\|\Delta\|^\omega} \geq c - \epsilon. \tag{2.25}$$

On the other hand we have for $\Delta \in \mathbf{\Delta}$, $\|\Delta\| \leq \rho$,

$$\frac{\mathrm{sv}_{(A,\lambda_j)}(A_\Delta)}{\|\Delta\|^\omega} \leq \frac{\mathrm{sv}_{\mathbf{\Delta}}^{\|\cdot\|}(A, B, C, \lambda_j, \|\Delta\|)}{\|\Delta\|^\omega} \leq c + \epsilon.$$

An hence,

$$\max_{\substack{0 \neq \Delta \in \mathbf{\Delta} \\ \|\Delta\| \leq \rho}} \frac{\mathrm{sv}_{(A,\lambda_j)}(A_\Delta)}{\|\Delta\|^\omega} \leq c + \epsilon. \tag{2.26}$$

The inequalities (2.25) and (2.26) imply that $|\, \rho^{-\omega} \operatorname{cond}_{\mathbf{\Delta},\omega}^{\|\cdot\|}(A, B, C, \lambda_j) - c \,| < \epsilon$. □

In order to state the main result of this section we need some additional notation. Let $\lambda_1, \ldots, \lambda_r$ be the pairwise different eigenvalues of $A \in \mathbb{C}^{n \times n}$. By the Jordan decomposition theorem we have

$$A = \sum_{j=1}^{r} (\, \lambda_j P_j + N_j \,), \tag{2.27}$$

where

- $P_1, \ldots, P_r \in \mathbb{C}^{n \times n}$ are the projectors onto the generalized eigenspaces of A, i.e.

$$\mathrm{range}(P_j) = \textstyle\bigcup_{\ell \in \mathbb{N}} \ker(\lambda_j I_n - A)^\ell \qquad \text{for } j \in \underline{r},$$

$$\sum_{j=1}^{r} P_j = I_n \qquad \text{and} \qquad P_j P_k = \begin{cases} 0 & \text{if } j \neq k, \\ P_j & \text{if } j = k \end{cases} \qquad \text{for } j, k \in \underline{r},$$

- $N_1, \ldots, N_r \in \mathbb{C}^{n \times n}$ are the nilpotent matrices

$$N_j := (A - \lambda_j I_n) P_j, \qquad j \in \underline{r}.$$

They satisfy the relations

$$N_j N_k = 0 \text{ if } k \neq j, \qquad \text{and} \qquad N_j P_k = P_k N_j = \begin{cases} 0 & \text{if } j \neq k, \\ N_j & \text{if } j = k \end{cases} \qquad \text{for all } j, k \in \underline{r}.$$

Let $\nu_j \in \mathbb{N}$ be the index of nilpotency of N_j, i.e. $\nu_j = \min\{\, \ell \in \mathbb{N} \mid N_j^\ell = 0 \,\}$ Then ν_j is the size of the largest Jordan block associated with the eigenvalue λ_j in the Jordan canonical form of A. If $\nu_j = 1$ (i.e. $N_j = 0$) then λ_j is called a semi-simple (nondefective) eigenvalue of A.

Using the above relations it is easily verified that for every $s \in \mathbb{C} \setminus \sigma(A)$,

$$(sI_n - A)^{-1} \;=\; \sum_{j=1}^{r} \left((s-\lambda_j)^{-1} P_j + \sum_{\ell=1}^{\nu_j - 1} (s-\lambda_j)^{-(\ell+1)} N_j^\ell \right). \tag{2.28}$$

Let B, C be such that $(A, B, C) \in L_{n,l,q}$. Then we have for $G(s) = C(sI_l - A)^{-1}B$ that

$$G(s) \;=\; \sum_{j=1}^{r} \left((s-\lambda_j)^{-1} C P_j B + \sum_{\ell=1}^{\ell_j - 1} (s-\lambda_j)^{-(\ell+1)} C N_j^\ell B \right), \tag{2.29}$$

where

$$\ell_j \;:=\; \begin{cases} 1 & \text{if } CN_j^\ell B = 0 \text{ for all } \ell \in \mathbb{N}, \\ 1 + \max\{\, \ell \in \mathbb{N} \mid CN_j^\ell B \neq 0 \,\} & \text{otherwise.} \end{cases} \tag{2.30}$$

If $l = q = n$ and the matrices B, C are nonsingular then $\ell_j = \nu_j$ for all $j \in \underline{r}$. The next theorem is the main result of this section. It gives the relationship between the growth of spectral value sets for small perturbations, Hölder condition numbers of eigenvalues, μ-functions and the Jordan structure of the perturbed matrix.

Theorem 2.6.6 *Let $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}^{\mathbb{C}}$ and $(A, B, C) \in L_{n,l,q}$, where A has the Jordan structure (2.27). For $j \in \underline{r}$ and $\rho > 0$ let $\mathcal{K}_j(\rho)$ denote the connected component of $\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A, B, C, \rho)$ which contains the eigenvalue $\lambda_j \in \sigma(A)$. Furthermore, let ℓ_j be defined as in (2.30).*

Suppose $\mu_{\mathbf{\Delta}}^{\|\cdot\|}(CP_jB) \neq 0$ and $\ell_j = 1$ (i.e. $CN^\ell B = 0$ for all $\ell \in \mathbb{N}$). Then

(i) To each $\epsilon > 0$ there is a $\rho_\epsilon > 0$ such that for all $\rho \in (0, \rho_\epsilon]$,

$$\mathcal{D}\left(\lambda_j\,,\, (1-\epsilon)\,\rho\,\mu_{\mathbf{\Delta}}^{\|\cdot\|}(CP_jB)\right) \subseteq \mathcal{K}_j(\rho) \subseteq \mathcal{D}\left(\lambda_j\,,\, (1+\epsilon)\,\rho\,\mu_{\mathbf{\Delta}}^{\|\cdot\|}(CP_jB)\right).$$

(ii) We have

$$\lim_{\rho \searrow 0} \frac{\mathcal{K}_j(\rho) - \lambda_j}{\rho} = \mathcal{D}\left(0\,,\, \mu_{\mathbf{\Delta}}^{\|\cdot\|}(CP_jB)\right),$$

where the limit is taken with respect to the Hausdorff metric.

(iii) The spectral variation satisfies

$$\mathrm{sv}_{\mathbf{\Delta}}^{\|\cdot\|}(A, B, C, \lambda_j, \rho) = \mu_{\mathbf{\Delta}}^{\|\cdot\|}(CP_jB)\,\rho + o(\rho).$$

(iv) We have

$$\mathrm{cond}_{\mathbf{\Delta},1}^{\|\cdot\|}(A, B, C, \lambda_j) = \mu_{\mathbf{\Delta}}^{\|\cdot\|}(CP_jB).$$

Suppose that $\ell_j \geq 2$ and $\mu_{\mathbf{\Delta}}^{\|\cdot\|}(CN_j^{\ell_j - 1}B) \neq 0$. Then

(v) *To each $\epsilon > 0$ there is a $\rho_\epsilon > 0$ such that for all $\rho \in (0, \rho_\epsilon]$,*

$$\mathcal{D}\left(\lambda_j\,,\ \left((1-\epsilon)\ \rho\ \mu_{\mathbf{\Delta}}^{\|\cdot\|}(CN_j^{\ell_j-1}B)\right)^{\frac{1}{\ell_j}}\right) \subseteq \mathcal{K}_j(\rho) \subseteq \mathcal{D}\left(\lambda_j\,,\ \left((1+\epsilon)\ \rho\ \mu_{\mathbf{\Delta}}^{\|\cdot\|}(CN_j^{\ell_j-1}B)\right)^{\frac{1}{\ell_j}}\right).$$

(vi) *We have*

$$\lim_{\rho\searrow 0} \frac{\mathcal{K}_j(\rho)-\lambda_j}{\rho^{1/\ell_j}} = \mathcal{D}\left(0\,,\ \left(\mu_{\mathbf{\Delta}}^{\|\cdot\|}(CN^{\ell_j-1}B)\right)^{\frac{1}{\ell_j}}\right),$$

where the limit is taken with respect to the Hausdorff metric.

(vii) *The spectral variation satisfies*

$$\mathrm{sv}_{\mathbf{\Delta}}^{\|\cdot\|}(A,B,C,\lambda_j,\rho) = \left(\mu_{\mathbf{\Delta}}^{\|\cdot\|}(CN^{\ell_j-1}B)\right)^{\frac{1}{\ell_j}}\ \rho^{\frac{1}{\ell_j}} + o(\rho^{\frac{1}{\ell_j}}).$$

(viii) *We have*

$$\mathrm{cond}_{\mathbf{\Delta},\,\frac{1}{\ell_j}}^{\|\cdot\|}(A,B,C,\lambda_j) = \left(\mu_{\mathbf{\Delta}}^{\|\cdot\|}(CN^{\ell_j-1}B)\right)^{\frac{1}{\ell_j}}.$$

The proof of Theorem 2.6.6 is based on the lemma below.

Lemma 2.6.7 *Let X be a complex Banach space and $\ell \in \mathbb{N}$. Let U be an open neighborhood of $0 \in \mathbb{C}$ and let $\Gamma : U \setminus \{0\} \to X$ be a continuous function such that the limit $a = \lim_{s\to 0} s^\ell\,\Gamma(s)$ with respect to the norm topology exists. Let $\phi : X \to \mathbb{R}$ be a continuous function satisfying $\phi(a) > 0$ and $\phi(sx) = |s|\phi(x)$ for all $s \in \mathbb{C}$ and all $x \in X$. For $\rho > 0$ set*

$$S_\rho := \{0\} \cup \{\ s \in U\setminus\{0\}\ |\ \ \phi(\Gamma(s)) \geq \rho^{-1}\ \}.$$

Let $\mathcal{K}_\rho$ denote the connected component of S_ρ which contains 0. Then to each $\epsilon > 0$ there is a $\rho_\epsilon > 0$ such that for all $\rho \in (0, \rho_\epsilon]$ the set $\mathcal{K}_\rho$ is compact and satisfies

$$\mathcal{D}(0, R_-(\rho,\epsilon)) \subseteq \mathcal{K}_\rho \subseteq \mathcal{D}(0, R_+(\rho,\epsilon)), \tag{2.31}$$

where $R_\pm(\rho,\epsilon) := (\,(1\pm\epsilon)\rho\,\phi(a)\,)^{\frac{1}{\ell}}$.

Proof: The compactness of $\mathcal{K}_\rho$ is immediate from the continuity of $\phi\circ\Gamma$ and the inclusion (2.31). The latter is a consequence of the following two facts. Let $\epsilon > 0$ be given. Then

(*a*) there is a $\rho_0 > 0$ such that $\mathcal{D}(0, R_-(\rho,\epsilon)) \subseteq S_\rho$ for all $\rho \in (0, \rho_0]$;

(*b*) there is a $\rho_0 > 0$ such that $S_\rho \cap \{\, s \in \mathbb{C}\ |\ \ |s| = R_+(\rho,\epsilon)\ \} = \emptyset$ for all $\rho \in (0, \rho_0]$.

We prove (*a*) and (*b*) by contradiction. Suppose that (*a*) fails. Then there are sequences $\rho_k > 0$ and $s_k \in \mathbb{C}$ satisfying $\lim_{k\to\infty}\rho_k = 0$, $|s_k| \leq R_-(\rho_k,\epsilon)$ and $\phi(\Gamma(s_k)) < \rho_k^{-1}$ for all $k \in \mathbb{N}$. This implies that for every $k \in \mathbb{N}$,

$$\phi(s_k^\ell\Gamma(s_k)) = |s_k|^\ell\phi(\Gamma(s_k)) < |s_k|^\ell\rho_k^{-1} \leq R_-(\rho_k,\epsilon)\rho_k^{-1} = (1-\epsilon)\phi(a). \tag{2.32}$$

Suppose that (*b*) fails. Then there are sequences $\rho_k > 0$ and $s_k \in \mathbb{C}$ satisfying $\lim_{k\to\infty}\rho_k = 0$, $|s_k| = R_+(\rho_k,\epsilon)$ and $\phi(\Gamma(s_k)) \geq \rho_k^{-1}$ for all $k \in \mathbb{N}$. This implies that for every $k \in \mathbb{N}$,

$$\phi(s_k^\ell\Gamma(s_k)) = |s_k|^\ell\phi(\Gamma(s_k)) \geq |s_k|^\ell\rho_k^{-1} = R_+(\rho_k,\epsilon)\rho_k^{-1} = (1+\epsilon)\phi(a). \tag{2.33}$$

In both cases we have that $\lim_{k\to\infty} s_k = 0$ and therefore $\lim_{k\to\infty} \phi(s_k^\ell \Gamma(s_k)) = \phi(a)$. The latter contradicts (2.32) and (2.33). $\square$

Proof of Theorem 2.6.6: It is enough to prove (i) and (v). The claims (ii) and (vi) are then obvious, and the remaining statements follow by combining (i) and (v) with Proposition 2.6.5.

Since $\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A+\lambda I_n, B, C, \rho) = \lambda + \sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A, B, C, \rho)$ for every $\lambda \in \mathbb{C}$, we may assume that $\lambda_j = 0$. Then the formula (2.29) yields for the rational function $G(s) = C(sI_n - A)^{-1}B$ that

$$G(s) = s^{-1}CP_jB + \textstyle\sum_{\ell=1}^{\ell_j-1} s^{-(\ell+1)}CN_j^\ell B + H(s),$$

where

$$H(s) = \sum_{\substack{i\in\underline{r}\\ \lambda_i\neq 0}} \left((s-\lambda_i)^{-1}CP_iB + \textstyle\sum_{\ell=1}^{\ell_i-1}(s-\lambda_i)^{-(\ell+1)}CN_i^\ell B \right).$$

For every $\ell \in \mathbb{N}$ we have $\lim_{s\to 0} s^\ell H(s) = 0$. Thus, the following holds.

(a) If $\ell_j = 1$ then $\lim_{s\to 0} s\,G(s) = CP_jB$.

(b) If $\ell_j \geq 2$ then $\lim_{s\to 0} s^{\ell_j}G(s) = CN_j^{\ell_j-1}B$.

By Theorem 2.5.2 we have

$$\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A, B, C, \rho) = \sigma(A) \cup \{\, s\in\mathbb{C} \mid \ \mu_{\mathbf{\Delta}}^{\|\cdot\|}(G(s)) \geq \rho^{-1} \,\}.$$

Since $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}^{\mathbb{C}}$ the function $\mu_{\mathbf{\Delta}}^{\|\cdot\|}(\cdot)$ is continuous by Corollary 2.4.4 and satisfies $\mu_{\mathbf{\Delta}}^{\|\cdot\|}(sM) = |s|\mu_{\mathbf{\Delta}}^{\|\cdot\|}(M)$ for all $s\in\mathbb{C}$ and $M \in \mathbb{C}^{q\times l}$. Thus the theorem follows by applying Lemma 2.6.7 to the functions $\phi(\cdot) = \mu_{\mathbf{\Delta}}^{\|\cdot\|}(\cdot)$ and $\Gamma(\cdot) = G(\cdot)$ and to the number $\ell = \ell_j$. $\square$

Chapter 3

Semi-algebraicity of spectral value sets

In this chapter we show that spectral value sets and μ-functions are semi-algebraic for a large class of perturbation structures. Semi-algebraic spectral value sets have piecewise analytic boundary, and semi-algebraic μ-functions are real analytic on an open and dense subset of $\mathbb{C}^{q\times l}$.

The structure of this chapter is a follows. In the first section we give a brief introduction to the basics of real algebraic geometry. Here we list the facts and proofs which are needed for the understanding of the main result. We discuss basic properties of semi-algebaic sets and functions and introduce the Tarski-Seidenberg theorem. Then we show that the most important vector and matrix norms are semi-algebraic functions. Finally, in Section 3.2 we present our main result.

For notational convenience we write elements of $\mathbb{R}^n$ as row vectors in the majority of cases.

3.1 Basic facts from real algebraic geometry

3.1.1 Semi-algebraic sets

A semi-algebraic set is a subset of $\mathbb{R}^n$ which can be defined by finitely many polynomial equations and inequalities. The precise definition is as follows.

Definition 3.1.1 *A semi-algebraic subset of $\mathbb{R}^n$ is an element of the Boolean algebra of subsets of $\mathbb{R}^n$ which is generated by the sets*

$$\{\ (x_1,\dots,x_n)\in\mathbb{R}^n \ \mid\ p(x_1,\dots,x_n)>0\ \},\qquad p\in\mathbb{R}[X_1,\dots,X_n],$$

where $\mathbb{R}[X_1,\dots,X_n]$ denotes the ring of real polynomials in the variables $X_1,\dots,X_n$.

It follows immediately from the definition that the sets $\{(x_1,\dots,x_n)\in\mathbb{R}^n \mid p(x_1,\dots,x_n)\bullet 0\}$, where $\bullet\in\{<,>,\geq,\leq,=,\neq\}$, $p\in\mathbb{R}[X_1,\dots,X_n]$, are semi-algebraic. It is then easily seen that a subset S of $\mathbb{R}^n$ is semi-algebraic and only if S is a finite union of sets of the form

$$\{x\in\mathbb{R}^n \mid \tilde{p}_1(x)=0,\dots,\tilde{p}_r(x)=0, q_1(x)>0,\dots,q_s(x)>0\ \}, \tag{3.1}$$
$$\tilde{p}_1,\dots,\tilde{p}_r,q_1,\dots,q_s\in\mathbb{R}[X_1,\dots,X_n].$$

Note that the condition $\tilde{p}_1(x)=\tilde{p}_2(x)=\ldots=\tilde{p}_r(x)=0$ is equivalent to the condition $g(x)=0$, where $g(x)=\sum_{k=1}^{r}(\widetilde{p}_k(x))^2$. If g is the zero polynomial then the set (3.1) is open. If g is not the zero polynomial then it can be factored into a product of nonzero real polynomials, which are irreducible in the ring $\mathbb{R}[X_1, \ldots, X_n]$. We have $g(x) = 0$ if and only if $p(x) = 0$ for at least one of the irreducible factors p of g. Thus the set (3.1) is a finite union of sets of the form

$$\{x \in \mathbb{R}^n \mid p(x) = 0, q_1(x) > 0, \ldots, q_s(x) > 0\}, \tag{3.2}$$

where $p \in \mathbb{R}[X_1, \ldots, X_n]$ is irreducible. Let S_1 denote the set (3.2), and let S_2 be a set of the same type,

$$S_2 = \{x \in \mathbb{R}^n \mid \tilde{p}(x) = 0, \tilde{q}_1(x) > 0, \ldots, \tilde{q}_{\tilde{s}}(x) > 0\}.$$

If $\tilde{p} = cp$ for some $c \in \mathbb{R} \setminus \{0\}$ then $S_1 \cup S_2 = \{x \in \mathbb{R}^n \mid p(x) = 0\} \cap B$, where B is the open set

$$B = \{x \in \mathbb{R}^n \mid (q_1(x) > 0, \ldots, q_s(x) > 0) \text{ or } (\tilde{q}_1(x) > 0, \ldots, \tilde{q}_{\tilde{s}}(x) > 0)\}.$$

We thus obtain the following result.

Lemma 3.1.2 *Each semi-algebraic subset S of $\mathbb{R}^n$ can be written in the form*

$$S = B_0 \cup \bigcup_{j \in \underline{\ell}} (\{x \in \mathbb{R}^n \mid p_j(x) = 0\} \cap B_j), \tag{3.3}$$

where $B_0, \ldots, B_\ell \subseteq \mathbb{R}^n$ are open semi-algebraic sets and $p_1, \ldots, p_\ell \in \mathbb{R}[X_1, \ldots, X_n] \setminus \{0\}$ are irreducible polynomials which are not scalar multiples of each other.

Note that the representation (3.3) is not unique. For instance, the empty set has the representations $\emptyset = \{x \in \mathbb{R}^n \mid \|x\|_2^2 + a^2 = 0\} = \{x \in \mathbb{R}^n \mid -\|x\|_2^2 - a^2 > 0\}$, where $a \neq 0$. Note further, that the set $S \setminus B_0$ is contained in the algebraic set $A = \{x \in \mathbb{R}^n \mid h(x) = 0\}$, where $h(x) = \prod_{j \in \underline{\ell}} p_j(x)$. Since h is not the zero polynomial, A has no interior points.

We now apply the lemma to show a proposition on the smoothness of semi-algebraic sets.

Proposition 3.1.3 *Let S be a semi-algebraic subset of $\mathbb{R}^{n+1}$ such that $\operatorname{int}(S) = \emptyset$. Then there is a polynomial $h \in \mathbb{R}[X_1, \ldots, X_n] \setminus \{0\}$ such that the following holds. For all $(x, y) \in \mathbb{R}^n \times \mathbb{R}$ which satisfy $(x, y) \in S$ and $h(x) \neq 0$ there exist open neighborhoods $V \subset \mathbb{R}^n$ of x and $W \subset \mathbb{R}$ of y and a real analytic function $\phi : V \to W$ such that $S \cap (V \times W) = \operatorname{graph}(\phi)$.*

Proof: Let us recall some basic facts of elimination theory. Let $\mathbb{F}$ be an arbitrary field and consider the polynomials

$$\begin{aligned} p(X, Y) &= a_r(X) Y^r + \ldots + a_1(X) Y + a_0(X), \\ q(X, Y) &= b_s(X) Y^s + \ldots + b_1(X) Y + b_0(X), \end{aligned}$$

where $r, s \in \mathbb{N}$, $X := (X_1, \ldots, X_n)$ and $a_j(X), b_j(X) \in \mathbb{F}[X]$. The Sylvester matrix associated with the polynomials p and q is defined by

$$\mathcal{S}(p, q)(X) := \begin{array}{c} \begin{bmatrix} a_r(X) & \ldots & \ldots & a_0(X) & & \\ & \ddots & & & \ddots & \\ & & a_r(X) & \ldots & \ldots & a_0(X) \\ b_s(X) & \ldots & \ldots & b_0(X) & & \\ & \ddots & & & \ddots & \\ & & b_s(X) & \ldots & \ldots & b_0(X) \end{bmatrix} \end{array} \begin{array}{l} \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} s \text{ rows} \\ \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} r \text{ rows} \end{array}$$

The resultant of p and q with respect to the variable Y is the polynomial $\mathcal{R}(p, q)(X) := \det(\mathcal{S}(p, q)(X)) \in \mathbb{F}[X]$. We need the following properties of the resultant [28, 68].

(*a*) Let $x_0 \in \mathbb{F}^n$ and suppose that $a_r(x_0) \neq 0$. Then $\mathcal{R}(p,q)(x_0) = 0$ if and only if the polynomials $p(x_0, Y), q(x_0, Y) \in \mathbb{F}[Y]$ have a common factor $f(Y) \in \mathbb{F}[Y]$ of degree $\deg f \geq 1$. Thus, if $a_r(x_0) \neq 0$ and $\mathcal{R}(p,q)(x_0) \neq 0$ then the polynomials $p(x_0, Y)$ and $q(x_0, Y)$ have no common zero y.

(*b*) Suppose that $a_r(X) \in \mathbb{F}[X]$ is not the zero polynomial. Then $\mathcal{R}(p,q)(X) \in \mathbb{F}[X]$ equals the zero polynomial if and only if p and q have a common factor $f \in \mathbb{F}[X,Y]$ with $\deg_Y f \geq 1$, where $\deg_Y f$ denotes the degree of f with respect to Y.

Now let S be a nonempty semi-algebraic subset of $\mathbb{R}^{n+1}$ which has no interior points. According to Lemma 3.1.2 we have $S = \bigcup_{j \in \underline{\ell}} (\, B_j \cap \{(x,y) \in \mathbb{R}^n \times \mathbb{R} \mid p_j(x,y) = 0\,\}\,)$, where the sets $B_j, j \in \underline{\ell}$, are open semi-algebraic subsets of $\mathbb{R}^{n+1}$ and the irreducible polynomials

$$p_j(X,Y) = \sum_{i=0}^{r_j} a_{ji}(X) Y^i \in \mathbb{R}[X,Y] \setminus \{0\}, \qquad r_j = \deg_Y p_j.$$

are not scalar multiples of each other. Set $J := \{ j \in \underline{\ell} \mid r_j \neq 0 \}$ and let

$$h(X) := \left(\textstyle\prod_{j \in \underline{\ell}} a_{j r_j}(X) \right) \left(\textstyle\prod_{\substack{j,k \in J \\ j \neq k}} \mathcal{R}(p_j, p_k)(X) \right) \left(\textstyle\prod_{j \in J} \mathcal{R}(p_j, \frac{\partial p_j}{\partial Y})(X) \right).$$

Then h is not the zero polynomial. This follows from (*b*) and the irreducibility of the polynomials p_j. Suppose now that $(x,y) \in S$ and $h(x) \neq 0$. Then (*a*) implies that $p_j(x,y) = 0$ for exactly one $j \in J$. Moreover, there is an open neighborhood $U \subset \mathbb{R}^{n+1}$ of (x,y) such that $S \cap U = p_j^{-1}(0) \cap U$, since otherwise we could find a sequence $(x_i, y_i) \in S$ such that $\lim_{i \to \infty}(x_i, y_i) = (x,y)$ and $p_{k_i}(x_i, y_i) = 0$, where $k_i \neq j$. The latter implies that $p_k(x,y) = 0$ for some $k \in \underline{\ell}$, $k \neq j$. This is impossible since $h(x) \neq 0$. From (*a*) and $h(x) \neq 0$ it follows that $\frac{\partial p_j}{\partial Y}(x,y) \neq 0$, so we can apply the implicit function theorem: there exist open neighborhoods $V \subset \mathbb{R}^n$ of x and $W \subset \mathbb{R}$ of y and a real analytic function $\phi : V \to W$ such that $V \times W \subseteq U$ and $p_j^{-1}(0) \cap (V \times W) = \mathrm{graph}(\phi)$. Thus $S \cap (V \times W) = \mathrm{graph}(\phi)$. □

3.1.2 The Tarski-Seidenberg theorem

A fundamental result in real algebraic geometry is the following so called Tarski-Seidenberg theorem.

Theorem 3.1.4 *Let $\pi : \mathbb{R}^{n+m} = \mathbb{R}^n \oplus \mathbb{R}^m \to \mathbb{R}^n$, $\pi(x,y) := x$, be the canonical projection. Let S is a semi-algebraic subset of $\mathbb{R}^{n+m}$. Then its projection*

$$\pi(S) = \{ x \in \mathbb{R}^n \mid \exists\, y \in \mathbb{R}^m : (x,y) \in S \,\}$$

is also a semi-algebraic subset of $\mathbb{R}^n$.

Proofs of the Tarski-Seidenberg theorem can be found in [87, 81, 11, 4] and in the appendix of [56].

Corollary 3.1.5 *If S is a semi-algebraic subset of $\mathbb{R}^{n+m}$ then the set*

$$\{ x \in \mathbb{R}^n \mid \forall\, y \in \mathbb{R}^m : (x,y) \in S \,\} \;=\; \mathbb{R}^n \setminus \pi(\mathbb{R}^{m+n} \setminus S)$$

is semi-algebraic.

There is an equivalent formulation of the Tarski-Seidenberg theorem in terms of logic. For this formulation we need

Definition 3.1.6 *An atomic formula in the variables* $\xi_1, \ldots, \xi_n$ *is a formula of the form*

$$p(\xi_1, \ldots, \xi_n) \bullet 0,$$

where $\bullet \in \{<, >, \geq, \leq, =, \neq\}$ *and* p *is a polynomial with real coefficients. A polynomial formula* $F(\xi_1, \ldots, \xi_n)$ *is an expression which is built up in a finite number of steps from atomic formulas by means of conjunction, disjunction, negation and the quantifiers* $\forall \xi_k \in \mathbb{R}$, $\exists\, \xi_k \in \mathbb{R}$, $k \in \underline{n}$.

The theorem below, also called Tarski-Seidenberg theorem, is equivalent to Theorem 3.1.4.

Theorem 3.1.7 *Let* $F(\xi_1, \ldots, \xi_n, \eta_1, \ldots, \eta_m)$ *be a polynomial formula in which the variables* $\eta_1, \ldots, \eta_m$ *are bound by quantifiers and* $\xi_1, \ldots, \xi_n$ *are free variables. Then there is a polynomial formula* $G(\xi_1, \ldots, \xi_n)$ *without quantifiers such that the equivalence*

$$F(x_1, \ldots, x_n, y_1, \ldots, y_m) \;\Leftrightarrow\; G(x_1, \ldots, x_n)$$

holds for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$.

An immediate consequence of the Tarski-Seidenberg theorem is

Proposition 3.1.8 *Let* $S \subseteq \mathbb{R}^n$ *is a semi-algebraic set. Then its interior, its closure and its boundary are also semi-algebraic sets.*

Proof: The closure of S satisfies

$$\mathrm{cl}(S) = \{x \in \mathbb{R}^n \mid \underbrace{\forall \epsilon \in \mathbb{R} : (\ \epsilon \leq 0 \text{ or } (\exists y \in \mathbb{R}^n : \|x - y\|_2^2 < \epsilon^2 \text{ and } y \in S\ ))}_{F(x,y,\epsilon)}\ \}.$$

By the Tarski-Seidenberg theorem the formula $F(x, y, \epsilon)$ is equivalent to a polynomial formula $G(x)$ without quantifiers. Thus, $\mathrm{cl}(S)$ is semi-algebraic. The interior and the boundary of S are semi-algebraic since $\mathrm{int}(S) = \mathbb{R}^n \backslash \mathrm{cl}(\mathbb{R}^n \backslash S)$ and $\partial S = \mathrm{cl}(S) \setminus \mathrm{int}(S)$. $\square$

3.1.3 Stratifications of semi-algebraic sets

An analytic stratification of a subset S of $\mathbb{R}^n$ is a decomposition of S into connected real analytic submanifolds of $\mathbb{R}^n$. Lojasiewicz [72] has shown that each nonempty semi-algebraic subset of $\mathbb{R}^n$ admits an analytic stratification with finitely many strata, which are also semi-algebraic. See also [4] for a proof. Consequently, each semi-algebraic set has finitely many semi-algebraic connected components [4, 11]. Since our main interest is in spectral value sets, which are subsets of $\mathbb{C} \cong \mathbb{R}^2$, we discuss semi-algebraic subsets of $\mathbb{R}^2$ in detail.

Proposition 3.1.9 *Let* S *be a semi-algebraic subset of* $\mathbb{R}^2$*. Then* S *can be written as a disjoint union* $S = \mathrm{int}(S) \,\dot\cup\, M \,\dot\cup\, P$*, such that*

(a) M *is either the empty set or a* 1*-dimensional real analytic submanifold of* $\mathbb{R}^2$*;*

(b) P *is a finite set (which may be empty).*

Furthermore, $\operatorname{int}(S)$ *and* M *are semi-algebraic sets and have finitely many connected components, which are also semi-algebraic.*

We sketch the proof for the first statement of the proposition.

Proof: Let $N = S \setminus \operatorname{int}(S)$. Then N is semi-algebraic, and by Lemma 3.1.2 we have

$$N = \bigcup_{j\in\underline{\ell}} \left(\{(x,y)\in\mathbb{R}^2 \mid p_j(x,y) = 0\} \cap B_j\right),$$

where $B_1, \dots, B_\ell$ are open semi-algebraic subsets of $\mathbb{R}^2$ and $p_1, \dots, p_\ell \in \mathbb{R}[X,Y] \setminus \{0\}$ are irreducible polynomials which are not scalar multiples of each other. Using resultants with respect to both the variable X and the variable Y it can be shown that the set

$$\begin{aligned} P &:= \bigcup_{j\in\underline{\ell}} \left\{ (x,y)\in N \;\middle|\; \tfrac{\partial p_j}{\partial X}(x,y) = \tfrac{\partial p_j}{\partial Y}(x,y) = p_j(x,y) = 0 \right\} \\ &\quad \cup \bigcup_{1\le k<j\le\ell} \{ (x,y)\in N \mid p_k(x,y) = p_j(x,y) = 0 \} \end{aligned}$$

contains at most finitely many points. Set $M := N \setminus P$. Let $(x_0,y_0)\in M$. Then $p_j(x_0,y_0) = 0$ for exactly one index $j \in \underline{\ell}$, and there exists an open neighborhood $U \subset \mathbb{R}^2$ of (x_0,y_0) such that $U \cap M = U \cap \left\{ (x,y)\in\mathbb{R}^2 \mid p_j(x,y) = 0 \right\}$. On $U \cap M$ the gradient of p_j does not vanish. Hence $U \cap M$ is an analytic submanifold of $\mathbb{R}^2$. Thus M is an analytic submanifold of $\mathbb{R}^2$. $\square$

Let S and N be as in the proof above. Recall that the boundary of an open subset of a topological space has no interior points. Thus the semi-algebraic set $\partial \operatorname{int}(S)$ has a representation $\partial \operatorname{int}(S) = \bigcup_{j\in\underline{\tilde\ell}} \left(\{(x,y)\in\mathbb{R}^2 \mid \tilde p_j(x,y) = 0\} \cap \tilde B_j\right)$ with irreducible polynomials $\tilde p_j$ and open semi-algebraic sets $\tilde B_j$. The relation $S = \operatorname{int}(S) \cup N$ implies that

$$\begin{aligned} \partial S &\subseteq \partial \operatorname{int}(S) \cup \partial N \\ &\subseteq \bigcup_{j\in\underline{\tilde\ell}} \{(x,y)\in\mathbb{R}^2 \mid \tilde p_j(x,y) = 0\} \cup \bigcup_{j\in\underline{\ell}} \{(x,y)\in\mathbb{R}^2 \mid p_j(x,y) = 0\}. \end{aligned}$$

Thus $\operatorname{int}(\partial S) = \emptyset$. Hence, we have the following corollary to Proposition 3.1.9.

Corollary 3.1.10 *Let* S *be a nonempty semi-algebraic subset of* $\mathbb{R}^2$. *Then* $\partial S = M \,\dot\cup\, P$, *where* P *is a finite set and* M *is either the empty set or a real analytic submanifold of* $\mathbb{R}^2$ *with finitely many connected components.*

3.1.4 Semi-algebraic functions

Let $S_k \subseteq \mathbb{R}^{n_k}$ be non-empty semi-algebraic sets, $k = 1, 2$. A function $f : S_1 \to S_2$ is said to be semi-algebraic if its graph is a semi-algebraic subset of $\mathbb{R}^{n_1+n_2}$. We mention some basic properties of semi-algebraic functions.

1. The composition of semi-algebraic functions is semi-algebraic. This can be seen as follows. Let $S_k \subseteq \mathbb{R}^{n_k}$, $k = 1, 2, 3$, be semi-algebraic sets and let $f : S_1 \to S_2$ and $g : S_2 \to S_3$ be semi-algebraic functions. We have

$$\operatorname{graph}(g \circ f) = \{ (x,y)\in\mathbb{R}^{n_1\times n_3} \mid \underbrace{\exists\, z\in \mathbb{R}^{n_2} : (x,z)\in\operatorname{graph}(f) \text{ and } (z,y)\in\operatorname{graph}(g)}_{F(x,y,z)} \}.$$

   By the Tarski-Seidenberg theorem the expression $F(x,y,z)$ is equivalent to a polynomial formula $G(x,y)$ in which the variables x, y are free.

2. Let $f : S_1 \to S_2$ be a semi-algebraic function, and let $A \subseteq S_1$ ($B \subseteq S_2$) be a semi-algebraic set. Then $f(A)$ ($f^{-1}(B)$) is semi-algebraic. Again, this follows from the Tarski-Seidenberg theorem since

$$\begin{aligned} f(A) &= \{\, y \in \mathbb{R}^{n_2} \mid \exists\, x \in \mathbb{R}^{n_1} : x \in A \text{ and } (x,y) \in \operatorname{graph}(f) \,\}, \\ f^{-1}(B) &= \{\, x \in \mathbb{R}^{n_1} \mid \exists\, y \in \mathbb{R}^{n_2} : y \in B \text{ and } (x,y) \in \operatorname{graph}(f) \,\}. \end{aligned}$$

3. The projections $\mathbb{R}^n \ni (x_1, \ldots, x_n) \overset{\pi_k}{\longmapsto} x_k$, being linear maps, are semi-algebraic. Thus the component functions $f_k = \pi_k \circ f$ of a semi-algebraic function $f : \mathbb{R}^n \supseteq S_1 \to \mathbb{R}^m$ are semi-algebraic. The converse also holds. If $f_1, \ldots, f_n : S_1 \to \mathbb{R}$ are semi-algabraic then $f = (f_1, \ldots f_n) : S_1 \to \mathbb{R}^n$ is semi-algebraic, too.

4. Let $f_1, f_2 : S_1 \to \mathbb{R}$ be semi-algebraic. Then the functions $f_1 + f_2$, $f_1 f_2 : S_1 \to \mathbb{R}$ are semi-algebraic. Moreover, the set $f_2^{-1}(\{0\})$ is semi-algebraic, and if $f_2^{-1}(\{0\}) \neq S_1$ then $\frac{f_1}{f_2} : S_1 \setminus f_2^{-1}(\{0\}) \to \mathbb{R}$ is semi-algebraic.

By means of the Tarski-Seidenberg-Theorem it can be shown that many sets associated with a semi-algebraic function are semi-algebraic. For instance we have the following result which is needed in Section 3.2.

Proposition 3.1.11 *Let $\emptyset \neq S_1 \subset \mathbb{R}^n$ be a semi-algebraic set and let $f : S_1 \to \mathbb{R}$ be a semi-algebraic function. Then the sets*

$$\begin{aligned} M_f &= \{\, x \in S_1 \mid f \text{ has a local minimum at } x \,\}, \\ D_f &= \{\, x \in S_1 \mid f \text{ is discontinuous at } x \,\} \end{aligned}$$

are semi-algebraic.

Proof: We have

$$\begin{aligned} M_f &= \{\, x \in S_1 \mid \exists \epsilon \in \mathbb{R} : \epsilon > 0 \text{ and } (\forall \xi \in \mathbb{R}^n : \xi \notin S_1 \text{ or } \|\xi - x\|_2^2 \geq \epsilon^2 \text{ or } f(\xi) - f(x) \geq 0) \,\}, \\ D_f &= \{\, x \in S_1 \mid \exists \epsilon \in \mathbb{R} : \epsilon > 0 \text{ and} \\ &\qquad \forall \delta \in \mathbb{R} : (\delta = 0 \text{ or } (\exists \xi \in \mathbb{R}^n : \xi \in S_1 \text{ and } \|\xi - x\|_2^2 \leq \delta^2 \text{ and } g(\xi, x, \epsilon) \geq 0)) \,\}, \end{aligned}$$

where $(\xi, x, \epsilon) \mapsto g(\xi, x, \epsilon) = \|f(\xi) - f(x)\|_2^2 - \epsilon^2$ is a semi-algebraic function. □

By applying Proposition 3.1.3 to the graph of a real valued semi-algebraic function, one obtains

Proposition 3.1.12 *Let $f : S_1 \to \mathbb{R}$ be a semi-algebraic function, where S_1 is a semi-algebraic subset of $\mathbb{R}^n$ with nonempty interior. Then there is a polynomial $h \in \mathbb{R}[X_1, \ldots, X_n] \setminus \{0\}$ such that the restriction $f|_{S_1 \setminus h^{-1}(0)}$ is a real analytic function.*

Now, we consider infima and suprema of semi-algebraic functions. To this end we have to extend the definition of semi-algebraicity to functions whose range is a subset of the extended real line $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$.

Definition 3.1.13 *Let S be a semi-algebraic subset of $\mathbb{R}^n$. A function $f : S \to \overline{\mathbb{R}}$ is called semi-algebraic if $f^{-1}(-\infty)$ and $f^{-1}(\infty)$ are semi-algebraic sets and the restriction $f|_{S \setminus f^{-1}(\{-\infty,\infty\})}$ is a semi-algebraic function.*

Proposition 3.1.14 *Let $S \subseteq \mathbb{R}^n \times \mathbb{R}$ be semi-algebraic and let $S_x := \{\, y \in \mathbb{R} \mid (x,y) \in S \,\}$ for all $x \in \mathbb{R}^n$. Then the sets*

$$\begin{aligned} E_S &:= \{\, x \in \mathbb{R}^n \mid S_x \neq \emptyset \,\} \\ A_S &:= \{\, x \in E_S \mid S_x \textit{ is bounded from above} \,\} \\ B_S &:= \{\, x \in E_S \mid S_x \textit{ is bounded from below} \,\} \end{aligned}$$

are semi-algebraic subsets of $\mathbb{R}^n$. The functions

$$\overline{f}_S, \underline{f}_S : \mathbb{R}^n \to \overline{\mathbb{R}} \qquad \overline{f}_S(x) := \sup S_x \qquad \underline{f}_S(x) := \inf S_x$$

are semi-algebraic and we have

$$\overline{f}_S^{-1}(-\infty) = \underline{f}_S^{-1}(\infty) = \mathbb{R}^n \setminus E_S, \qquad \overline{f}_S^{-1}(\infty) = \mathbb{R}^n \setminus A_S, \qquad \underline{f}_S^{-1}(-\infty) = \mathbb{R}^n \setminus B_S. \tag{3.4}$$

Proof: The relations (3.4) are obvious. The other assertions follow from the Tarski-Seidenberg theorem since the sets E_S, S_x are semi-algebraic and we have

$$\begin{aligned} A_S &= E_S \cap \{\, x \in \mathbb{R}^n \mid \exists z \in \mathbb{R} : \forall y \in \mathbb{R} : z \geq y \text{ or } y \notin S_x \,\}, \\ B_S &= E_S \cap \{\, x \in \mathbb{R}^n \mid \exists z \in \mathbb{R} : \forall y \in \mathbb{R} : z \leq y \text{ or } y \notin S_x \,\}, \end{aligned}$$

$$\begin{aligned} \mathrm{graph}(\overline{f}_S) \setminus ((E_S \cup A_S) \times \{-\infty, \infty\}) &= \{(x,z) \in \mathbb{R}^n \times \mathbb{R} \mid \forall y \in \mathbb{R} : (z \geq y \text{ or } y \notin S_x) \\ &\qquad \text{and } \forall \tilde{z} : (\exists\, y \in \mathbb{R} : (y \in S_x \text{ and } y \geq \tilde{z}) \text{ or } \tilde{z} \leq z) \}, \\ \mathrm{graph}(\underline{f}_S) \setminus ((E_S \cup B_S) \times \{-\infty, \infty\}) &= \{(x,z) \in \mathbb{R}^n \times \mathbb{R} \mid \forall y \in \mathbb{R} : (z \leq y \text{ or } y \notin S_x) \\ &\qquad \text{and } \forall \tilde{z} : (\exists\, y \in \mathbb{R} : (y \in S_x \text{ and } y \leq \tilde{z}) \text{ or } \tilde{z} \geq z) \}. \end{aligned}$$

□

Corollary 3.1.15 *If $f : S_1 \times S_2 \to \mathbb{R}$ is semi-algebraic then the functions*

$$\underset{\sim}{f}, \tilde{f} : S_1 \to \overline{\mathbb{R}} \qquad \underset{\sim}{f}(x) := \inf_{y \in S_2} f(x,y), \quad \tilde{f}(x) := \sup_{y \in S_2} f(x,y)$$

are also semi-algebraic.

Proof: We have $\underset{\sim}{f} = \underline{f}_S$ and $\tilde{f} = \bar{f}_S$, where

$$S := \{\, (x,z) \in \mathbb{R}^n \times \mathbb{R} \mid x \in S_1 \text{ and } \exists\, y \in S_2 : ((x,y),z) \in \mathrm{graph}(f) \,\}.$$ □

Corollary 3.1.16 *Let $f : \mathbb{R}^m \to \mathbb{R}$ and $S \subseteq \mathbb{R}^n \times \mathbb{R}^m$ be semi-algebraic. Then the function*

$$F : \mathbb{R}^n \to \overline{\mathbb{R}} \qquad F(x) := \inf \{\, f(y) \mid (x,y) \in S \,\}$$

is semi-algebraic. An analogous statement holds for sup.

Proof: We have $F(x) = \inf \tilde{S}_x$, $x \in \mathbb{R}^n$, where

$$\tilde{S} := \{\, (x,z) \in \mathbb{R}^n \times \mathbb{R} \mid \exists y \in \mathbb{R}^m : (\, (x,y) \in S \text{ and } (y,z) \in \mathrm{graph}(f)) \,\}.$$ □

3.1.5 Semi-algebraicity of norms

The goal of this section is to show that the most important vector and matrix norms are semi-algebraic functions. We first have to extend the definition of semi-algebraicity.

Definition 3.1.17 *Let V be a finite dimensional real vector space and let $\phi : V \to \mathbb{R}^n$ be be a linear isomorphism. A subset S of V is said to be semi-algebraic if $\phi(S)$ is a semi-algebraic subset of $\mathbb{R}^n$. A function $f : V \supseteq S \to \overline{\mathbb{R}}$ is semi-algebraic if the composition $f \circ \phi^{-1} : \phi(S) \to \overline{\mathbb{R}}$ is semi-algebraic. These definitions do not depend on the choice of the isomorphism ϕ. A function between semi-algebraic subsets of finite dimensional real vector spaces V and W is said to be semi-algebraic if its graph is a semi-algebraic subset of $V \times W$.*

In particular semi-algebraicity of subsets of $\mathbb{C}^{m\times n}$ is well defined.

Proposition 3.1.18 *Let $f : [0,\infty)^n \to [0,\infty)$ be a convex and positively homogeneous function such that $f(y) = 0$ if and only if $y = 0$. Suppose that f is semi-algebraic. Then the absolute norm*

$$\|\cdot\| : \mathbb{C}^n \to [0,\infty) \qquad \|x\| := f(|x|) \tag{3.5}$$

is a semi-algebraic function.

Proof: The function $\mathbb{C} \ni z \mapsto |z| = \sqrt{(\Re z)^2 + (\Im z)^2}$ is semi-algebraic. Thus the map $\mathbb{C}^n \ni x \mapsto |x|$ is semi-algebraic, too. Hence, the norm (3.5) is a composition of semi-algebraic functions, and therefore semi-algebraic. □

By specializing the last proposition to the semi-algebraic functions $[0,\infty)^n \ni y \mapsto \max_{k\in\underline{n}} |y_k|$ and $[0,\infty)^n \ni y \mapsto \left(\sum_{k\in\underline{n}} y_k^p\right)^{1/p}$, where $p \geq 1$ is rational, we obtain

Corollary 3.1.19 *The Hölder norms $\|\cdot\|_p : \mathbb{C}^n \to [0,\infty)$ are semi-algebraic functions for all $p \in [1,\infty) \cap \mathbb{Q}$ and $p = \infty$.*

Proposition 3.1.20 *Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$. Suppose that the norms $\|\cdot\|_\alpha : \mathbb{F}^n \to [0,\infty)$ and $\|\cdot\|_\beta : \mathbb{F}^n \to [0,\infty)$ are semi-algebraic functions. Then the operator norm $\|\cdot\|_{\alpha,\beta} : \mathbb{F}^{m\times n} \to [0,\infty)$ is also semi-algebraic.*

Proof: The semi-algebraicity of the norms $\|\cdot\|_\alpha$, $\|\cdot\|_\beta$ yields that the function

$$f : \mathbb{F}^{m\times n} \times (\mathbb{F}^n \backslash \{0\}) \to [0,\infty) \qquad f(A,x) := \frac{\|Ax\|_\beta}{\|x\|_\alpha}$$

is semi-algebraic. Thus the semi-algebraicity of the operator norm $\|\cdot\|_{\alpha,\beta} = \sup_{x\in\mathbb{C}^n\backslash\{0\}} f(\cdot,x)$ follows from Corollary 3.1.15. □

Proposition 3.1.21 *Let $\mathcal{H}_n$ denote set set of all hermitian $n\times n$-matrices, and for $H \in \mathcal{H}_n$ let $\lambda_k(H)$, $k \in \underline{n}$, denote the eigenvalues of H in decreasing order. The functions $\lambda_k(\cdot) : \mathcal{H}_n \to \mathbb{R}$ are semi-algebraic.*

Proof: The sets $\mathcal{H}_n$ and $\mathcal{S}_{n,k} := \{\ X \in \mathbb{C}^{n\times k} \ |\ X^*X = I_k\ \}$, $k \in \underline{n}$, and the functions

$$f_{n,k} : \mathbb{C}^k \setminus \{0\} \times \mathcal{S}_{n,k} \times \mathcal{H}_n \to \mathbb{R} \qquad f_{n,k}(\xi, X, H) := \frac{\xi^* X^* H\, X\xi}{\xi^* X^* X\xi}$$

are semi-algebraic. By the Courant-Fischer max-min principle we have for $H \in \mathcal{H}(n)$ and $k \in \underline{n}$,

$$\lambda_k(H) = \sup_{\dim \mathcal{X} = k} \ \inf_{x \in \mathcal{X}\setminus\{0\}} \frac{x^* H\, x}{x^* x} = \sup_{X \in \mathcal{S}_{n,k}} \ \inf_{\xi \in \mathbb{C}^k \setminus \{0\}} f_{n,k}(\xi, X, H)$$

Thus the proposition follows by applying Corollary 3.1.15 twice. □

Corollary 3.1.22 *The singular value functions $\sigma_k : \mathbb{C}^{m\times n} \to [0, \infty)$, $1 \le k \le \min\{m, n\}$ are semi-algebraic.*

Proof: This follows from the above proposition and the fact that

$$\sigma_k(A) = \lambda_k\left(\begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix}\right) \qquad \text{for all } A \in \mathbb{C}^{m\times n},\ 1 \le k \le \min\{m, n\}.$$ □

Corollary 3.1.23 *Suppose that the absolute and symmetric norm $|\!|\!| \cdot |\!|\!| : \mathbb{R}^{\min\{m,n\}} \to [0, \infty)$ is semi-algebraic. Then the unitarily invariant norm*

$$\|\cdot\| : \mathbb{C}^{m\times n} \to [0, \infty) \qquad \|A\| := |\!|\!|\ [\,\sigma_1(A), \ldots, \sigma_{\min\{m,n\}}(A)\,]^T\ |\!|\!|$$

is semi-algebraic as well.

By combining Corollaries 3.1.23 and 3.1.19 we obtain

Corollary 3.1.24 *The Schatten p-norms $\|\cdot\|_{(p)} : \mathbb{C}^{n\times m} \to [0, \infty)$ are semi-algebraic for all $p \in [1, \infty) \cap \mathbb{Q}$ and $p = \infty$.*

3.2 Main result

The following theorem is a consequence of the Tarski-Seidenberg theorem and the definitions of μ-functions and spectral value sets.

Theorem 3.2.1 *Let $(\boldsymbol{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}$. Suppose that $\boldsymbol{\Delta}$ is a semi-algebraic subset of $\mathbb{C}^{l\times q} \cong \mathbb{R}^{4lq}$ and that $\|\cdot\| : \operatorname{span}_{\mathbb{R}} \boldsymbol{\Delta} \to [0, \infty)$ is a semi-algebraic function. Then*

(i) The spectral value sets $\sigma_{\boldsymbol{\Delta}}^{\|\cdot\|}(A, B, C, \rho)$, where $(A, B, C) = (\,[a_{jk}], [b_{jk}], [c_{jk}]\,) \in L_{n,l,q}$, $\rho > 0$, are semi-algebraic subsets of $\mathbb{C} \cong \mathbb{R}^2$. There exists a polynomial formula F in the real variables

$$\begin{array}{ll} x, y, \rho, & \\ \Re a_{jk}, \Im a_{jk}, & (j,k) \in \underline{n} \times \underline{n}, \\ \Re b_{jk}, \Im b_{jk}, & (j,k) \in \underline{n} \times \underline{l}, \\ \Re c_{jk}, \Im c_{jk}, & (j,k) \in \underline{q} \times \underline{n}, \end{array}$$

without quantifiers *(i.e. all these variables are free variables of F) such that the following equivalence holds for all $x, y \in \mathbb{R}$ and all $(A, B, C, \rho) \in L_{n,l,q} \times [0, \infty)$.*

$$x + iy \in \sigma_{\boldsymbol{\Delta}}^{\|\cdot\|}(A, B, C, \rho) \quad \Leftrightarrow \quad F(\,\rho, x, y, [\Re a_{jk}], [\Im a_{jk}], [\Re b_{jk}], [\Im b_{jk}], [\Re c_{jk}], [\Im c_{jk}]\,).$$

(ii) *The function* $\mu_{\mathbf{\Delta}}^{\|\cdot\|} : \mathbb{C}^{q\times l} \to [0,\infty)$ *is semi-algebraic.*

(iii) *If, in addition,* $l = q = n$ *then the function* $\widetilde{\mu}_{\mathbf{\Delta}}^{\|\cdot\|} : \mathbb{C}^{n\times n} \to [0,\infty]$ *is semi-algebraic.*

Proof: (i). We have $x + iy \in \sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A, B, C, \rho)$ if and only if the following condition $(*)$ holds.

$(*)\quad \exists\, \Delta \in \mathbb{C}^{q\times l} : (\ \Delta \in \mathbf{\Delta}\ \text{and}\ \ \|\Delta\| \le \rho\ \ \text{and}\ \ \det(A + B\Delta C - (x + iy)I_n) = 0\ )$

The semi-algebraicity of $\mathbf{\Delta}$ and $\|\cdot\|$ yields that the condition $(\Delta \in \mathbf{\Delta}$ and $\|\Delta\| \le \rho)$ in $(*)$ can be replaced by an equivalent polynomial formula $F_0(\Re\Delta, \Im\Delta, \rho)$. Thus, claim (i) follows from the Tarski-Seidenberg theorem.

(ii). Let $S = \{\, (\Delta, M) \mid \Delta \in \mathbf{\Delta},\ M \in \mathbb{C}^{q\times l},\ \det(I_l - \Delta M) = 0\, \}$. Then S is semi-algebraic and $\mu_{\mathbf{\Delta}}^{\|\cdot\|}(M)^{-1} = \inf\{\, \|\Delta\| \mid (\Delta, M) \in S\, \}$. By Corollary 3.1.16 the function $\mathbb{C}^{q\times l} \ni M \mapsto \mu_{\mathbf{\Delta}}^{\|\cdot\|}(M)^{-1}$ is semi-algebraic. Thus $\mu_{\mathbf{\Delta}}^{\|\cdot\|}(\cdot)$ is semi-algebraic. The proof of (iii) is analogous.

□

Corollary 3.2.2 *Suppose that the assumptions of Theorem 3.2.1 hold. Then*

(i) *The spectral value sets* $\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A, B, C, \rho)$ *have piecewise analytic boundary. More precisely, there exists a finite set* $P_{A,B,C,\rho} \subset \mathbb{C}$ *such that* $\partial\sigma_{\mathbf{\Delta}}^{\|\cdot\|}(A, B, C, \rho) \setminus P_{A,B,C,\rho}$ *is either empty or a 1-dimensional real analytic submanifold of* $\mathbb{C} \cong \mathbb{R}^2$ *with finitely many connected components.*

(ii) *For each* $(A, B, C) \in L_{n,l,q}$ *the function*

$$f : \mathbb{C}\backslash\sigma(A) \to [0,\infty) \qquad f(s) = \mu_{\mathbf{\Delta}}^{\|\cdot\|}(G(s)), \quad G(s) = C(sI_n - A)^{-1}B$$

is semi-algebraic. The sets $C_f(\rho), Z_f, M_f, D_f$*, defined in Proposition 2.4.6 are semi-algebraic. There exists an open and dense semi-algebraic subset* $S \subset \mathbb{C}$ *such that the restriction* $f|_S$ *is a real analytic function.*

(iii) *There exists an open and dense semi-algebraic subset* $S \subseteq \mathbb{C}^{q\times l} \cong \mathbb{R}^{4lq}$ *such that the restriction* $\mu_{\mathbf{\Delta}}^{\|\cdot\|}|_S$ *is a real analytic function.*

Proof: This follows from Propositions 3.1.11, 3.1.12 and Corollary 3.1.10. □

The reader will have no difficulties in proving the following result on the stability radius.

Theorem 3.2.3 *Let* $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}$*. Suppose that* $\mathbf{\Delta}$ *is a semi-algebraic subset of* $\mathbb{C}^{l\times q} \cong \mathbb{R}^{4lq}$ *and that* $\|\cdot\| : \mathrm{span}_{\mathbb{R}}\, \mathbf{\Delta} \to [0,\infty)$ *is a semi-algebraic function. Let* $\mathbb{C}_g \ne \emptyset$ *be an open semi-algebraic subset of* $\mathbb{C}$*. Then the set* $\mathcal{S} = \{\, A \in \mathbb{C}^{n\times n} \mid A \text{ is } \mathbb{C}_g\text{-stable}\, \}$ *is an open semi-algebraic subset of* $\mathbb{C}^{n\times n}$*. Moreover, the stability radius function* $r_{\mathbf{\Delta}}^{\|\cdot\|}(\cdot, \mathbb{C}_g) : L_{n,l,q} \to [0,\infty]$,

$$(A, B, C) \longmapsto r_{\mathbf{\Delta}}^{\|\cdot\|}(A, B, C, \mathbb{C}_g) \ = \ \inf\{\, \|\Delta\| \ \mid\ \Delta \in \mathbf{\Delta},\ \ \sigma(A + B\Delta C) \not\subset \mathbb{C}_g\ \}$$

is semi-algebraic.

Chapter 4

Spectral value sets of block diagonal matrices

In this chapter we study spectral value sets for structured perturbations $A \rightsquigarrow A_\Delta = A + B\Delta C$ where A, B, C are block diagonal matrices. The perturbations Δ are not necessarily block diagonal but are they assumed to have zero blocks at some prescribed positions. The matrices A_Δ can be interpreted as the system matrices of composite linear systems. We solve the associated μ-problem for two classes of norms. By specializing our results to the diagonal case we obtain a new proof for the eigenvalue inclusion theorem of Brualdi.

4.1 The setting

Let us introduce some necessary notation. Throughout this chapter the letters q, l denote finite sequences $q = (q_1, \dots, q_m)$, $l = (l_1, \dots, l_m)$, where $q_j, l_j, m \in \mathbb{N}$. For $M_j \in \mathbb{C}^{q_j \times l_j}$, $j \in \underline{m}$, we set

$$\bigoplus_{j=1}^m M_j := \operatorname{diag}(M_1, \dots, M_m) = \begin{bmatrix} M_1 & & & 0 \\ & M_2 & & \\ & & \ddots & \\ 0 & & & M_m \end{bmatrix}.$$

For an index set $\mathcal{I} \subseteq \underline{m} \times \underline{m}$ let $\mathbf{\Delta}_{\mathcal{I},q,l}$ denote the set of block matrices Δ of the form

$$\Delta = \begin{bmatrix} \Delta_{11} & \dots & \Delta_{1m} \\ \vdots & & \vdots \\ \Delta_{m1} & \dots & \Delta_{mm} \end{bmatrix}, \quad \Delta_{jk} \in \mathbb{C}^{l_j \times q_k}, \text{ and } \Delta_{jk} = 0 \text{ if } (j,k) \notin \mathcal{I}. \tag{4.1}$$

In the sequel we will abbreviate the left equation in (4.1) by $\Delta = [\Delta_{jk}]$. In this chapter we consider spectral value sets

$$\sigma^{\|\cdot\|}_{\mathbf{\Delta}_{\mathcal{I},q,l}}\Big(\bigoplus_{j=1}^m A_j,\, \bigoplus_{j=1}^m B_j,\, \bigoplus_{j=1}^m C_j,\, \rho\Big), \tag{4.2}$$

where $(A_j, B_j, C_j) \in L_{n_j,l_j,q_j}$ and $\|\cdot\|$ is a norm on the complex vector space $\mathbf{\Delta}_{\mathcal{I},q,l}$. Spectral value sets of this type have the following system theoretic interpretation. Consider the linear sytems

$$\dot{x}_j(t) = A_j x_j(t) + B_j u_j(t), \qquad y_j(t) = C_j x_j(t), \qquad j \in \underline{m},$$

where $u_j(\cdot)$ is the input, $y_j(\cdot)$ is the output and $x_j(\cdot)$ is the state of the jth system. By introducing the couplings

$$u_j(t) = \sum_{k\in\underline{m},\,(j,k)\in\mathcal{I}} \Delta_{jk}\, y_k(t), \qquad \Delta_{jk}\in\mathbb{C}^{l_j\times q_j} \tag{4.3}$$

one obtains the composite systems

$$\begin{bmatrix}\dot{x}_1\\ \vdots\\ \dot{x}_m\end{bmatrix} = A_\Delta \begin{bmatrix}x_1\\ \vdots\\ x_m\end{bmatrix}, \qquad A_\Delta = \left(\bigoplus_{j=1}^m A_j\right) + \left(\bigoplus_{j=1}^m B_j\right)\Delta\left(\bigoplus_{j=1}^m C_j\right), \qquad \Delta\in\boldsymbol{\Delta}_{\mathcal{I},q,l}. \tag{4.4}$$

The pairs $(j,k)\in\mathcal{I}$ can be regarded as the oriented edges of a directed graph $\Gamma(\underline{m},\mathcal{I})$ whose vertices are the numbers $1,\dots,m$. This is illustrated in Figure 4.1. Observe that in Figure 4.1 the endpoint of the edge (j,k) is the first component, j. This orientation reflects the interconnection structure (4.3).

The spectral value set (4.2) is the set of eigenvalues of all system matrices A_Δ with $\Delta\in\boldsymbol{\Delta}_{\mathcal{I},q,l}$ and $\|\Delta\|\le\rho$. For $s\in\mathbb{C}\backslash\sigma(A_j)$ and $j\in\underline{m}$ let $G_j(s) := C_j(sI_{n_j}-A_j)^{-1}B_j$. Then

$$\bigoplus_{j=1}^m G_j(s) = \left(\bigoplus_{j=1}^m C_j\right)\left(sI - \bigoplus_{j=1}^m A_j\right)^{-1}\left(\bigoplus_{j=1}^m B_j\right).$$

Theorem 2.4.5 yields that

$$\begin{aligned}\sigma^{\|\cdot\|}_{\boldsymbol{\Delta}_{\mathcal{I},q,l}}&\left(\bigoplus_{j=1}^m A_j,\, \bigoplus_{j=1}^m B_j,\, \bigoplus_{j=1}^m C_j,\, \rho\right)\\ &= \sigma\left(\bigoplus_{j=1}^m A_j\right) \cup \left\{\, s\in\mathbb{C}\setminus\sigma\left(\bigoplus_{j=1}^m A_j\right) \;\middle|\; \mu^{\|\cdot\|}_{\boldsymbol{\Delta}_{\mathcal{I},q,l}}\left(\bigoplus_{j=1}^m G_j(s)\right) \ge \rho^{-1}\,\right\}\end{aligned} \tag{4.5}$$

Motivated by (4.5) we study the μ-functions $\mu^{\|\cdot\|}_{\boldsymbol{\Delta}_{\mathcal{I},q,l}}(\bigoplus_{j=1}^m M_j)$, where $M_j\in\mathbb{C}^{q_j\times l_j}$, $j\in\underline{m}$. Note that by Proposition 2.4.3,

$$\mu^{\|\cdot\|}_{\boldsymbol{\Delta}_{\mathcal{I},q,l}}(\bigoplus_{j=1}^m M_j) \;=\; \max_{\substack{\Delta\in\boldsymbol{\Delta}_{\mathcal{I},q,l}\\ \|\Delta\|=1}} \varrho\left(\Delta\left(\bigoplus_{j=1}^m M_j\right)\right). \tag{4.6}$$

In Sections 4.2 and 4.3 we provide formulas for the computation of these functions with respect to two classes of norms. In Section 4.4 we discuss the question which eigenvalues of $A = \bigoplus_{j=1}^m A_j$ remain fixed under perturbations of the form $A \rightsquigarrow A_\Delta$, $\Delta\in\boldsymbol{\Delta}_{\mathcal{I},q,l}$. In Section 4.5 we consider the μ-functions for off-diagonal perturbations, i.e. for the index set $\mathcal{I} = \{\,(j,k)\mid j,k\in\underline{m},\ j\ne k\,\}$. Finally, in Sections 4.6-4.8 we relate our results to the classical eigenvalue inclusion theorems of Gershgorin, Brauer and Brualdi.

4.2 The first class of norms

For $A = [a_{jk}]\in\mathbb{C}^{n\times n}$, $n\in\mathbb{N}$, we set

$$\mathcal{I}_A := \{\,(j,k)\in\underline{n}\times\underline{n} \mid a_{jk}\ne 0\,\}.$$

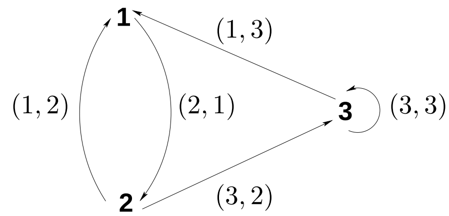
Let $R=[r_{jk}]\in\mathbb{R}^{m\times m}$ be a nonnegative matrix. For $j,k\in\underline{m}$ let $\|\cdot\|_{\alpha_j}$ be a norm on $\mathbb{C}^{q_j}$ and let $\|\cdot\|_{\beta_k}$ be a norm on $\mathbb{C}^{l_k}$. Let $\mathcal{I} := \mathcal{I}_R = \{\,(j,k)\in\underline{m}\times\underline{m}\mid r_{jk}>0\,\}$. To these data we define a weighted maximum norm on $\boldsymbol{\Delta}_{\mathcal{I},q,l}$ by the formula

$$\|\Delta\| := \max_{(j,k)\in\mathcal{I}} r_{jk}^{-1}\|\Delta_{jk}\|_{\alpha_j,\beta_k}, \tag{4.7}$$

index set

$$\mathcal{I} = \{\, (1,2),\ (1,3) \\ (2,1), \\ (3,2), (3,3)\}$$

directed graph $\Gamma(\underline{3}, \mathcal{I})$



interconnection structure

$$\begin{bmatrix} 0 & \Delta_{12} & \Delta_{13} \\ \Delta_{21} & 0 & 0 \\ 0 & \Delta_{32} & \Delta_{33} \end{bmatrix} \in \mathbf{\Delta}_{\mathcal{I},q,l}$$

block diagram

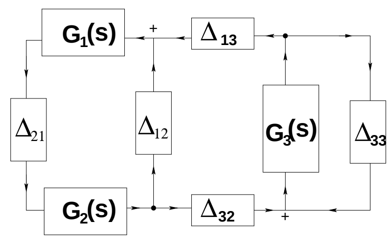


Figure 4.1: Composite systems

where $\Delta = [\Delta_{jk}]$ is a matrix as in (4.1). Note that the following equivalence holds for all $\Delta \in \mathbf{\Delta}_{\mathcal{I},q,l}$.

$$\|\Delta\| \leq 1 \quad \Leftrightarrow \quad \|\Delta_{jk}\|_{\alpha_j,\beta_k} \leq r_{jk} \text{ for all } (j,k) \in \mathcal{I}. \tag{4.8}$$

In this section we show

Theorem 4.2.1 *Let $R = [r_{jk}] \in \mathbb{R}^{m\times m}$ be a nonnegative matrix, and let $M_j \in \mathbb{C}^{q_j \times l_j}$, $j \in \underline{m}$. Then, with respect to the norm* (4.7),

$$\mu^{\|\cdot\|}_{\mathbf{\Delta}_{\mathcal{I},q,l}} \left(\bigoplus_{j=1}^m M_j \right) = \varrho(\, R \operatorname{diag}(\|M_1\|_{\beta_1,\alpha_1}, \ldots, \|M_m\|_{\beta_m,\alpha_m})\,).$$

In order to show Theorem 4.2.1 we need the following well known properties of the spectral radius of nonnegative matrices [61, Corollaries 8.1.19 and 8.3.3 and Theorem 8.3.1].

(ϱ1) For any nonnegative matrix $0 \leq A \in \mathbb{R}^{n\times n}$,

$$\varrho(A) = \max_{\substack{v \in \mathbb{R}^n \setminus \{0\} \\ v \geq 0}} \ \min_{\substack{j \in \underline{n} \\ v_j \neq 0}} \frac{(Av)_j}{v_j},$$

where v_j denotes the jth entry of v and $(Av)_j$ denotes the jth entry of the vector Av. Moreover, there is a nonnegative vector $0 \leq x \in \mathbb{R}^n \setminus \{0\}$ such that $Ax = \varrho(A)x$.

(ϱ2) Let $A_1, A_2 \in \mathbb{R}^{n\times n}$. If $0 \leq A_1 \leq A_2$ then $\varrho(A_1) \leq \varrho(A_2)$.

Observe that property (ϱ2) follows from (ϱ1). In the proof of Theorem 4.2.1 we will apply the lemma below, which is also a consequence of property (ϱ1).

Lemma 4.2.2 *Let $Y_{jk} \in \mathbb{C}^{l_j \times l_k}$, $j,k \in \underline{m}$. Let $\|\cdot\|_{\beta_j}$ be a norm on $\mathbb{C}^{l_j}$. Then*

$$\varrho\left(\begin{bmatrix} Y_{11} & \dots & Y_{1m} \\ \vdots & & \vdots \\ Y_{m1} & \dots & Y_{mm} \end{bmatrix}\right) \leq \varrho\left(\begin{bmatrix} \|Y_{11}\|_{\beta_1,\beta_1} & \dots & \|Y_{1m}\|_{\beta_m,\beta_1} \\ \vdots & & \vdots \\ \|Y_{m1}\|_{\beta_m,\beta_1} & \dots & \|Y_{mm}\|_{\beta_m,\beta_m} \end{bmatrix}\right).$$

Proof: Let $\lambda \in \mathbb{C}$ be an eigenvalue of the block matrix $[Y_{jk}]$, i.e.

$$\begin{bmatrix} Y_{11} & \dots & Y_{1m} \\ \vdots & & \vdots \\ Y_{m1} & \dots & Y_{mm} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}, \qquad x_j \in \mathbb{C}^{l_j}.$$

where $x_j \in \mathbb{C}^{l_j}$, $j \in \underline{m}$, and $x_j \neq 0$ for at least one j. Then $\lambda\, x_j = \sum_{k=1}^m Y_{jk} x_k$ for all $j \in \underline{m}$. This implies

$$|\lambda| \leq \frac{\sum_{k=1}^r \|Y_{jk}\|_{\beta_k,\beta_j}\, \|x_k\|_{\beta_k}}{\|x_j\|_{\beta_j}} \qquad \text{for all } j \text{ with } x_j \neq 0.$$

Thus

$$|\lambda| \leq \max_{\substack{v \in \mathbb{R}^m \setminus \{0\} \\ v \geq 0}} \min_{\substack{j \in \underline{m} \\ v_j \neq 0}} \frac{\sum_{k=1}^m \|Y_{jk}\|_{\beta_k,\beta_j}\, v_k}{v_j} \underset{(\varrho 1)}{=} \varrho\left(\begin{bmatrix} \|Y_{11}\|_{\beta_1,\beta_1} & \dots & \|Y_{1m}\|_{\beta_m,\beta_1} \\ \vdots & & \vdots \\ \|Y_{m1}\|_{\beta_1,\beta_m} & \dots & \|Y_{mm}\|_{\beta_m,\beta_m} \end{bmatrix}\right).$$

$\square$

Proof of Theorem 4.2.1 We already mentioned that by Proposition 2.4.3,

$$\mu^{\|\cdot\|}_{\mathbf{\Delta}_{\mathcal{I},q,l}}(\textstyle\bigoplus_{j=1}^m M_j) \quad = \quad \displaystyle\max_{\substack{\Delta \in \mathbf{\Delta}_{\mathcal{I},q,l} \\ \|\Delta\|=1}} \varrho\left(\Delta(\textstyle\bigoplus_{j=1}^m M_j)\right). \tag{4.9}$$

For the sake of brevity set $P := R\ \mathrm{diag}(\|M_1\|_{\beta_1,\alpha_1}, \dots, \|M_m\|_{\beta_m,\alpha_m}) \in \mathbb{R}^{m \times m}$. Let $\Delta = [\Delta_{jk}] \in \mathbf{\Delta}_{\mathcal{I},q,l}$ with $\|\Delta\| \leq 1$. Then

$$\begin{aligned} \varrho(\ \Delta\,(\textstyle\bigoplus_{k=1}^m M_k)\ ) &= \varrho(\ [\Delta_{jk} M_k]\ ) \\ &\leq \varrho(\ [\ \|\Delta_{jk}\, M_k\|_{\beta_k,\beta_j}\ ]\ ) \qquad \text{(by Lemma 4.2.2)} \\ &\leq \varrho(\ [\ \|\Delta_{jk}\|_{\alpha_k,\beta_j}\, \|M_k\|_{\beta_k,\alpha_k}\ ]\ ) \qquad \text{(by Property } (\varrho 2)) \\ &\leq \varrho(\ [\ r_{jk} \|M_k\|_{\beta_k,\alpha_k}\ ]\ ) \qquad \text{(by (4.8) and Property } (\varrho 2)) \\ &= \varrho(P) \end{aligned}$$

Thus,

$$\max_{\substack{\Delta \in \mathbf{\Delta}_{\mathcal{I},q,l} \\ \|\Delta\|=1}} \varrho\,(\,\Delta\,(\textstyle\bigoplus_{k=1}^m M_k)\,) \leq \varrho(P).$$

It remains to show that the latter inequality is actually an equality. For $k \in \underline{m}$ let $x_k \in \mathbb{C}^{q_k}$ and $y_k \in \mathbb{C}^{l_k}$ be such that $\|x_k\|_{\beta_k} = \|y_k\|^D_{\alpha_k} = 1$ and $y_k^T M_k x_k = \|M_k\|_{\beta_k,\alpha_k}$. Let $\Delta_0 = [\Delta_{jk}]$, where $\Delta_{jk} = r_{jk} x_j y_k^T$. Then $\Delta_0 \in \mathbf{\Delta}_{\mathcal{I},q,l}$ and $\|\Delta_0\| = 1$. Since P is nonnegative there is a nonnegative vector $\xi = [\xi_1, \dots, \xi_m]^T \in \mathbb{R}^m$ such that $P\xi = \varrho(P)\xi$. Define $w = [\xi_1 x_1^T, \dots, \xi_m x_m^T]^T$. Then a straightforward computation yields $\Delta_0\ (\bigoplus_{k=1}^m M_k)\, w = \varrho(P) w$. Thus $\varrho(\Delta_0\ (\bigoplus_{k=1}^m M_k)) \geq \varrho(P)$, and the proof is complete. □

Theorem 4.2.1 yields the following result for spectral value sets of a diagonal matrix with respect to componentwise perturbations of its entries.

Corollary 4.2.3 *Let $A, B, C \in \mathbb{C}^{n\times n}$ be diagonal matrices and let $R \in \mathbb{R}^{n\times n}$ be a nonnegative matrix. Then*

$$\bigcup_{\substack{|\Delta|\le R\\ \Delta\in\mathbb{C}^{n\times n}}} \sigma(A+B\Delta C) \;=\; \sigma(A)\cup\{\, s\in\mathbb{C}\setminus\sigma(A) \mid \;\varrho(\,R\,|B(sI_n-A)^{-1}C|\,)\ge 1\;\}. \quad (4.10)$$

Proof: The set on the left hand side of Equation (4.10) coincides with the spectral value set $\sigma^{\|\cdot\|}_{\mathbf{\Delta}_{\mathcal{I},q,l}}(A,B,C,1)$, where $\mathcal{I}=\mathcal{I}_R$, $l=q=(1,\ldots,1)$ and $\|\Delta\| = \max_{(j,k)\in\mathcal{I}} r_{jk}^{-1}|\delta_{jk}|$ for all $\Delta=[\delta_{jk}]\in\mathbf{\Delta}_{\mathcal{I},q,l}$. Hence, (4.10) is a consequence of Theorem 4.2.1 and Equation (4.5). □

4.3 The second class of norms

Let $\|\cdot\|_\alpha$ be an arbitrary norm on $\mathbb{C}^{m\times m}$ and set $\mathcal{N}_\alpha(\cdot) := \|\cdot\|_{\alpha,\alpha}$. For $j,k\in\underline{m}$ let $\|\cdot\|_{\alpha_j}$ be a norm on $\mathbb{C}^{q_j}$ and let $\|\cdot\|_{\beta_k}$ be a norm on $\mathbb{C}^{l_k}$. To these data we define a norm on $\mathbf{\Delta}_{\mathcal{I},q,l}$ by the formula

$$\|\Delta\| := \mathcal{N}_\alpha\left(\begin{bmatrix} \|\Delta_{11}\|_{\alpha_1,\beta_1} & \cdots & \|\Delta_{1m}\|_{\alpha_m,\beta_1} \\ \vdots & & \vdots \\ \|\Delta_{m1}\|_{\alpha_1,\beta_m} & \cdots & \|\Delta_{mm}\|_{\alpha_m,\beta_m}\end{bmatrix}\right), \quad (4.11)$$

where $\Delta=[\Delta_{jk}]$ is a matrix as in (4.1). In order to determine the corresponding μ-functions we need a graph theoretic notion. A finite sequence $\gamma=(j_1,\ldots,j_\ell)$ of mutually distinct integers is said to be a cycle in the directed graph $\Gamma(\underline{m},\mathcal{I})$ of length $|\gamma|=\ell$ if $(j_i,j_{i+1})\in\mathcal{I}$ for all $i\in\underline{\ell-1}$ and $(j_\ell,j_1)\in\mathcal{I}$. We will write $j\in\gamma$ if $j=j_i$ for some $i\in\underline{\ell}$. By $\mathcal{Z}(\mathcal{I})$ we denote the set of all cycles in $\Gamma(\underline{m},\mathcal{I})$.

The following two theorems are the main results of this section.

Theorem 4.3.1 *Let $M_j\in\mathbb{C}^{q_j\times l_j}$, $j\in\underline{m}$, and $\mathcal{I}\subseteq\underline{m}\times\underline{m}$. If $\mathcal{Z}(\mathcal{I})\ne\emptyset$ then with respect to the norm* (4.11),

$$\mu^{\|\cdot\|}_{\mathbf{\Delta}_{\mathcal{I},q,l}}\left(\bigoplus_{j=1}^m M_j\right) = \max_{\gamma\in\mathcal{Z}(\mathcal{I})}\left(\prod_{j\in\gamma}\|M_j\|_{\beta_j,\alpha_j}\right)^{\frac{1}{|\gamma|}}.$$

Theorem 4.3.2 *Let $M_j\in\mathbb{C}^{q_j\times l_j}$, $j\in\underline{m}$, and $\mathcal{I}\subseteq\underline{m}\times\underline{m}$. Let $\|\cdot\|$ be an arbitrary norm on $\mathbf{\Delta}_{\mathcal{I},q,l}$. If $\mathcal{Z}(\mathcal{I})=\emptyset$ then*

$$\mu^{\|\cdot\|}_{\mathbf{\Delta}_{\mathcal{I},q,l}}\left(\bigoplus_{j=1}^m M_j\right) = 0.$$

For the proof of these theorems we need the two propositions below. Let us introduce a further notation.

Notation 4.3.3 *For $A = [a_{jk}] \in \mathbb{R}^{m\times m}$ let $\mathcal{I}_A := \{ (j,k) \in \underline{m} \times \underline{m} \mid a_{jk} \neq 0 \}$. Let $\gamma = (j_1, \dots, j_\ell) \in \mathcal{Z}(\underline{m} \times \underline{m})$. Then the cycle product over γ is defined as*

$$\prod\nolimits_\gamma A := \prod_{i=1}^{\ell} a_{j_i j_{i+1}}, \qquad \text{where } j_{\ell+1} := j_1.$$

Note that if $\gamma = (j)$ is a cycle of length 1 then $\prod_\gamma A = a_{jj}$. Let $B = [b_{jk}] \in \mathbb{R}^{m\times m}$. Then we denote by $A \circ B$ the Hadamard-product of A and B, $A \circ B = [a_{jk} b_{jk}] \in \mathbb{R}^{m\times m}$.

Proposition 4.3.4 *Let $A = [a_{jk}] \in \mathbb{R}^{m\times m}$ and $\Delta \in \mathbb{R}^{m\times m}$. Suppose that $A, \Delta \geq 0$. If $\mathcal{Z}(\mathcal{I}_A) = \emptyset$ then $\varrho(\Delta \circ A) = 0$. Otherwise we have*

$$\varrho(\Delta \circ A) \leq \varrho(\Delta) \max_{\gamma \in \mathcal{Z}(\mathcal{I}_A)} \left(\prod\nolimits_\gamma A\right)^{\frac{1}{|\gamma|}}.$$

Proof: (following the proof of Theorem 5.7.21 in [62]) If a cycle $\gamma \in \mathcal{Z}(\underline{m} \times \underline{m})$ contains an edge $(j,k) \notin \mathcal{I}_{\mathcal{A}}$ then $\prod_\gamma A = 0$. Thus it suffices to show that

$$\varrho(\Delta \circ A) \leq \varrho(\Delta) \max_{\gamma \in \mathcal{Z}(\underline{m} \times \underline{m})} \left(\prod\nolimits_\gamma A\right)^{\frac{1}{|\gamma|}}. \tag{4.12}$$

Suppose first that $A > 0$. Since the claim is homogeneous in the entries of A we may assume that $\max_{\gamma \in \mathcal{Z}(\underline{m} \times \underline{m})} \left(\prod_\gamma A\right)^{\frac{1}{|\gamma|}} = 1$. We construct a positive diagonal matrix $D = \operatorname{diag}(d_1, \dots, d_m)$ such that $\|DAD^{-1}\|_{\max} \leq 1$, where $\|\cdot\|_{\max}$ denotes the componentwise maximum norm. Set $d_1 = 1$. For $2 \leq k \leq m$ define

$$d_k := \max a_{j_1 j_2} a_{j_2 j_3} \dots a_{j_{\ell-1}, j_\ell},$$

where the maximum is taken over all sequences $j_1, j_2, \dots, j_{\ell-1}, j_\ell \in \underline{m}$ with $j_1 = 1$, $j_\ell = k$ and $\ell \in \mathbb{N}$ arbitrary. The maximum is well defined since only the finite number of sequences without repetition of indices need to be considered, because, by assumption, each cycle product is ≤ 1. The entry in the jth row and kth column of DAD^{-1} is $\frac{d_j a_{jk}}{d_k}$. This quotient is ≤ 1, since the numerator is a sequence product from 1 to k while the denominator is the maximum sequence product from 1 to k. Thus $\|DAD^{-1}\|_{\max} \leq 1$. Now, we have

$$D(\Delta \circ A)D^{-1} = \Delta \circ (DAD^{-1}) \leq \Delta \|DAD^{-1}\|_{\max} \leq \Delta,$$

and therefore $\varrho(\Delta \circ A) = \varrho(D(\Delta \circ A)D^{-1}) \leq \varrho(\Delta)$. Hence, the claim is proved for the case $A > 0$. However, both sides of the inequality (4.12) depend continuously on A. Thus (4.12) holds for all $A \geq 0$. □

In the proof of our main results we use the block cyclic matrices

$$Z(Y_1, \dots, Y_\ell) := \begin{bmatrix} 0 & Y_1 & & & 0 \\ 0 & & Y_2 & & \vdots \\ \vdots & & & \ddots & \vdots \\ 0 & \dots & & & Y_{\ell-1} \\ Y_\ell & \dots & & \dots & 0 \end{bmatrix}, \qquad Y_j \in \mathbb{C}^{m_j \times m_j}.$$

In particular, we need the spectral radius of a positive cyclic matrix with scalar blocks (see [32, chap. 8, § 2]):

Proposition 4.3.5 *Let $c_1, \ldots, c_\ell > 0$, and let $\omega_1, \ldots, \omega_\ell$ denote the ℓth roots of unity, $\omega_k = e^{2\pi i \frac{k-1}{\ell}}$, $k \in \underline{\ell}$. Let $\xi^{(k)} := \left[\xi_1^{(k)}, \ldots, \xi_\ell^{(k)}\right]^T \in \mathbb{C}^\ell$, where*

$$\xi_j^{(k)} := \omega_k^{j-1} \left(\prod_{\nu=1}^{j-1} c_\nu\right)^{-1} \left(\prod_{\nu=1}{}^{\ell} c_\nu\right)^{\frac{j-1}{\ell}}, \qquad j, k \in \underline{\ell}. \tag{4.13}$$

Then the vectors $\xi^{(1)}, \ldots, \xi^{(\ell)}$ form a basis of eigenvectors of $Z(c_1, \ldots, c_\ell)$. More precisely, we have

$$Z(c_1, \ldots, c_\ell)\; \xi^{(k)} = \omega_k \left(\prod_{\nu=1}^{\ell} c_\nu\right)^{1/\ell} \xi^{(k)}, \qquad k \in \underline{\ell}.$$

Therefore,

$$\varrho(\, Z(c_1, \ldots, c_\ell)\,) = \left(\prod_{\nu=1}^{\ell} c_\nu\right)^{1/\ell}.$$

Proposition 4.3.5 can be verified by straightforward computation. For instance, definition (4.13) reads for the case $\ell = 3$,

$$\xi_1^{(k)} = 1, \qquad \xi_2^{(k)} = \omega_k\, c_1^{-1} (c_1 c_2 c_3)^{1/3}, \qquad \xi_3^{(k)} = \omega_k^2\, (c_1 c_2)^{-1} (c_1 c_2 c_3)^{2/3}, \qquad k = 1, 2, 3,$$

where $\omega_1, \omega_2, \omega_3$ are the third roots of unity. Thus,

$$\frac{c_1\, \xi_2^{(k)}}{\xi_1^{(k)}} = \frac{c_2\, \xi_3^{(k)}}{\xi_2^{(k)}} = \frac{c_3\, \xi_1^{(k)}}{\xi_3^{(k)}} = \omega_k\, (c_1 c_2 c_3)^{1/3}.$$

Equivalently,

$$\begin{bmatrix} 0 & c_1 & 0 \\ 0 & 0 & c_2 \\ c_3 & 0 & 0 \end{bmatrix} \begin{bmatrix} \xi_1^{(k)} \\ \xi_2^{(k)} \\ \xi_3^{(k)} \end{bmatrix} = \omega_k\, (c_1 c_2 c_3)^{1/3} \begin{bmatrix} \xi_1^{(k)} \\ \xi_2^{(k)} \\ \xi_3^{(k)} \end{bmatrix}, \qquad k = 1, 2, 3.$$

We are now in the position to give the proofs of the main results of this section.

Proof of Theorems 4.3.1 and 4.3.2. Let

$$c := \begin{cases} 0 & \text{if } \mathcal{Z}(\mathcal{I}) = \emptyset, \\ \max_{\gamma \in \mathcal{Z}(\mathcal{I})} \left(\prod_{j \in \gamma} \|M_j\|_{\beta_j, \alpha_j}\right)^{\frac{1}{|\gamma|}} & \text{otherwise.} \end{cases}$$

We show that

$$\max_{\substack{\Delta \in \boldsymbol{\Delta}_{\mathcal{I},q,l} \\ \|\Delta\|=1}} \varrho\left(\Delta\left(\bigoplus_{j=1}^m M_j\right)\right) \leq c. \tag{4.14}$$

Let $E = [e_{jk}] \in \mathbb{R}^{m \times m}$, where $e_{jk} = 1$ if $(j,k) \in \mathcal{I}$ and $e_{jk} = 0$ otherwise. Set

$$A := E\, \mathrm{diag}(\|M_1\|_{\beta_1, \alpha_1} \ldots \|M_m\|_{\beta_m, \alpha_m}) = [\, e_{jk} \|M_k\|_{\beta_k, \alpha_k}\,] \in \mathbb{R}^{m \times m}.$$

Then $\mathcal{I}_A \subseteq \mathcal{I}$, and for every cycle $\gamma \in \mathcal{Z}(\mathcal{I})$ we have that $\prod_\gamma A = \prod_{j \in \gamma} \|M_j\|_{\beta_j, \alpha_j}$.
If $\gamma \in \mathcal{Z}(\mathcal{I}) \setminus \mathcal{Z}(\mathcal{I}_A)$ then $M_j = 0$ for at least one $j \in \gamma$, and hence $\prod_{j \in \gamma} \|M_j\|_{\beta_j, \alpha_j} = 0$. Thus

$$c = \begin{cases} 0 & \text{if } \mathcal{Z}(\mathcal{I}_A) = \emptyset, \\ \max_{\gamma \in \mathcal{Z}(\mathcal{I}_A)} \left(\prod_\gamma A\right)^{\frac{1}{|\gamma|}} & \text{otherwise.} \end{cases}$$

Let $\Delta = [\Delta_{jk}] \in \mathbf{\Delta}_{\mathcal{I},q,l}$ with $\|\Delta\| = 1$. Then

$$\begin{aligned}
\varrho\left(\Delta\left(\bigoplus_{j=1}^m M_j\right)\right) &= \varrho(\,[\,\Delta_{jk}M_k\,]\,) \\
&\le \varrho(\,[\,\|\Delta_{jk}M_k\|_{\beta_k,\beta_j}\,]\,) \qquad \text{(by Lemma 4.2.2)} \\
&\le \varrho(\,[\,\|\Delta_{jk}\|_{\alpha_j,\beta_k}\,\|M_k\|_{\beta_k,\alpha_k}\,]\,) \qquad \text{(by Property } (\varrho 2)) \\
&= \varrho(\,[\,\|\Delta_{jk}\|_{\alpha_j,\beta_k}\,e_{jk}\,\|M_k\|_{\beta_k,\alpha_k}\,]\,) \qquad (\text{since } \Delta_{jk}=0 \text{ if } e_{jk}=0) \\
&= \varrho\left(\,[\,\|\Delta_{jk}\|_{\alpha_j,\beta_k}\,] \circ A\,\right) \\
&\le \varrho(\,[\,\|\Delta_{jk}\|_{\alpha_j,\beta_k}\,]\,)\,c \qquad \text{(by Proposition 4.3.4)} \\
&\le \underbrace{\mathcal{N}_\alpha(\,[\,\|\Delta_{jk}\|_{\alpha_j,\beta_k}\,]\,)}_{=\|\Delta\|=1}\,c \\
&= c.
\end{aligned}$$

Thus (4.14) is proved. If $c = 0$ then equality holds in (4.14) and therefore $\mu^{\|\cdot\|}_{\mathbf{\Delta}_{\mathcal{I},q,l}}\left(\bigoplus_{j=1}^m M_j\right) = 0$ for all norms on $\mathbf{\Delta}_{\mathcal{I},q,l}$. Thus the proof of Theorem 4.3.2 is complete. It remains to show that for all norms of type (4.11) the inequality (4.14) is actually an equality. To this end we construct, for each cycle $\gamma \in \mathcal{Z}(\mathcal{I}_A)$, a matrix $\Delta_\gamma \in \mathbf{\Delta}_{\mathcal{I},q,l}$ such that $\|\Delta_\gamma\| = 1$ and

$$\varrho\left(\Delta_\gamma\left(\bigoplus_{j=1}^m M_k\right)\right) \ge \left(\prod_{j\in\gamma}\|M_j\|_{\beta_j,\alpha_j}\right)^{\frac{1}{|\gamma|}}. \tag{4.15}$$

The construction of Δ_γ is as follows. Suppose that $\gamma = (j_1,\ldots,j_\ell)$. For $j \in \underline{m}$ let $x_j \in \mathbb{C}^{q_j}$ and $y_j \in \mathbb{C}^{l_j}$ be such that $\|x_j\|_{\beta_j} = \|y_j\|^D_{\alpha_j} = 1$ and $y_j^T M_j x_j = \|M_j\|_{\beta_j,\alpha_j}$. Let $\Delta_\gamma := [\Delta^{(\gamma)}_{jk}]$, where

$$\Delta^{(\gamma)}_{jk} := \begin{cases} x_{j_i}y^T_{j_{i+1}} & \text{if } (j,k) = (j_i, j_{i+1}),\ i \in \underline{\ell-1}, \\ x_{j_\ell}y^T_{j_1} & \text{if } (j,k) = (j_\ell, j_1), \\ 0 \in \mathbb{C}^{l_j\times q_k} & \text{otherwise.}\end{cases}$$

For instance, if m=4,

$$\Delta_{(2)} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & x_2y_2^T & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad \Delta_{(2,4)} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & x_2y_4^T \\ 0 & 0 & 0 & 0 \\ 0 & x_4y_2^T & 0 & 0 \end{bmatrix} = \Delta_{(4,2)},$$

$$\Delta_{(1,2,3)} = \begin{bmatrix} 0 & x_1y_2^T & 0 & 0 \\ 0 & 0 & x_2y_3^T & 0 \\ x_3y_1^T & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad \Delta_{(1,3,2)} = \begin{bmatrix} 0 & 0 & x_1y_3^T & 0 \\ x_2y_1^T & 0 & 0 & 0 \\ 0 & x_3y_2^T & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\Delta_{(1,3.4)} = \begin{bmatrix} 0 & 0 & x_1y_3^T & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & x_3y_4^T \\ x_4y_1^T & 0 & 0 & 0 \end{bmatrix}, \qquad \Delta_{(3,4,2,1)} = \begin{bmatrix} 0 & 0 & x_1y_3^T & 0 \\ x_2y_1^T & 0 & 0 & 0 \\ 0 & 0 & 0 & x_3y_4^T \\ 0 & x_4y_2^T & 0 & 0 \end{bmatrix}.$$

We claim that Δ_γ has the required properties. To see this, note first that in each block row and each block column of Δ_γ there is at most one nonzero block. All nonzero blocks have norm 1. Therefore,

$$1 = \mathcal{N}_\alpha(\,[\,\|\Delta^{(\gamma)}_{jk}\|_{\alpha_j,\beta_k}]\,) = \|\Delta_\gamma\|.$$

Let us show (4.15). Observe that the principal block submatrix of $\Delta_\gamma\left(\bigoplus_{j=1}^m M_k\right)$ corresponding to the block rows and columns with numbers $j_1,\ldots,j_\ell$ is permutation similar to the block cyclic matrix

$$Z(x_{j_1}y_{j_2}^TM_{j_2},x_{j_2}y_{j_3}^TM_{j_3},\ldots,x_{j_\ell}y_{j_1}^TM_{j_1}) = Z(x_{j_1}y_{j_2}^T,x_{j_2}y_{j_3}^T,\ldots,x_{j_\ell}y_{j_1}^T)\left(\bigoplus_{i=1}^\ell M_{j_i}\right).$$

By Proposition 4.3.5 the product $\left(\prod_{j\in\gamma}\|M_j\|_{\beta_j,\alpha_j}\right)^{1/|\gamma|}$ is the spectral radius of the matrix

$$Z(\|M_{j_1}\|_{\beta_{j_1},\alpha_{j_1}},\|M_{j_2}\|_{\beta_{j_2},\alpha_{j_2}},\ldots,\|M_{j_\ell}\|_{\beta_{j_\ell},\alpha_{j_\ell}}).$$

Let $\xi=[\xi_{j_1},\ldots,\xi_{j_\ell}]^T\in\mathbb{R}^\ell$ be a corresponding eigenvector. Then the following relations hold.

$$\|M_{j_i}\|_{\beta_{j_i},\alpha_{j_i}}\xi_{j_{i+1}} = \left(\prod_{j\in\gamma}\|M_j\|_{\beta_j,\alpha_j}\right)^{\frac{1}{|\gamma|}}\xi_{j_i}, \qquad i\in\underline{\ell},\ j_{\ell+1}:=j_1. \tag{4.16}$$

Now, define $w:=[w_1^T,\ldots,w_m^T]^T$, where $w_j=\begin{cases}\xi_{j_i}x_{j_i} & \text{if } j=j_i,\ i\in\underline{\ell}\\ 0\in\mathbb{C}^{q_j} & \text{otherwise.}\end{cases}$

Moreover let $\widehat{w}:=[w_{j_1}^T,w_{j_2}^T,\ldots,w_{j_\ell}^T]^T$. Then

$$\begin{aligned}
&Z(x_{j_1}y_{j_2}^T,x_{j_2}y_{j_3}^T,\ldots,x_{j_\ell}y_{j_1}^T)\left(\bigoplus_{i=1}^\ell M_{j_i}\right)\widehat{w}\\
&= \begin{bmatrix} 0 & x_{j_1}y_{j_2}^TM_{j_2} & & & 0 & \vdots\\ 0 & & x_{j_2}y_{j_3}^TM_{j_3} & & & \\ \vdots & & \ddots & & \vdots & \\ 0 & \ldots & & 0 & x_{j_{\ell-1}}y_{j_\ell}^TM_{j_\ell} & \\ x_{j_\ell}y_{j_1}^TM_{j_1} & 0 & \ldots & \ldots & 0 & \end{bmatrix}\begin{bmatrix}\xi_{j_1}x_{j_1}\\ \xi_{j_2}x_{j_2}\\ \vdots\\ \vdots\\ \xi_{j_\ell}x_{j_\ell}\end{bmatrix}\\
&= \begin{bmatrix}\|M_{j_2}\|_{\beta_{j_2},\alpha_{j_2}}\,\xi_{j_2}\,x_{j_1}\\ \|M_{j_3}\|_{\beta_{j_3},\alpha_{j_3}}\,\xi_{j_3}\,x_{j_2}\\ \vdots\\ \vdots\\ \|M_{j_\ell}\|_{\beta_{j_\ell},\alpha_{j_\ell}}\,\xi_{j_\ell}\,x_{j_{\ell-1}}\\ \|M_{j_1}\|_{\beta_{j_1},\alpha_{j_1}}\,\xi_{j_1}\,x_{j_\ell}\end{bmatrix} \underset{(4.16)}{=} \left(\prod_{j\in\gamma}\|M_j\|_{\beta_j,\alpha_j}\right)^{\frac{1}{|\gamma|}}\begin{bmatrix}\xi_{j_1}\,x_{j_1}\\ \xi_{j_2}\,x_{j_2}\\ \vdots\\ \vdots\\ \xi_{j_{\ell-1}}\,x_{j_{\ell-1}}\\ \xi_{j_\ell}\,x_{j_\ell}\end{bmatrix}\\
&= \left(\prod_{j\in\gamma}\|M_j\|_{\beta_j,\alpha_j}\right)^{\frac{1}{|\gamma|}}\widehat{w}.
\end{aligned}$$

This implies that $\Delta_\gamma\left(\bigoplus_{j=1}^m M_j\right)w=\left(\prod_{j\in\gamma}\|M_j\|_{\beta_j,\alpha_j}\right)^{\frac{1}{|\gamma|}}w$. Thus (4.15) holds, and the proof is complete. □

4.4 Fixed eigenvalues

Let $\mathcal{I}\subseteq\underline{m}\times\underline{m}$ and suppose that $\mathcal{Z}(\mathcal{I})=\emptyset$. Then Theorem 4.3.2 implies that

$$\sigma^{\|\cdot\|}_{\mathbf{\Delta}_{\mathcal{I},q,l}}\left(\bigoplus_{j=1}^m A_j,\ \bigoplus_{j=1}^m B_j,\ \bigoplus_{j=1}^m C_j,\ \rho\right)=\sigma\left(\bigoplus_{j=1}^m A_j\right) \tag{4.17}$$

for all $\rho \geq 0$. In other words: if the graph $\Gamma(\underline{m}, \mathcal{I})$ has no cycles then all eigenvalues remain fixed under perturbations of the form

$$\textstyle\bigoplus_{j=1}^m A_j \rightsquigarrow \bigoplus_{j=1}^m A_j + (\bigoplus_{j=1}^m B_j)\,\Delta\,(\bigoplus_{j=1}^m C_j), \qquad \Delta \in \mathbf{\Delta}_{\mathcal{I},q,l}.$$

The theorem below is a refinement of this result.

Theorem 4.4.1 *Let $\mathcal{I} \subseteq \underline{m} \times \underline{m}$ and let $A_\Delta = \left(\bigoplus_{j=1}^m A_j\right) + \left(\bigoplus_{j=1}^m B_j\right)\Delta\left(\bigoplus_{j=1}^m C_j\right)$, where $(A_j, B_j, C_j) \in L_{n_j,l_j,q_j}$ and $\Delta \in \mathbf{\Delta}_{\mathcal{I},q,l}$. Let $j_0 \in \underline{m}$, and suppose there exists no cycle $\gamma \in \mathcal{Z}(\mathcal{I})$ such that $j_0 \in \gamma$. Then $\sigma(A_{j_0}) \subseteq \sigma(A_\Delta)$.*

Proof: A path in the graph $\Gamma(\underline{m}, \mathcal{I})$ joining the vertices j_0 and $k \in \underline{m}$ is a finite sequence $j_0, j_1, \ldots, j_\nu \in \underline{m}$ such that $j_\nu = k$ and $(j_{i-1}, j_i) \in \mathcal{I}$ for all $i \in \underline{\nu}$. Let V_1 denote the set of indices $k \in \underline{m} \setminus \{j_0\}$ such that there exists a path joining j_0 and k. Set $V_2 := \underline{m} \setminus (\{j_0\} \cup V_1)$. Since j_0 is not contained in a cycle $\gamma \in \mathcal{Z}(\mathcal{I})$ the blocks of $\Delta = [\Delta_{jk}] \in \mathbf{\Delta}_{\mathcal{I},q,l}$ satisfy

$$\begin{aligned} \Delta_{jk} &= 0 \quad \text{for } (j,k) \in (\{j_0\} \cup V_1) \times V_2, \\ \Delta_{jj_0} &= 0 \quad \text{for } j \in \{j_0\} \cup V_1. \end{aligned} \tag{4.18}$$

We have $A_\Delta = [(A_\Delta)_{jk}]$, where $(A_\Delta)_{jk} = \begin{cases} A_j + B_j\Delta_{jj}C_j & \text{if } j = k \\ B_j\Delta_{jk}C_k & \text{if } j \neq k \end{cases}$.

The conditions (4.18) imply

$$\begin{aligned} (A_\Delta)_{jk} &= 0 \qquad \text{for } (j,k) \in (\{j_0\} \cup V_1) \times V_2, \\ (A_\Delta)_{jj_0} &= 0 \qquad \text{for } j \in V_1, \\ (A_\Delta)_{j_0j_0} &= A_{j_0}. \end{aligned} \tag{4.19}$$

Let $\pi : \underline{n} \to \underline{n}$ be a permutation satisfying

$$\pi(1) = j_0, \qquad \pi(\{2, \ldots, 1 + \#V_1\}) = V_1, \qquad \pi(\{2 + \#V_1, \ldots, n\}) = V_2.$$

The block matrix

$$\widetilde{A}_\Delta = \begin{bmatrix} (A_\Delta)_{\pi(1),\pi(1)} & \cdots & (A_\Delta)_{\pi(1),\pi(m)} \\ \vdots & & \vdots \\ (A_\Delta)_{\pi(m),\pi(1)} & \cdots & (A_\Delta)_{\pi(m),\pi(m)} \end{bmatrix}$$

is similar to A_Δ. The relations (4.19) imply that $\widetilde{A}_\Delta$ has the following block structure.

$$\widetilde{A}_\Delta = \begin{bmatrix} A_{j_0} & X_{01} & 0 \\ 0 & X_{11} & 0 \\ X_{20} & X_{21} & X_{22} \end{bmatrix}.$$

If $V_1 = \emptyset$ or $V_2 = \emptyset$ then the corresponding block matrices X_{ij} are not present. It follows that

$$\sigma(A_{j_0}) \subseteq \sigma\left(\begin{bmatrix} A_{j_0} & X_{01} \\ 0 & X_{11} \end{bmatrix}\right) \subseteq \sigma\left(\begin{bmatrix} A_{j_0} & X_{01} \\ 0 & X_{11} \end{bmatrix}\right) \cup \sigma(X_{22}) = \sigma(\widetilde{A}_\Delta) = \sigma(A_\Delta).$$

$\square$

4.5 Off-diagonal perturbation structures

In this section we discuss μ-functions for the index set $\mathcal{I}_{\text{off}} := \{\,(j,k)\in \underline{m}\times\underline{m} \mid\ j\neq k\,\}$. The corresponding perturbation class $\mathbf{\Delta}_{\mathcal{I}_{\text{off}},q,l}$ is the set of all $m\times m$ block matrices $\Delta=[\Delta_{jk}]$ such that $\Delta_{jk}\in\mathbb{C}^{l_j\times q_k}$ and $\Delta_{jj}=0$ for all $j,k\in\underline{m}$. Recall that the following inequality holds for the geometric mean.

Lemma 4.5.1 *Let $c_1\geq c_2\geq\ldots\geq c_\ell\geq 0$. Then, for all $k\in\underline{\ell}$, $\left(\prod_{j=1}^{\ell} c_j\right)^{\frac{1}{\ell}} \leq \left(\prod_{j=1}^{k} c_j\right)^{\frac{1}{k}}$.*

The proposition below is a special case of Theorem 4.3.1.

Proposition 4.5.2 *Let $M_j\in\mathbb{C}^{q_j\times l_j}$, $j\in\underline{m}$. Then with respect to the norm* (4.11),

$$\mu^{\|\cdot\|}_{\mathbf{\Delta}_{\mathcal{I}_{\text{off}},q,\ell}}\left(\bigoplus_{j=1}^m M_j\right) = \max_{1\leq j<k\leq m}\sqrt{\|M_j\|_{\beta_j,\alpha_j}\ \|M_k\|_{\beta_k,\alpha_k}}.$$

Proof: Each pair $(j,k)\in\underline{m}\times\underline{m}$, $j\neq k$, is a cycle in the graph associated with $\mathcal{I}_{\text{off}}$. Thus

$$\max_{1\leq j<k\leq m}\sqrt{\|M_j\|_{\beta_j,\alpha_j}\ \|M_k\|_{\beta_k,\alpha_k}} \leq \max_{\gamma\in\mathcal{Z}(\mathcal{I}_{\text{off}})}\left(\prod_{j\in\gamma}\|M_j\|_{\beta_j,\alpha_j}\right)^{\frac{1}{|\gamma|}}. \tag{4.20}$$

Each cycle $\gamma\in\mathcal{Z}(\mathcal{I}_{\text{off}})$ has length $|\gamma|\geq 2$. Let $\gamma=(j_1,\ldots,j_\ell)$, and let $j_i,j_r\in\gamma$ be such that $i\neq r$ and $\|M_{j_i}\|_{\beta_{j_i},\alpha_{j_i}}\geq\|M_{j_r}\|_{\beta_{j_r},\alpha_{j_r}}\geq\|M_{j_\nu}\|_{\beta_{j_\nu},\alpha_{j_\nu}}$ for all $\nu\in\underline{\ell}\setminus\{i,r\}$. By Lemma 4.5.1 we have that $\left(\prod_{j\in\gamma}\|M_j\|_{\beta_j,\alpha_j}\right)^{\frac{1}{|\gamma|}}\leq\sqrt{\|M_{j_i}\|_{\beta_{j_i},\alpha_{j_i}}\ \|M_{j_r}\|_{\beta_{j_r},\alpha_{j_r}}}$. Thus, equality holds in (4.20). Now, the proposition follows from Theorem 4.3.1. □

An analogous result holds if the underlying norm on $\mathbf{\Delta}_{\mathcal{I}_{\text{off}},q,l}$ is an operator norm subordinate to a Hölder p-norm. More precisely, we have

Theorem 4.5.3 *Let $p\in[1,\infty]$. Let $M_j\in\mathbb{C}^{q_j\times l_j}$, $j\in\underline{m}$. Then*

$$\mu^{(p)}_{\mathbf{\Delta}_{\mathcal{I}_{\text{off}},q,l}}\left(\bigoplus_{j=1}^m M_j\right) = \max_{1\leq j<k\leq m}\sqrt{\|M_j\|_p\ \|M_k\|_p}.$$

For the proof we need the following Lemma.

Lemma 4.5.4 *Let $m_j,n_j\in\mathbb{N}$, $j=1,2$. Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be absolute norms on $\mathbb{C}^{m_1+m_2}$ and $\mathbb{C}^{n_1+n_2}$ respectively. Let $X\in\mathbb{C}^{m_1\times n_2}$, $Y\in\mathbb{C}^{m_2\times n_1}$ and $Z\in\mathbb{C}^{m_2\times n_2}$. Then for all $t\geq 1$,*

$$\left\|\begin{bmatrix}0 & tX\\ tY & Z\end{bmatrix}\right\|_{\beta,\alpha}\leq t\left\|\begin{bmatrix}0 & X\\ Y & Z\end{bmatrix}\right\|_{\beta,\alpha}.$$

Proof: The function $\mathbb{R}\ni\zeta\mapsto f(\zeta):=\left\|\begin{bmatrix}0 & X\\ Y & \zeta Z\end{bmatrix}\right\|_{\beta,\alpha}$ is convex, and since $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ are absolute we have that $f(-\zeta)=\left\|\begin{bmatrix}I_{m_1} & 0\\ 0 & -I_{m_2}\end{bmatrix}\begin{bmatrix}0 & X\\ Y & \zeta Z\end{bmatrix}\begin{bmatrix}-I_{n_1} & 0\\ 0 & I_{n_2}\end{bmatrix}\right\|_{\beta,\alpha}=f(\zeta)$ for all $\zeta\in\mathbb{R}$. From this it follows by convexity that f is a non-decreasing function on $[0,\infty)$. Thus for all $t\geq 1$, $\left\|\begin{bmatrix}0 & tX\\ tY & Z\end{bmatrix}\right\|_{\beta,\alpha}=tf\left(\frac{1}{t}\right)\leq tf(1)=t\left\|\begin{bmatrix}0 & X\\ Y & Z\end{bmatrix}\right\|_{\beta,\alpha}$. □

Proof of Theorem 4.5.3 For brevity set $M := \bigoplus_{j=1}^m M_j$. Without loss of generality we may assume that $\|M_1\|_p \geq \|M_2\|_p \geq \|M_j\|_p$ for $j \geq 3$. Let $\Delta \in \mathbf{\Delta}_{\mathcal{I}_{\rm off},q,l}$. Then $\Delta = \begin{bmatrix} 0 & X \\ Y & Z \end{bmatrix}$ for some $X \in \mathbb{C}^{l_1 \times q'}, Y \in \mathbb{C}^{q_1 \times l'}, Z \in \mathbb{C}^{q' \times l'}$, where $q' = \sum_{j=2}^m q_j$, $l' = \sum_{j=2}^m l_j$. Suppose first that $M_2 = 0$. Then all eigenvalues of $\Delta M = \begin{bmatrix} 0 & 0 \\ YM_1 & 0 \end{bmatrix}$ are zero for all $\Delta \in \mathbf{\Delta}_{\mathcal{I}_{\rm off},q,l}$. Consequently, $\mu^{(p)}_{\mathbf{\Delta}_{\mathcal{I}_{\rm off},q,l}}(M) = 0 = \sqrt{\|M_1\|_p\|M_2\|_p}$. Suppose now that $M_2 \neq 0$ and let

$$F_t := \operatorname{diag}(tI_{q_1}, I_{q'}), \qquad G_t := \operatorname{diag}(tI_{l_1}, I_{l'}), \qquad t := \sqrt{\frac{\|M_1\|_p}{\|M_2\|_p}} \geq 1.$$

Then $\|t^{-2}M_1\|_p = \|M_2\|_p$ and therefore

$$\|F_t^{-1}MG_t^{-1}\|_p = \|\operatorname{diag}\left(t^{-2}M_1, M_2, M_3, \ldots, M_m\right)\|_p = \|M_2\|_p.$$

Lemma 4.5.4 yields that $\|G_t\Delta F_t\|_p \leq t\,\|\Delta\|_p$ for all $\Delta \in \mathbf{\Delta}_{\mathcal{I}_{\rm off},q,l}$. Suppose that $\|\Delta\|_p = 1$. Then we have

$$\begin{aligned}
\varrho(\Delta M) &\leq \|G_t\Delta MG_t^{-1}\|_p \\
&= \|(G_t\Delta F_t)(F_t^{-1}MG_t^{-1})\|_p \\
&\leq \|G_t\Delta F_t\|_p\,\|F_t^{-1}MG_t^{-1}\|_p \\
&= \|G_t\Delta F_t\|_p\,\|M_2\|_p \\
&\leq \|\Delta\|_p\,t\,\|M_2\|_p \\
&= \sqrt{\|M_1\|_p\|M_2\|_p}\,.
\end{aligned}$$

Therefore $\mu^{(p)}_{\mathbf{\Delta}_{\mathcal{I}_{\rm off},q,l}}(M) \leq \sqrt{\|M_1\|_p\|M_2\|_p}$. We show that equality holds. Let $x_j \in \mathbb{C}^{l_j}$, $y_j \in \mathbb{C}^{q_j}$, $j = 1, 2$, be such that $\|x_j\|_p = \|y_j\|_p^D = 1$ and $y_j^T M_j x_j = \|M_j\|_p$, $j = 1, 2$. Define

$$\Delta_0 \;:=\; \operatorname{diag}\left(\begin{bmatrix} 0 & x_1y_2^T \\ x_2y_1^T & 0 \end{bmatrix}, 0\right) \in \mathbf{\Delta}_{\mathcal{I}_{\rm off},q,l}, \qquad x := \begin{bmatrix} \sqrt{\|M_2\|_p}\,x_1 \\ \sqrt{\|M_1\|_p}\,x_2 \\ 0 \end{bmatrix} \in \mathbb{C}^{l_1+\ldots+l_m}.$$

Then $\|\Delta_0\|_p = 1$ and $\Delta_0 Mx = \sqrt{\|M_1\|_p\,\|M_2\|_p}\,x$. Thus $\mu^{(p)}_{\mathbf{\Delta}_{\mathcal{I}_{\rm off},q,l}}(M) \geq \sqrt{\|M_1\|_p\,\|M_2\|_p}\,.$ $\square$

We have not been able to extend Theorem 4.5.3 to arbitrary index sets $\mathcal{I} \subseteq \underline{m} \times \underline{m}$. We formulate our conjecture in this respect as an open question.

Open question: Does the identity

$$\mu^{(p)}_{\mathbf{\Delta}_{\mathcal{I},q,l}}\left(\bigoplus_{j=1}^m M_j\right) = \max_{\gamma \in \mathcal{Z}(\mathcal{I})} \left(\prod_{j\in\gamma} \|M_j\|_p\right)^{\frac{1}{|\gamma|}}$$

hold for arbitrary index sets $\mathcal{I} \subseteq \underline{m} \times \underline{m}$?

4.6 The theorems of Gershgorin, Brauer and Brualdi

In this section we recall the classical eigenvalue inclusion theorems of Gershgorin, Brauer and Brualdi. We show that Brualdi's theorem implies Brauer's and Gershgorin's. Then, in the next section we prove a slight extension of Brualdi's result. For $A = [a_{jk}] \in \mathbb{C}^{n\times n}$ set $R_j(A) := \sum_{\substack{k\in\underline{n}\\ k\neq j}} |a_{jk}|$ and $\mathcal{G}_A := \bigcup_{j\in\underline{n}} \mathcal{D}(a_{jj}, R_j(A))$.

Theorem 4.6.1 (Gershgorin) *Let $A\in\mathbb{C}^{n\times n}$. Then $\sigma(A) \subseteq \mathcal{G}_A$.*

The theorem of Gershgorin has been improved by Brauer [16]. He uses inclusion regions for the eigenvalues of the following type.

$$\mathcal{C}(z_1, z_2, \rho) := \{\, s\in\mathbb{C} \mid\ |s - z_1|\,|s - z_2| \leq \rho \,\} \qquad z_1, z_2\in\mathbb{C},\ \rho \geq 0.$$

The sets $\mathcal{C}(z_1, z_2, \rho)$ and their boundaries are called the ovals of Cassini. The sets $\mathcal{C}(1, -1, \rho)$ are shown in Figure 4.2. For $A = [a_{jk}]\in\mathbb{C}^{n\times n}$ let

$$\mathcal{C}_A := \bigcup_{1\leq j<k\leq n} \mathcal{C}(\, a_{jj},\, a_{kk},\, R_j(A)R_k(A)\,).$$

Theorem 4.6.2 (Brauer) *Let $A\in\mathbb{C}^{n\times n}$. Then $\sigma(A) \subseteq \mathcal{C}_A$.*

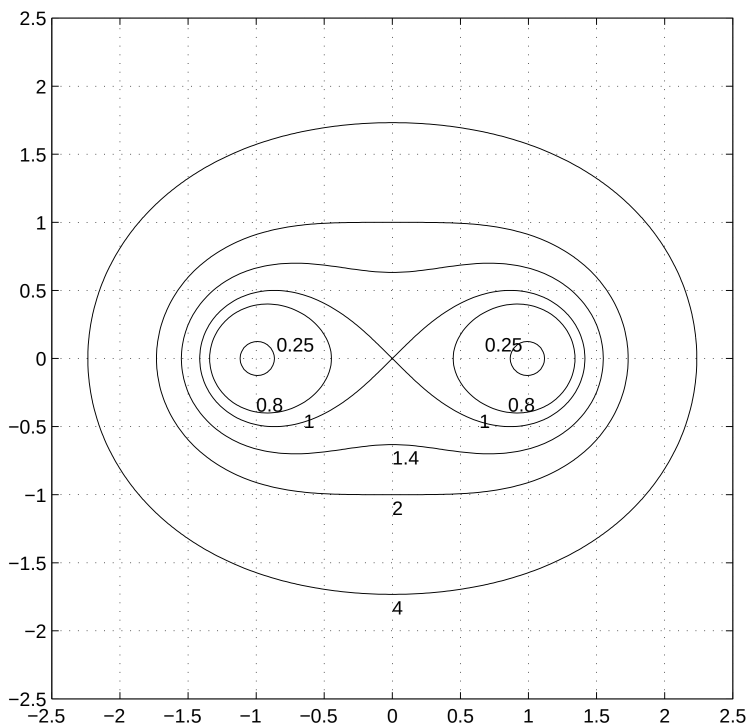


Figure 4.2: The ovals of Cassini $\mathcal{C}(1, -1, \rho)$, $\rho = 0.25, 0.8, 1, 1.4, 2, 4$.

A refinement of Brauer's Theorem is the eigenvalue inclusion theorem of Brualdi [18]. He uses inclusion sets of the form

$$\mathcal{B}\,(z_1, \ldots, z_\ell, \rho) := \left\{\, s\in\mathbb{C} \;\middle|\; \prod_{j\in\underline{\ell}} |s - z_j| \leq \rho \,\right\} \qquad z_1, \ldots, z_\ell\in\mathbb{C},\ \rho \geq 0.$$

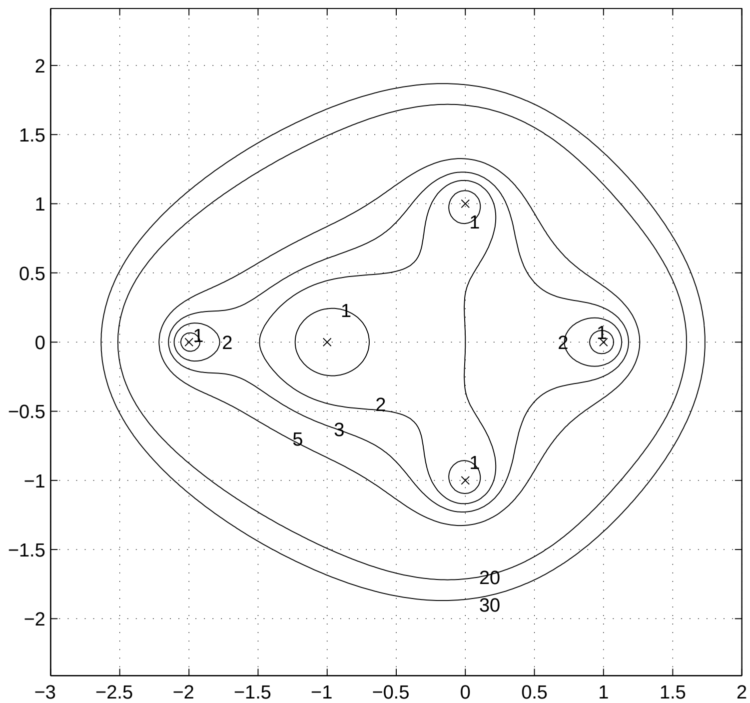


Figure 4.3: The boundaries of the Brualdi sets $\mathcal{B}(-2,-1,1,-i,i,\rho)$, $\rho = 1,2,3,5,20,30$.

Examples of Brualdi sets $\mathcal{B}(z_1,\ldots,z_\ell,\rho)$ are shown in Figure 4.3. For $A = [a_{jk}] \in \mathbb{C}^{n\times n}$ let

$$\mathcal{B}_A := \bigcup_{(j_1,\ldots,j_\ell)\in\mathcal{Z}(\mathcal{I}_A),\ \ell\geq 2} \mathcal{B}\left(a_{j_1j_1},\ldots,a_{j_\ell j_\ell},\ \prod_{i=1}^{\ell} R_{j_i}(A)\right). \tag{4.21}$$

A cycle $\gamma\in\mathcal{Z}(\mathcal{I}_A)$ is said to be nontrivial if $|\gamma| \geq 2$. Note that the union in 4.21 is taken over all nontrivial cycles.

Theorem 4.6.3 (Brualdi) *Let $A\in\mathbb{C}^{n\times n}$. Suppose that to each index $j\in\underline{n}$ there is a nontrivial cycle $\gamma\in\mathcal{Z}(\mathcal{I}_A)$ such that $j\in\gamma$. Then $\sigma(A)\subseteq\mathcal{B}_A$.*

For $A = [a_{jk}]\in\mathbb{C}^{n\times n}$ let

$$\begin{aligned}\sigma_0(A) \ &:= \ \{\ a_{jj}\ |\ j\in\underline{n} \text{ is not contained in a nontrivial cycle } \gamma\in\mathcal{Z}(\mathcal{I}_A)\ \}\\ &= \ \{\ a_{jj}\ |\ \text{if } j\in\gamma\in\mathcal{Z}(\mathcal{I}_A) \text{ then } \gamma = (j)\ \}\end{aligned}$$

In the next section we will show the following extension of Brualdi's theorem.

Theorem 4.6.4 *Let $A\in\mathbb{C}^{n\times n}$. Then $\sigma_0(A)\subseteq\sigma(A)$ and $\sigma(A)\backslash\sigma_0(A)\subseteq\mathcal{B}_A$.*

We close this section with some remarks on Brualdi-sets. First note that $\mathcal{B}(z_1,\rho) = \mathcal{D}(z_1,\rho)$ and $\mathcal{B}(z_1,z_2,\rho) = \mathcal{C}(z_1,z_2,\rho)$. Moreover, the Brualdi-sets satisfy the following condition.

Proposition 4.6.5 *Let $z_1,\ldots,z_\ell\in\mathbb{C}$ and $\rho_1,\ldots,\rho_\ell\geq 0$. Then*

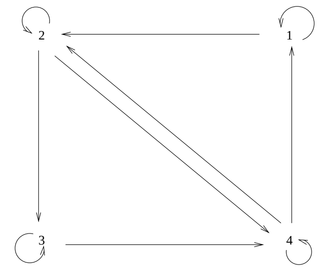
$$\mathcal{B}\left(z_1,\ldots,z_\ell,\prod_{j\in\underline{\ell}}\rho_j\right)\subseteq\bigcup_{k\in\underline{\ell}}\mathcal{B}\left(z_1,\ldots,\widehat{z}_k,\ldots,z_\ell,\prod_{j\in\underline{\ell},j\neq k}\rho_j\right),$$

where $\mathcal{B}\left(z_1,\ldots,\widehat{z}_k,\ldots,z_\ell,\prod_{j\in\underline{\ell},j\neq k}\rho_j\right) = \left\{\ s\in\mathbb{C}\ \middle|\ \prod_{j\in\underline{\ell},j\neq k}|s-z_j|\leq\prod_{j\in\underline{\ell},j\neq k}\rho_j\ \right\}.$

Matrix A

$$A = \begin{bmatrix} 1 & 1.5 & 0 & 0 \\ 0 & i & 0.5 & -0.5i \\ 0 & 0 & -2 & 2 \\ 0.25 & 0.25i & 0 & -2i \end{bmatrix}$$

The graph $\Gamma(\underline{4}, \mathcal{I}_A)$



The eigenvalue inclusion regions $\mathcal{G}_A$, $\mathcal{C}_A$ and $\mathcal{B}_A$

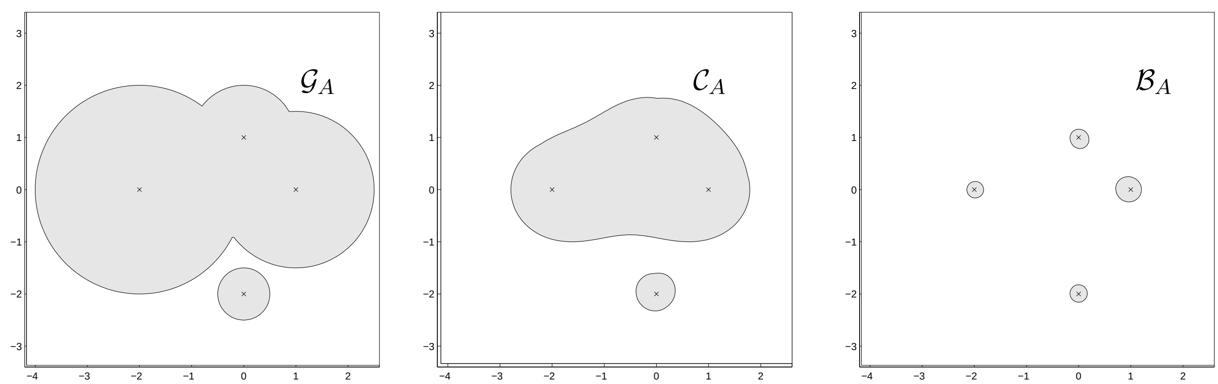


Figure 4.4: Example: the graph associated with $A \in \mathbb{C}^{n\times n}$ and comparison of the regions $\mathcal{G}_A, \mathcal{C}_A$ and $\mathcal{B}_A$. The crosses mark the diagonal elements of A.

Proof: Suppose that $s \notin \bigcup_{k\in\underline{\ell}} \mathcal{B}\left(z_1, \ldots, \widehat{z}_k, \ldots, z_\ell, \prod_{j\in\underline{\ell}, j\neq k} \rho_j\right)$. Then we have for all $k \in \underline{\ell}$, $\prod_{j\in\underline{\ell},j\neq k} |s - z_j| > \prod_{j\in\underline{\ell},j\neq k} \rho_j$. By multiplying these ℓ inequalities we obtain $\left(\prod_{j\in\underline{\ell}} |s - z_j|\right)^{\ell-1} > \left(\prod_{j\in\underline{\ell}} \rho_j\right)^{\ell-1}$. Thus $s \notin \mathcal{B}\left(z_1, \ldots, z_\ell, \prod_{j\in\underline{\ell}} \rho_j\right)$. □

Corollary 4.6.6 *Let $z_1, \ldots, z_\ell \in \mathbb{C}$ and $\rho_1, \ldots, \rho_\ell \geq 0$. Then*

$$\mathcal{B}\left(z_1, \ldots, z_\ell, \prod_{j\in\underline{\ell}} \rho_j\right) \subseteq \bigcup_{1\leq j<k\leq \ell} \mathcal{C}(z_j, z_k, \rho_j\rho_k) \subseteq \bigcup_{j\in\underline{\ell}} \mathcal{D}(z_j, \rho_j).$$

Corollary 4.6.6 implies that for all $A \in \mathbb{C}^{n\times n}$, $n \geq 2$,

$$\mathcal{B}_A \subseteq \mathcal{C}_A \subseteq \mathcal{G}_A. \tag{4.22}$$

Thus the theorems of Brauer and Gershgorin are consequences of Theorem 4.6.4. The first inclusion in (4.22) has been shown by Varga [96]. The second has been shown by Brauer [16].

Each Brualdi-set can be represented as the intersection of a family of sets which are unions of disks. More precisely, we have

Proposition 4.6.7 *Let* $z_1, \dots, z_\ell \in \mathbb{C}$ *and* $\rho > 0$. *Then*

$$\mathcal{B}\left(z_1, \dots, z_\ell, \rho\right) = \bigcap_{\substack{r_1,\dots,r_\ell>0,\\ \prod_{j\in\underline{\ell}} r_j=1}} \left(\bigcup_{j\in\underline{\ell}} \mathcal{D}(z_j, \rho^{\frac{1}{\ell}} r_j) \right). \tag{4.23}$$

Proof: Let D denote the set on the right hand side of (4.23). Suppose that $s \notin D$. Then there are $r_1, \dots, r_\ell > 0$ such that $\prod_{j\in\underline{\ell}} r_j = 1$ and $|s - z_j| > \rho^{\frac{1}{\ell}} r_j$ for all $j \in \underline{\ell}$. Multiplying the latter inequalities we obtain that $\prod_{j\in\underline{\ell}} |s - z_j| > \rho$. Thus $s \notin \mathcal{B}\left(z_1, \dots, z_\ell, \rho\right)$. Hence $\mathcal{B}\left(z_1, \dots, z_\ell, \rho\right) \subseteq D$. The inclusion $D \subseteq \mathcal{B}\left(z_1, \dots, z_\ell, \rho\right)$ is obvious. □

From the relation (4.23) one can derive the following upper bound for the connected components of Brualdi sets.

Proposition 4.6.8 *Let* $z_1, \dots, z_\ell \in \mathbb{C}$ *and* $\rho > 0$. *Then*

(*i*) *Each connected component of* $\mathcal{B}(z_1, \dots, z_\ell, \rho)$ *contains at least one of the points* z_j, $j \in \underline{\ell}$.

(*ii*) *Let* K_j *denote the connected component of* $\mathcal{B}(z_1, \dots, z_\ell, \rho)$ *which contains* z_j. *Let* $\epsilon > 0$ *and suppose that*

$$\min_{k\in\underline{\ell}, k\neq j} |z_j - z_k| > \rho^{\frac{1}{\ell}} \left(\epsilon + \epsilon^{-\frac{1}{\ell-1}}\right). \tag{4.24}$$

Then $K_j \subseteq \mathcal{D}(z_j, \epsilon\rho^{\frac{1}{\ell}})$.

Proof: (*i*). Set $f(s) := \prod_{j\in\underline{\ell}}(s - z_j)$. Let K be a connected component of $\mathcal{B}(z_1, \dots, z_\ell, \rho)$. Then K is compact. Hence there exists $s_0 \in K$ such that $|f(s_0)| = \min_{s\in K} |f(s)|$. Let U be an open neighborhood of s_0 such that $U \cap \mathcal{B}(z_1, \dots, z_\ell, \rho) = U \cap K$. Then $|f(s)| > \rho$ for all $s \in U \setminus K$. Thus $|f(s_0)| = \min_{s\in U} |f(s)|$. The latter implies that $f(s_0) = 0$ since f is holomorphic. Thus $s_0 = z_j$ for some $j \in \underline{\ell}$.
(*ii*). For $i \in \underline{\ell}$ set

$$r_i = \begin{cases} \epsilon & \text{if } i = j, \\ \epsilon^{-\frac{1}{\ell-1}} & \text{otherwise.} \end{cases}$$

Then $\prod_{i\in\underline{\ell}} r_i = 1$ and Proposition 4.6.7 yields that $\mathcal{B}(z_1, \dots, z_\ell, \rho) \subseteq \bigcup_{i\in\underline{\ell}} \mathcal{D}(z_i, \rho^{\frac{1}{\ell}} r_i)$. The condition (4.24) implies that $\mathcal{D}(z_j, \rho^{\frac{1}{\ell}} r_j) \cap \mathcal{D}(z_k, \rho^{\frac{1}{\ell}} r_k) = \emptyset$ for all $k \neq j$. Thus $K_j \subseteq \mathcal{D}(z_j, \rho^{\frac{1}{\ell}} r_j)$. □

Roughly speaking, the above proposition states that if the distance of $z_j \in \mathbb{C}$ from the numbers $z_k \in \mathbb{C}$, $k \neq j$, is large then the connected component K_j of $\mathcal{B}(z_1, \dots, z_\ell, \rho)$ is a small set. This is illustrated in Figure 4.5.

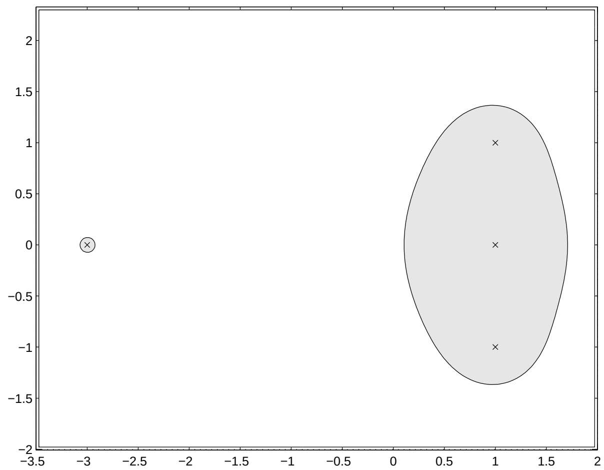


Figure 4.5: The Brualdi set $\mathcal{B}(-3, 1, 1+i, 1-i, \rho)$, $\rho = 5$.

4.7 Proof of Brualdi's theorem

In this section we prove Theorem 4.6.4. It is a consequence of the proposition below, which in turn follows from the results of Section 4.3.

Proposition 4.7.1 *Let $\mathcal{I} \subseteq \underline{n} \times \underline{n}$, and let $\boldsymbol{\Delta}_{\mathcal{I}} = \{\ [\delta_{jk}] \in \mathbb{C}^{n\times n}\ |\ \ \delta_{jk} = 0$ if $(j,k) \notin \mathcal{I}\ \}$. Let $\|\cdot\|_\alpha$ be an arbitray norm on $\mathbb{C}^n$. Let $a_j, b_j, c_j \in \mathbb{C}$, $j \in \underline{n}$. For $\Delta = [\delta_{jk}] \in \boldsymbol{\Delta}_{\mathcal{I}}$ let*

$$\|\Delta\| \ := \ \|\,|\Delta|\,\|_\alpha = \left\| \begin{bmatrix} |\delta_{11}| & \dots & |\delta_{1n}| \\ \vdots & & \vdots \\ |\delta_{n1}| & \dots & |\delta_{nn}| \end{bmatrix} \right\|_\alpha,$$

$$A_\Delta \ := \ \operatorname{diag}(a_1, \dots, a_n) + \operatorname{diag}(b_1, \dots, b_n)\,\Delta\,\operatorname{diag}(c_1, \dots, c_n).$$

Then the following statements hold.

(i) *For all $\rho \geq 0$,*

$$\sigma^{\|\cdot\|}_{\boldsymbol{\Delta}_{\mathcal{I}}}(\operatorname{diag}(a_1, \dots, a_n), \operatorname{diag}(b_1, \dots, b_n), \operatorname{diag}(c_1, \dots, c_n),\ \rho\ )$$

$$= \{a_1, \dots, a_n\} \cup \bigcup_{(j_1,\dots,j_\ell)\in\mathcal{Z}(\mathcal{I})} \mathcal{B}\left(a_{j_1}, \dots, a_{j_\ell},\ \rho^\ell \textstyle\prod_{i=1}^{\ell} |b_{j_i} c_{j_i}|\right). \quad (4.25)$$

(ii) *Let $j_0 \in \underline{n}$ and suppose there exists no cycle $\gamma \in \mathcal{Z}(\mathcal{I})$ such that $j_0 \in \gamma$. Then $a_{j_0} \in \sigma(A_\Delta)$ for all $\Delta \in \boldsymbol{\Delta}_{\mathcal{I}}$.*

Proof: Claim (ii) is a special case of Theorem 4.4.1. Suppose that $\mathcal{Z}(\mathcal{I}) = \emptyset$. Then claim (ii) yields that

$$\sigma^{\|\cdot\|}_{\boldsymbol{\Delta}_{\mathcal{I}}}(\operatorname{diag}(a_1, \dots, a_n), \operatorname{diag}(b_1, \dots, b_n), \operatorname{diag}(c_1, \dots, c_n),\ \rho\ ) = \{a_1, \dots, a_n\}.$$

Hence claim (i) holds for $\mathcal{Z}(\mathcal{I}) = \emptyset$. Suppose now that $\mathcal{Z}(\mathcal{I}) \neq \emptyset$. If $\rho = 0$ then there is nothing to prove. Let $\rho > 0$ and $s \in \mathbb{C} \setminus \{a_1, \ldots, a_n\}$. Then we have the following equivalences.

$$
\begin{aligned}
& s \in \sigma^{\|\cdot\|}_{\mathbf{\Delta}_{\mathcal{I}}}(\, \mathrm{diag}(a_1, \ldots, a_n), \mathrm{diag}(b_1, \ldots, b_n), \mathrm{diag}(c_1, \ldots, c_n);\ \rho\,) \\
& \qquad \Leftrightarrow \quad \mu^{\|\cdot\|}_{\mathbf{\Delta}_{\mathcal{I}}}(\mathrm{diag}(c_1(s-a_1)^{-1}b_1, \ldots, c_n(s-a_n)^{-1}b_n)) \geq \rho^{-1} \\
& \qquad \Leftrightarrow \quad \max_{\gamma \in \mathcal{Z}(\mathcal{I})} \left(\prod_{j\in\gamma} |c_j(s-a_j)^{-1}b_j|\right)^{\frac{1}{|\gamma|}} \geq \rho^{-1} \qquad\qquad \text{(by Theorem 4.3.1)} \\
& \qquad \Leftrightarrow \quad \exists\, \gamma \in \mathcal{Z}(\mathcal{I}) : \left(\prod_{j\in\gamma} |c_j(s-a_j)^{-1}b_j|\right)^{\frac{1}{|\gamma|}} \geq \rho^{-1} \\
& \qquad \Leftrightarrow \quad \exists\, \gamma \in \mathcal{Z}(\mathcal{I}) : \prod_{j\in\gamma} |s-a_j| \leq \rho^{|\gamma|} \prod_{j\in\gamma} |b_j c_j|
\end{aligned}
$$

Hence, (4.25) holds. □

Proof of Theorem 4.6.3 If $A \in \mathbb{C}^{n\times n}$ is a diagonal matrix then there is nothing to prove. Assume that A is nondiagonal and has offdiagonal row sums $R_j(A)$. Set $\Delta = [\delta_{jk}] \in \mathbb{C}^{n\times n}$, where

$$
\delta_{jk} := \begin{cases} R_j(A)^{-1}a_{jk} & \text{if } j \neq k \text{ and } R_j(A) \neq 0, \\ 0 & \text{otherwise.} \end{cases}
$$

Then $A = \mathrm{diag}(a_{11}, \ldots, a_{nn}) + \mathrm{diag}(R_1(A), \ldots, R_n(A))\,\Delta$, and $\|\Delta\|_1 = 1$. Furthermore, we have that $\Delta \in \Delta_{\mathcal{I}}$, where $\mathcal{I} = \mathcal{I}_A \setminus \{(j,j) \mid j \in \underline{n}\,\}$. Thus

$$
\sigma(A) \quad \subseteq \quad \sigma^{\|\cdot\|}_{\mathbf{\Delta}_{\mathcal{I}}}(\, \mathrm{diag}(a_{11}, \ldots, a_{nn}), \mathrm{diag}(R_1(A), \ldots, R_n(A)), I_n,\ 1\,).
$$

Thus Theorem 4.7.1 implies the result. □

4.8 A theorem of Brauer type

Let $A = [a_{jk}] \in \mathbb{C}^{n\times n}$ and $E = A - \mathrm{diag}(a_{11}, \ldots, a_{nn})$. Then the off-diagonal row-sums $R_j(A) = \sum_{\substack{k\in\underline{n} \\ k\neq j}} |a_{jk}|$ satisfy $R_j(A) \leq \|E\|_\infty$. Thus, Brauer's Theorem yields

$$
\sigma(A) \subseteq \bigcup_{1\leq j<k\leq n} \mathcal{C}(\, a_{jj},\, a_{kk},\, \|E\|_\infty^2\,). \tag{4.26}
$$

The theorem below is a consequence of Theorem 4.5.3. It extends (4.26) to all operator norms subordinate to a Hölder p-norm.

Theorem 4.8.1 *Let $A = [a_{jk}] \in \mathbb{C}^{n\times n}$. Let $E = A - \mathrm{diag}(a_{11}, \ldots, a_{nn})$. Let $B = \mathrm{diag}(b_1, \ldots, b_n)$, $C = \mathrm{diag}(c_1, \ldots, c_n)$, where $c_j, b_j > 0$ for all $j \in \underline{n}$. Then we have for every $p \in [1, \infty]$,*

$$
\sigma(A) \subseteq \bigcup_{1\leq j<k\leq n} \mathcal{C}(a_{jj}, a_{kk},\, \|B^{-1}EC^{-1}\|_p^2\, b_j b_k c_j c_k\,).
$$

Proof: Set $\Delta = B^{-1}EC^{-1}$. Then $A = \mathrm{diag}(a_{11}, \ldots, a_{nn}) + B\Delta C$. Thus for all $p \in [1, \infty]$, $\sigma(A) \subseteq \sigma^{(p)}_{\mathbf{\Delta}_{\mathcal{I}_{\mathrm{off}},n,n}}(\mathrm{diag}(a_{11}, \ldots, a_{nn}), B, C, \|\Delta\|_p)$. If $\Delta = 0$ then there is nothing to prove. Let $\Delta \neq 0$ and $s \notin \{a_{11}, \ldots, a_{nn}\}$. Then the following equvalences hold.

$$
\begin{aligned}
& s \in \sigma^{(p)}_{\mathbf{\Delta}_{\mathcal{I}_{\text{off}},n,n}}(\operatorname{diag}(a_{11},\dots,a_{nn}),B,C,\|\Delta\|_p) \\
& \Leftrightarrow \quad \mu^{(p)}_{\mathbf{\Delta}_{\mathcal{I}_{\text{off}},n,n}}(\operatorname{diag}(c_1(s-a_{11})^{-1}b_1,\dots,c_n(s-a_{nn})^{-1}b_n)) \geq \|\Delta\|_p^{-1} \\
& \Leftrightarrow \quad \max_{1\leq j<k\leq n} \sqrt{|c_j(s-a_{jj})^{-1}b_j|\,|c_k(s-a_{kk})^{-1}b_k|} \geq \|\Delta\|_p^{-1} \qquad \text{(by Theorem 4.5.3)} \\
& \Leftrightarrow \quad \min_{1\leq j<k\leq n} (|s-a_{jj}|\,|s-a_{kk}|(b_jb_kc_jc_k)^{-1}) \leq \|\Delta\|_p^2 \\
& \Leftrightarrow \quad \bigcup_{1\leq j<k\leq n} \mathcal{C}(a_{jj},a_{kk},\,\|\Delta\|_p^2\,b_jb_kc_jc_k\,)
\end{aligned}
$$

$\square$

Chapter 5

Complex pseudospectra

5.1 Introduction

This chapter deals with complex pseudospectra. These are spectral value sets of the form

$$\sigma_{\mathbb{C}}^{\|\cdot\|}(A,\rho) = \sigma_{\mathbb{C}}^{\|\cdot\|}(A, I_n, I_n, \rho) = \bigcup_{\substack{\Delta\in\mathbb{C}^{n\times n}\\ \|\Delta\|\le\rho}} \sigma(A+\Delta),$$

where $A\in\mathbb{C}^{n\times n}$. Most of our results are for the case that the underlying norm $\|\cdot\|$ is unitarily invariant or an operator norm subordinate to given vector norms $\|\cdot\|_\alpha$, $\|\cdot\|_\beta$ on $\mathbb{C}^n$. Recall the notation

$$\sigma_{\mathbb{C}}^{(\alpha,\beta)}(A,\rho) = \bigcup_{\substack{\Delta\in\mathbb{C}^{n\times n}\\ \|\Delta\|_{\alpha,\beta}\le\rho}} \sigma(A+\Delta) \qquad \text{and} \qquad \sigma_{\mathbb{C}}^{(\alpha)}(A,\rho) = \sigma_{\mathbb{C}}^{(\alpha,\alpha)}(A,\rho).$$

By Corollary 2.5.4 the μ-functions

$$\begin{aligned} \mu_{\mathbb{C}}^{(\alpha,\beta)}(M) &= \left(\min\{\, \|\Delta\|_{\alpha,\beta} \mid \Delta\in\mathbb{C}^{n\times n},\ 1\in\sigma(\Delta M)\,\}\right)^{-1} \\ \widetilde{\mu}_{\mathbb{C}}^{(\alpha,\beta)}(M) &= \min\{\, \|\Delta\|_{\alpha,\beta} \mid \Delta\in\mathbb{C}^{n\times n},\ A-\Delta \text{ is singular}\,\} \end{aligned}$$

satisfy $\mu_{\mathbb{C}}^{(\alpha,\beta)}(M) = \|M\|_{\beta,\alpha}$ and $\widetilde{\mu}_{\mathbb{C}}^{(\alpha,\beta)}(M) = \inf_{\alpha,\beta}(M)$. Hence Theorem 2.4.5 yields

$$\begin{aligned} \sigma_{\mathbb{C}}^{(\alpha,\beta)}(A,\rho) &= \sigma(A)\cup\{\, s\in\mathbb{C} \mid \|(sI_n-A)^{-1}\|_{\beta,\alpha} \ge \rho^{-1}\,\} && (5.1)\\ &= \{\, s\in\mathbb{C} \mid \inf_{\alpha,\beta}(sI_n-A)\le\rho\,\}. && (5.2) \end{aligned}$$

By Corollary 2.5.5 we have for any normalized unitarily invariant norm that $\mu_{\mathbb{C}}^{\|\cdot\|}(M) = \sigma_{\max}(M)$ and $\widetilde{\mu}_{\mathbb{C}}^{\|\cdot\|}(M) = \sigma_{\min}(M)$. Therefore

$$\begin{aligned} \sigma_{\mathbb{C}}^{\|\cdot\|}(A,\rho) &= \{\, s\in\mathbb{C} \mid \sigma_{\max}((sI_n-A)^{-1}) \ge \rho^{-1}\,\} \\ &= \{\, s\in\mathbb{C} \mid \sigma_{\min}(sI_n-A)\le\rho\,\} \\ &= \sigma_{\mathbb{C}}^{(2)}(A,\rho). && (5.3) \end{aligned}$$

Let us mention some basic invariance properties of pseudospectra. From the equivalence $s\in\sigma(A+\Delta) \Leftrightarrow z+ws\in\sigma((zI_n+wA)+w\Delta)$, where $s,z,w\in\mathbb{C}$, it follows that

$$\sigma_{\mathbb{C}}^{\|\cdot\|}(zI_n+wA, |w|\rho) = z + w\,\sigma_{\mathbb{C}}^{\|\cdot\|}(A,\rho).$$

Let $U \in \mathbb{C}^{n\times n}$ be unitary. Then $\sigma(A+\Delta) = \sigma(U^*(A+\Delta)U) = \sigma(U^*AU + U^*\Delta U)$. If $\|\cdot\|$ is unitarily invariant then $\|U^*\Delta U\| = \|\Delta\|$. Hence,

$$\sigma_{\mathbb{C}}^{\|\cdot\|}(U^*AU, \rho) = \sigma_{\mathbb{C}}^{\|\cdot\|}(A, \rho). \tag{5.4}$$

If the underlying norm $\|\cdot\|$ is an operator norm subordinate to absolute norms then the identities $\|U^*\Delta U\| = \|\Delta\|$ and (5.4) hold for all unitary diagonal matrices U. If the underlying norm satisfies $\|\Delta\| = \|\overline{\Delta}\|$ for all $\Delta \in \mathbb{C}^{n\times n}$ then $\sigma_{\mathbb{C}}^{\|\cdot\|}(\overline{A}, \rho) = \overline{\sigma_{\mathbb{C}}^{\|\cdot\|}(A, \rho)}$.

We now give an outline of the chapter. Let $A \in \mathbb{C}^{n\times n}$ be a normal matrix (i.e. $A^*A = AA^*$) and let $S \in \mathbb{C}^{n\times n}$ be nonsingular. Then $\sigma_{\mathbb{C}}^{\|\cdot\|}(A, \rho) \subseteq \sigma_{\mathbb{C}}^{\|\cdot\|}(SAS^{-1}, \rho)$. Thus, the eigenvalues of SAS^{-1} are at least as sensitive to perturbations as the eigenvalues of A. Moreover, if A is not a scalar multiple of the identity matrix then for fixed $\rho > 0$ the pseudospectrum $\sigma_{\mathbb{C}}^{\|\cdot\|}(SAS^{-1}, \rho)$ can be made arbitrary large by an appropriate choice of S, see Section 5.6. Thus, as is well known, the eigenvalues of a nonnormal matrix can be very sensitive to perturbations. From this observation the question arises what the precise connection between nonnormality of a matrix and the sensitivity of its eigenvalues is. In order to discuss this issue one first has to introduce a measure of nonnormality of a matrix. Various measures have been proposed in the literature, see e.g. [24] and the references therein. In this chapter we concentrate on the so called departure from normality, a measure of nonnormality which has been introduced by Henrici [38]. Among other things we investigate whether the departure from normality fullfills Chatelin's requirement that 'a good measure of nonnormality should reflect how unstable the spectral decomposition is' [22]. In Section 5.3 we obtain a simple upper bound for pseudospectra in terms of the perturbation level and the departure from normality with respect to spectral norm. Then, in Section 5.4 we prove that the pseudospectra of Jordan matrices come arbitrarily close to this upper bound. In Section 5.5 we investigate the pseudospectra of 2×2 matrices. Here we show that a large departure from normality does not guarantee that the eigenvalues are sensitive to perturbations. We briefly discuss some other measures of nonnormality and show that these measures also do not satisfy Chatelin's requirement. Section 5.6 deals with pseudospectra of scaled block triangular matrices. The upper bound for pseudospectra derived in Section 5.3 can be improved if one takes the dimension of the matrix into account. This has been done by Henrici [38]. In Section 5.7 we prove a generalization of Henrici's theorem in terms of pseudospectra. In the last section we extend a theorem of Ruhe [80], which is based on Henrici's result. Ruhe's theorem provides an upper bound for the sensitivity of the spectrum of a matrix A in terms of its departure from normality and the mutual distance of its eigenvalues.

The content of this chapter serves as the background of our consideration concerning real perturbations in Section 7.5.

5.2 Basic upper and lower bounds

This section contains basic upper and lower bounds for pseudospectra. Most of the statements in this section are variants or extensions of results in [25].

In the Proposition below $\mathcal{B}_\rho(K)$ denotes the closed ρ-neighborhood of a closed set $K \subseteq \mathbb{C}$, i.e.

$$\mathcal{B}_\rho(K) := \{\, s \in \mathbb{C} \mid \exists\, s_0 \in K : |s - s_0| \leq \rho \,\}.$$

Proposition 5.2.1 *Let $\|\cdot\|$ be a norm on $\mathbb{C}^{n\times n}$ satisfying $\|I_n\| = 1$. Let $A \in \mathbb{C}^{n\times n}$ and $\rho, \rho_0 \geq 0$. Then*

$$B_\rho\left(\sigma_{\mathbb{C}}^{\|\cdot\|}(A, \rho_0)\right) \subseteq \sigma_{\mathbb{C}}^{\|\cdot\|}(A, \rho_0 + \rho).$$

Proof: Let $s \in \mathcal{B}_\rho\left(\sigma_{\mathbb{C}}^{\|\cdot\|}(A,\rho_0)\right)$. Then there are $s_0 \in \sigma_{\mathbb{C}}^{\|\cdot\|}(A,\rho_0)$ and $z \in \mathbb{C}$ such that $s = s_0 + z$ and $|z| \le \rho$. We have $s_0 \in \sigma(A+\Delta)$ for some $\Delta \in \mathbb{C}^{n\times n}$ with $\|\Delta\| \le \rho_0$. Let $\Delta_1 := \Delta + zI_n$. Then $s \in \sigma(A+\Delta_1)$ and $\|\Delta_1\| \le \|\Delta\| + |z|\|I_n\| \le \rho_0 + |z| \le \rho_0 + \rho$. Thus $s \in \sigma_{\mathbb{C}}^{\|\cdot\|}(A,\rho_0+\rho)$. □

Set in Proposition 5.2.1 $\rho_0 = 0$. Then one obtains the following important lower bound for pseudospectra.

Corollary 5.2.2 *Let $\|\cdot\|$ be a norm on $\mathbb{C}^{n\times n}$ with the property that $\|I_n\| = 1$. Then for all $A \in \mathbb{C}^{n\times n}$ and all $\rho > 0$,*

$$\bigcup_{\lambda \in \sigma(A)} \mathcal{D}(\lambda,\,\rho) \subseteq \sigma_{\mathbb{C}}^{\|\cdot\|}(A,\rho).$$

The lower bound in Corollary 5.2.2 is attained for pseudospectra of diagonal matrices with respect to a large class of operator norms. More precisely, the following holds.

Proposition 5.2.3 *Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be absolute norms on $\mathbb{C}^n$ such that $\|I_n\|_{\alpha,\beta} = 1$ and $\|e_j\|_\alpha = \|e_j\|_\beta$ for all standard basis vectors $e_1,\ldots,e_n$ of $\mathbb{C}^n$. Then we have for all $\lambda_1,\ldots,\lambda_n \in \mathbb{C}$ and every $\rho > 0$,*

$$\sigma_{\mathbb{C}}^{(\alpha,\beta)}(\,\mathrm{diag}(\lambda_1,\ldots,\lambda_n)\,,\,\rho) = \textstyle\bigcup_{j\in\underline{n}}\mathcal{D}(\lambda_j,\rho).$$

Proof: By (5.2) the following equivalence holds for $s \in \mathbb{C}$.

$$\begin{aligned} s \in \sigma_{\mathbb{C}}^{(\alpha,\beta)}(\,\mathrm{diag}(\lambda_1,\ldots,\lambda_n)\,,\,\rho) \quad \Leftrightarrow \quad \rho \quad &\ge \quad \inf\nolimits_{\alpha,\beta}(sI_n - \mathrm{diag}(\lambda_1,\ldots,\lambda_n)) \\ &= \quad \inf\nolimits_{\alpha,\beta}(\mathrm{diag}(s-\lambda_1,\ldots,s-\lambda_n)) \\ &= \quad \min_{j\in\underline{n}} |s-\lambda_j| \qquad \text{(by Proposition 1.5.2).} \end{aligned}$$

□

The next result extends Proposition 5.2.3 to the block diagonal case.

Proposition 5.2.4 *Let $A_j \in \mathbb{C}^{n_j\times n_j}$, $j \in \underline{m}$. Let $1 \le p_1 \le p_2 \le \infty$. Then for any $\rho \ge 0$,*

$$\sigma_{\mathbb{C}}^{(p_2,p_1)}(\mathrm{diag}(A_1,\ldots,A_m),\,\rho\,) = \bigcup_{j\in\underline{m}} \sigma_{\mathbb{C}}^{(p_2,p_1)}(A_j,\rho).$$

Proof: For all $s \in \mathbb{C}$ we have that

$$\begin{aligned} s \in \sigma_{\mathbb{C}}^{(p_2,p_1)}(\,\mathrm{diag}(A_1,\ldots,A_m)\,,\,\rho) \quad \Leftrightarrow \quad & \rho \ge \inf\nolimits_{p_2,p_1}(sI_n - \mathrm{diag}(A_1,\ldots,A_m)) \\ & = \inf\nolimits_{p_2,p_1}(\mathrm{diag}(sI_{n_1}-A_1,\ldots,sI_{n_m}-A_m)) \\ & = \min\nolimits_{j\in\underline{m}} \inf\nolimits_{p_2,p_1}(sI_{n_j}-A_j) \qquad \text{(by Proposition 1.5.3)} \\ \Leftrightarrow \quad & s \in \textstyle\bigcup_{j\in\underline{m}} \sigma_{\mathbb{C}}^{(p_2,p_1)}(A_j,\rho) \end{aligned}$$

□

We now give some basic inclusion principles for pseudospectra.

Proposition 5.2.5 *Let $A \in \mathbb{C}^{n\times n}$ and $\rho > 0$.*

(*i*) *For every norm on* $\mathbb{C}^{n\times n}$ *and every* $E \in \mathbb{C}^{n\times n}$ *we have,*

(*a*) $$\sigma_{\mathbb{C}}^{\|\cdot\|}(A+E,\rho) \subseteq \sigma_{\mathbb{C}}^{\|\cdot\|}(A,\rho+\|E\|),$$

and, if $\|E\| < \rho$,

(*b*) $$\sigma_{\mathbb{C}}^{\|\cdot\|}(A,\rho-\|E\|) \subseteq \sigma_{\mathbb{C}}^{\|\cdot\|}(A+E,\rho).$$

(*ii*) *For all submultiplicative norms on* $\mathbb{C}^{n\times n}$ *and all nonsingular* $S \in \mathbb{C}^{n\times n}$,

$$\sigma_{\mathbb{C}}^{\|\cdot\|}(S^{-1}AS,\rho) \subseteq \sigma_{\mathbb{C}}^{\|\cdot\|}(A,\ \rho\,\|S^{-1}\|\,\|S\|\,).$$

(*iii*) *For every* $B\in\mathbb{C}^{n\times m}, C\in\mathbb{C}^{m\times m}$ *and every* $p_1,p_2\in[1,\infty]$,

$$\sigma_{\mathbb{C}}^{(p_1,p_2)}(A,\rho) \subseteq \sigma_{\mathbb{C}}^{(p_1,p_2)}\left(\begin{bmatrix} A & B \\ 0 & C\end{bmatrix},\rho\right).$$

Proof: (*i*). Suppose that $s\in\sigma_{\mathbb{C}}^{\|\cdot\|}(A+E,\rho)$. Then $s\in\sigma(A+E+\Delta)$ for some $\Delta\in\mathbb{C}^{n\times n}$ with $\|\Delta\| \le \rho$. Let $\widetilde{\Delta} := \Delta + E$. Then $s \in \sigma(A+\widetilde{\Delta})$ and $\|\widetilde{\Delta}\| \le \|\Delta\| + \|E\| \le \rho + \|E\|$. Thus $s\in\sigma_{\mathbb{C}}^{\|\cdot\|}(A,\rho+\|E\|)$, and (*a*) is proved. The proof of (*b*) is analogous. (*ii*) follows from the fact that $\sigma(S^{-1}AS+\Delta) = \sigma(S^{-1}(A+S\Delta S^{-1})S) = \sigma(A+S\Delta S^{-1})$ and $\|S\Delta S^{-1}\| \le \|S\|\|\Delta\|\|S^{-1}\|$. (*iii*). We have for all $\Delta\in\mathbb{C}^{n\times n}$ that $\sigma(A+\Delta) \subseteq \sigma\left(\begin{bmatrix} A+\Delta & B \\ 0 & C\end{bmatrix}\right) = \sigma\left(\begin{bmatrix} A & B \\ 0 & C\end{bmatrix} + \widetilde{\Delta}\right)$, where $\widetilde{\Delta} := \begin{bmatrix} \Delta & 0 \\ 0 & 0\end{bmatrix} \in\mathbb{C}^{(m+n)\times(m+n)}$. It is easily seen that $\|\Delta\|_{p_1,p_2} = \|\widetilde{\Delta}\|_{p_1,p_2}$ for all $p_1,p_2\in[1,\infty]$. $\square$

By combining Proposition 5.2.3 and claim (*i*) of Proposition 5.2.5 one obtains

Corollary 5.2.6 *Let* $\|\cdot\|_\alpha$ *and* $\|\cdot\|_\beta$ *be absolute norms on* $\mathbb{C}^n$ *such that* $\|I_n\|_{\alpha,\beta} = 1$ *and* $\|e_k\|_\alpha = \|e_k\|_\beta$ *for all standard basis vectors* $e_1,\ldots,e_n$ *of* $\mathbb{C}^n$. *Then for all* $A = [a_{jk}]\in\mathbb{C}^{n\times n}$ *and all* $\rho \ge 0$,

$$\sigma_{\mathbb{C}}^{(\alpha,\beta)}(A,\rho) \subseteq \bigcup_{j\in\underline{n}} \mathcal{D}(a_{jj},\ \rho+\|E\|_{\alpha,\beta}\,), \quad \textit{where} \quad E := A - \operatorname{diag}(a_{11},\ldots,a_{nn}).$$

Note that the special case $\alpha = \beta = \infty$, $\rho = 0$ in Corollary 5.2.6 follows from Gershgorin's theorem.

By combining Proposition 5.2.3 and claim (*ii*) of Proposition 5.2.5 one obtains the following well known eigenvalue inclusion theorem of Bauer and Fike [2] (see also [34, Theorem 7.2-2],[10, Theorem VIII.3.1],[86, Theorem IV.3.3]).

Corollary 5.2.7 *Suppose that* $A\in\mathbb{C}^{n\times n}$ *is diagonalizable and let* $S\in\mathbb{C}^{n\times n}$ *be a matrix whose columns form a basis of eigenvectors of* A. *Let* $\|\cdot\| = \|\cdot\|_{\alpha,\alpha}$ *be an operator norm subordinate to an arbitrary norm* $\|\cdot\|_\alpha$ *on* $\mathbb{C}^n$. *Then for all* $\rho \ge 0$,

$$\sigma_{\mathbb{C}}^{\|\cdot\|}(A,\rho) \subseteq \bigcup_{\lambda\in\sigma(A)} \mathcal{D}(\lambda,\ \rho\,\|S^{-1}\|\,\|S\|\,).$$

Corollary 5.2.8 *Let $V \in \mathbb{C}^{n\times k}$ be a matrix whose columns form an orthonormal basis of an invariant subspace of $A\in\mathbb{C}^{n\times n}$. Then we have for every unitarily invariant norm and every $\rho \geq 0$, $\sigma_{\mathbb{C}}^{\|\cdot\|}(V^*AV,\rho) \subseteq \sigma_{\mathbb{C}}^{\|\cdot\|}(A,\rho)$.*

Proof: Without loss of generality we may assume that $\|\cdot\|$ is normalized. Then by (5.3), $\sigma_{\mathbb{C}}^{\|\cdot\|}(A,\rho) = \sigma_{\mathbb{C}}^{(2)}(A,\rho)$. Let $W \in \mathbb{C}^{n\times(n-k)}$ be a matrix whose columns form an orthonormal basis of the orthogonal complement of range(V). Then the matrix $U := \begin{bmatrix} V & W\end{bmatrix}$ is unitary and $U^*AU = \begin{bmatrix} V^*AV & V^*AW \\ W^*AV & W^*AW \end{bmatrix}$. By assumption we have $A(\text{range}(V)) \subseteq \text{range}(V)$ and therefore $W^*AV = 0$. Thus, by claim (*iii*) of Proposition 5.2.5, $\sigma_{\mathbb{C}}^{(2)}(V^*AV,\rho) \subseteq \sigma_{\mathbb{C}}^{(2)}(U^*AU,\rho) = \sigma_{\mathbb{C}}^{(2)}(A,\rho)$. □

5.3 The departure from normality

The departure from normality of a matrix is a measure of nonnormality which has been introduced by Henrici [38]. Its definition is based on the following well known theorem of Schur.

Theorem 5.3.1 *Let $\Lambda\in\mathbb{C}^{n\times n}$ be a diagonal matrix whose diagonal elements are the eigenvalues of $A \in \mathbb{C}^{n\times n}$ (in any order but counting multiplicities). Then there exist a unitary matrix $U\in\mathbb{C}^{n\times n}$ and a strictly upper triangular matrix $R\in\mathbb{C}^{n\times n}$ such that*

$$U^*AU = \Lambda + R.$$

The matrix $\Lambda + R$ is said to be a Schur form of A.

Definition 5.3.2 *Let $\mathcal{R}(A)$ denote the set of all strictly upper triangular matrices $R\in\mathbb{C}^{n\times n}$ which occur in a Schur form of $A\in\mathbb{C}^{n\times n}$. Furthermore let $\|\cdot\|$ be a unitarily invariant norm. Then*

$$\delta_{\|\cdot\|}(A) \quad := \quad \inf_{R\in\mathcal{R}(A)} \|R\| \tag{5.5}$$

is called the departure from normality of A with respect to the norm $\|\cdot\|$.

The set $\mathcal{R}(A)$ is closed and we have $\mathcal{R}(A) = \{0_{n\times n}\}$ if and only if A is normal, where $0_{n\times n}$ denotes the $n\times n$ zero matrix. Hence, the infimum in (5.5) is a minimum and A is normal if and only if $\delta_{\|\cdot\|}(A) = 0$. Furthermore, it follows immediately from the definition of $\mathcal{R}(\cdot)$ that $\mathcal{R}(U^*AU) = \mathcal{R}(A)$ for all unitary matrices $U\in\mathbb{C}^{n\times n}$ and all $A\in\mathbb{C}^{n\times n}$. Consequently, $\delta_{\|\cdot\|}(U^*AU) = \delta_{\|\cdot\|}(A)$. In other words, $\delta_{\|\cdot\|}$ is invariant under unitary similarity transformations. It is also easily seen that $\delta_{\|\cdot\|}(zI_n + wA) = |w|\delta_{\|\cdot\|}(A)$ for all $z, w\in\mathbb{C}$. If all eigenvalues of $A\in\mathbb{C}^{n\times n}$ are simple then one has to compute at most $n!$ Schur forms of A in order to calculate $\delta_{\|\cdot\|}(A)$ as the following proposition shows.

Proposition 5.3.3 *Suppose that $A\in\mathbb{C}^{n\times n}$ has pairwise different eigenvalues $\lambda_1,\ldots,\lambda_n \in \mathbb{C}^n$. Let Π_n denote the set of all perturbations of the numbers $1,\ldots,n$ and to each permutation $\pi\in\Pi_n$ let $R_\pi\in\mathbb{C}^{n\times n}$ be a strictly upper tringular matrix such that* $\text{diag}(\lambda_{\pi(1)},\lambda_{\pi(2)},\ldots,\lambda_{\pi(n)}) + R_\pi$ *is a Schur form of A. Then for any unitarily invariant norm,*

$$\delta_{\|\cdot\|}(A) = \min_{\pi\in\Pi_n} \|R_\pi\|.$$

Proof: Let $\Lambda_\pi := \operatorname{diag}(\lambda_{\pi(1)}, \ldots, \lambda_{\pi(n)})$ and suppose that $\Lambda_\pi + R_\pi^{(k)}$, $k = 1, 2$ are Schur forms of A. Then the matrices $\Lambda_\pi + R_\pi^{(1)}$ and $\Lambda_\pi + R_\pi^{(2)}$ are unitarily similar. Moreover, by assumption the diagonal elements of Λ_π are pairwise different. Hence, by [83, Theorem 2.3], there exists a unitary diagonal matrix $U \in \mathbb{C}^{n\times n}$ such that $U^* R_\pi^{(1)} U = R_\pi^{(2)}$. Thus $\|R_\pi^{(1)}\| = \|R_\pi^{(2)}\|$. □

If A has a single eigenvalue then only one Schur form is needed in order to determine $\delta_{\|\cdot\|}(A)$. We have

Proposition 5.3.4 *If $\lambda \in \mathbb{C}$ is the only eigenvalue of $A \in \mathbb{C}^{n\times n}$ then $\delta_{\|\cdot\|}(A) = \|A - \lambda I_n\|$.*

Proof: Let $\lambda I_n + R$ be a Schur form of A. Then $A = U^*(\lambda I_n + R)U = \lambda I_n + U^* R U$ for some unitary matrix $U \in \mathbb{C}^{n\times n}$. Thus $\|A - \lambda I_n\| = \|U^* R U\| = \|R\|$. □

Corollary 5.3.5 *If $A \in \mathbb{C}^{n\times n}$ is nilpotent then $\delta_{\|\cdot\|}(A) = \|A\|$.*

For the departure from normality with respect to the Frobenius norm one has the following well known formula.

Proposition 5.3.6 *Let $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ be the eigenvalues of $A \in \mathbb{C}^{n\times n}$. Then*

$$\delta_{\|\cdot\|_F}(A) = \sqrt{\|A\|_F^2 - \textstyle\sum_{k=1}^n |\lambda_k|^2},$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Proof: Let $\Lambda + R$ be a Schur form of A. Then $\|A\|_F^2 = \|\Lambda + R\|_F^2 = \|\Lambda\|_F^2 + \|R\|_F^2 = \sum_{k=1}^n |\lambda_k|^2 + \|R\|_F^2$. □

From now on we denote the departure from normality of $A \in \mathbb{C}^{n\times n}$ with respect to spectral norm by $\delta_2(A)$. The latter yields a simple upper bound for pseudospectra with respect to unitarily invariant norms.

Proposition 5.3.7 *Let $A \in \mathbb{C}^{n\times n}$ and $\rho > 0$. Then for all normalized unitarily invariant norms,*

$$\bigcup_{\lambda\in\sigma(A)} \mathcal{D}(\lambda,\, \rho) \subseteq \sigma_{\mathbb{C}}^{\|\cdot\|}(A, \rho) \subseteq \bigcup_{\lambda\in\sigma(A)} \mathcal{D}(\lambda,\, \rho + \delta_2(A)\,).$$

Proof: The first inclusion holds by Proposition 5.2.2. We show the second. Let $\Lambda + R$ be a Schur form of $A \in \mathbb{C}^{n\times n}$ such that $\|R\|_2 = \delta_2(A)$. Then for all normalized unitarily invariant norms and all $\rho > 0$, $\sigma_{\mathbb{C}}^{\|\cdot\|}(A, \rho) = \sigma_{\mathbb{C}}^{(2)}(\Lambda + R, \rho) \subseteq \sigma_{\mathbb{C}}^{(2)}(\Lambda, \rho + \|R\|_2) = \bigcup_{\lambda\in\sigma(A)} \mathcal{D}(\lambda, \rho + \delta_2(A))$. □

Corollary 5.3.8 *Let $A \in \mathbb{C}^{n\times n}$ be a normal matrix. Then for every normalized unitarily invariant norm and every $\rho > 0$,*

$$\sigma_{\mathbb{C}}^{\|\cdot\|}(A, \rho) = \bigcup_{\lambda\in\sigma(A)} \mathcal{D}(\lambda,\, \rho\,).$$

5.4 Pseudospectra and Jordan structure

In this section we discuss the connection of the pseudospectrum of a matrix $A\in\mathbb{C}^{n\times n}$ with its Jordan structure. First, we consider generalized Jordan blocks of the form

$$J_n(\lambda,d) := \begin{bmatrix} \lambda & d & & & 0 \\ & \lambda & d & & \\ & & \ddots & \ddots & \\ & 0 & & \lambda & d \\ & & & & \lambda \end{bmatrix} \in \mathbb{C}^{n\times n}, \qquad n\geq 2.$$

We set $J_n(\lambda) := J_n(\lambda,1)$ and $J_1(\lambda) = J_1(\lambda,d) = [\lambda]\in\mathbb{C}^{1\times 1}$. Let $D_n(d) := \operatorname{diag}(d,d^2,\dots,d^n)$. Then for any $d\neq 0$, $D_n(d)J_n(\lambda)D_n(d)^{-1} = J_n(\lambda,d)$. For the permutation matrix $P_n = \begin{bmatrix} 0 & 1 \\ I_{n-1} & 0\end{bmatrix} \in \mathbb{C}^{n\times n}, n\geq 2$ we have $J_n(0,d)P_n = \begin{bmatrix} dI_{n-1} & 0 \\ 0 & 0\end{bmatrix}$. This implies $\|J_n(0,d)\|_p = \|J_n(0,d)P_n\|_p = \|dI_{n-1}\|_p = |d|$ for $n\geq 2$, $p\in[1,\infty]$. Proposition 5.3.4 then yields that $\delta_2(\,J_n(\lambda,d)\,) = \|J_n(0,d)\|_2 = |d|$, and from Corollary 5.2.6 we obtain

$$\sigma_{\mathbb{C}}^{(p)}(J_n(\lambda,d),\,\rho\,) \subseteq \mathcal{D}(\lambda,\,\rho+|d|\,), \qquad p\in[1,\infty]. \tag{5.6}$$

The theorem below shows that $\sigma_{\mathbb{C}}^{(p)}(J_n(\lambda,d),\,\rho\,)$ comes arbitrarily close to this upper bound as n tends to infinity.

Theorem 5.4.1 *Let $\|\cdot\|_{\alpha,\beta}$ be an operator norm induced by absolute norms. Then for every $\lambda,d\in\mathbb{C},\rho>0,n\in\mathbb{N}$, the pseudospectrum $\sigma_{\mathbb{C}}^{(\alpha,\beta)}(J_n(\lambda,d),\rho)$ is a disk about λ. Let $r_n^{(\alpha,\beta)}(\rho,d)$ denote the radius of that disk. Then*

$$r_n^{(\alpha,\beta)}(\rho,d) = r_n^{(\alpha,\beta)}(\rho,|d|) = \max\{r>0 \mid \ \|J_n(r,|d|)^{-1}\|_{\beta,\alpha}\geq \rho^{-1}\,\}.$$

For every $p_1,p_2\in[1,\infty]$ the sequence $n\mapsto r_n^{(p_1,p_2)}(\rho,d)$ is increasing. Moreover, we have for all $p\in[1,\infty]$ that $r_1^{(p,p)}(\rho,d)=\rho$ and

$$\lim_{n\to\infty} r_n^{(p,p)}(\rho,d) = \rho+|d|. \tag{5.7}$$

Proof: For $d=0$ or $n=1$ we have $J_n(\lambda,d)=\lambda I_n$. The corresponding pseudospectra are disks since

$$\sigma_{\mathbb{C}}^{(\alpha,\beta)}(\lambda I_n,\rho) \;=\; \{\lambda\}\cup\{\,s\in\mathbb{C}\backslash\{\lambda\}\mid\ \|sI_n-\lambda I_n\|_{\beta,\alpha}^{-1}\geq\rho^{-1}\,\} = \mathcal{D}(\lambda,\,\rho/\|I_n\|_{\beta,\alpha}\,).$$

Let $n\geq 2$ and $d,z\in\mathbb{C}\backslash\{0\}$. For notational convenience set $\|\cdot\| := \|\cdot\|_{\beta,\alpha}$ and $N_n := J_n(0,1)$. The unitary diagonal matrix $D(\frac{z\bar d}{|zd|}) = \operatorname{diag}(\frac{z\bar d}{|zd|},(\frac{z\bar d}{|zd|})^2,\dots,(\frac{z\bar d}{|zd|})^n)$ satisfies $D(\frac{z\bar d}{|zd|})^{-1}N_n\,D(\frac{z\bar d}{|zd|}) = \frac{z\bar d}{|zd|}\,N_n$. Thus

$$\begin{aligned} \|J_n(z,d)^{-1}\| &= \|(zI_n+dN_n)^{-1}\| = \|D(\tfrac{z\bar d}{|zd|})\,(zI_n+dN_n)^{-1}\,D(\tfrac{z\bar d}{|zd|})^{-1}\| \\ &= \left\|(z/|z|)^{-1}\,(|z|I_n+|d|N_n)^{-1}\right\| = \|\,(|z|I_n+|d|N_n)^{-1}\,\| \\ &= \|J_n(|z|,|d|)^{-1}\|, \end{aligned}$$

and hence

$$\begin{aligned}\sigma_{\mathbb{C}}^{(\alpha,\beta)}(J_n(\lambda,d),\rho) &= \{\lambda\}\cup\{\, s\in\mathbb{C}\backslash\{\lambda\} \mid \ \|(J_n(\lambda,d)-sI_n)^{-1}\| \geq \rho^{-1}\,\} \\ &= \{\lambda\}\cup\{\, s\in\mathbb{C}\backslash\{\lambda\} \mid \ \|J_n(\lambda-s,d)^{-1}\| \geq \rho^{-1}\,\} \\ &= \{\lambda\}\cup\{\, s\in\mathbb{C}\backslash\{\lambda\} \mid \ \|J_n(|\lambda-s|,|d|)^{-1}\| \geq \rho^{-1}\,\} \\ &= \{\lambda\}\cup\{\, \lambda+z \mid \ z\in\mathbb{C}\backslash\{0\},\ \|J_n(|z|,|d|)^{-1}\| \geq \rho^{-1}\,\}.\end{aligned}$$

Thus $\sigma_{\mathbb{C}}^{(\alpha,\beta)}(J_n(\lambda,d),\rho)$ has rotational symmetry with respect to the center λ. However, $J_n(\lambda,d)$ has only one eigenvalue, so $\sigma_{\mathbb{C}}^{(\alpha,\beta)}(J_n(\lambda,d),\rho)$ has only one connected component. Thus the set $\sigma_{\mathbb{C}}^{(\alpha,\beta)}(J_n(\lambda,d),\rho)$ is a disk of radius $r_n^{(\alpha,\beta)}(\rho,d)=\max\{r>0 \mid \ \|J_n(r,|d|)^{-1}\|\geq\rho^{-1}\,\}$. We have $J_{n+1}(\lambda,d) = \begin{bmatrix} J_n(\lambda,d) & d\,e_n \\ 0 & \lambda\end{bmatrix}$, where e_n denotes the nth standard basis vector of $\mathbb{C}^n$. Thus, by Proposition 5.2.5, $\sigma_{\mathbb{C}}^{(p_1,p_2)}(J_n(\lambda,d),\rho)\subseteq\sigma_{\mathbb{C}}^{(p_1,p_2)}(J_{n+1}(\lambda,d),\rho)$ for every $p_1,p_2\in[1,\infty]$,. Thus the sequence $n\mapsto r_n^{(p_1,p_2)}(\rho,d)$ is increasing. Equation (5.6) implies $r_n^{(p,p)}(\rho,d)\leq\rho+|d|$ for $p\in[1,\infty]$.

Now, we prove (5.7) by showing that to each $r\in(|d|,\rho+|d|)$ there exists an $n\in\mathbb{N}$ such that $\|J_n(r,|d|)^{-1}\|_p>\rho^{-1}$. First, we assume that $p\neq\infty$. Let $x_n\in\mathbb{C}^n$ denote the vector whose entries are all 1. Then $J_n(r,-|d|)\,x_n=(r-|d|)x_n+|d|e_n$. Therefore

$$\|J_n(r,|d|)^{-1}\|_p = \|J_n(r,-|d|)^{-1}\|_p \geq \frac{\|x_n\|_p}{\|(r-|d|)x_n+|d|e_n\|_p} = \left\|(r-|d|)\frac{x_n}{\|x_n\|_p}+\frac{|d|e_n}{\|x_n\|_p}\ \right\|_p^{-1}.$$

As n tends to infinity the right term converges to $(r-|d|)^{-1}>\rho^{-1}$(since $\lim_{n\to\infty}\|x_n\|_p=\infty$). Thus we have shown (5.7) for $p\neq\infty$. Now we consider the case $p=\infty$. We have

$$J_n(r,|d|)^{-1} \ = \ (rI_n+|d|N_n)^{-1}=r^{-1}\left(I_n+\frac{|d|}{r}N_n\right)^{-1}=r^{-1}\sum_{j=0}^{n-1}(-1)^j\left(\frac{|d|}{r}N_n\right)^j.$$

Thus, the element in the first row and jth column of $J_n(r,|d|)^{-1}$ is $r^{-1}(-|d|/r)^{j-1}$. Hence,

$$\begin{aligned}\|J_n(r,|d|)^{-1}\|_\infty &\geq r^{-1}\sum_{j=0}^{n-1}(|d|/r)^j \\ &= r^{-1}\left(1-\frac{|d|}{r}\right)^{-1}\left(1-\left(\frac{|d|}{r}\right)^n\right) \\ &= (r-|d|)^{-1}\left(1-\left(\frac{|d|}{r}\right)^n\right)\ \underset{n\to\infty}{\longrightarrow}\ (r-|d|)^{-1}>\rho^{-1}.\end{aligned}$$

□

Figure 5.1 shows some examples of functions of the form $\rho\mapsto r_n^{(2,2)}(\rho,d)$ and $0\leq d\mapsto r_n^{(2,2)}(\rho,d)$. Note that the latter functions are increasing. We will give a proof of his fact in Section 5.6. Figure 5.2 shows real and complex pseudospectra of a 6×6 Jordan matrix. Note that the real pseudospectra are no disks.

We are now going to extend the result of Bauer and Fike to matrices which are not necessarily diagonalizable. Let $\lambda_1,\ldots,\lambda_r$ be the pairwise different eigenvalues of $A\in\mathbb{C}^{n\times n}$, and let, as in Section 2.6

$$A \ = \ \sum_{j\in\underline{r}}(\,\lambda_j P_j+N_j\,) \tag{5.8}$$

be the Jordan decomposition of A, where P_j is the projector onto the generalized eigenspace associated with λ_j and $N_j = (A - \lambda_j I_n)P_j$. Let ν_j denote the index of nilpotency of N_j. Let $S \in \mathbb{C}^{n\times n}$ be a matrix whose columns form a Jordan basis of A such that $S^{-1}AS = \bigoplus_{j\in\underline{r}} \bigoplus_{k\in\underline{m_j}} J_{n_{jk}}(\lambda_j)$. Then

$$\nu_j = \max\{\, n_{jk} \mid 1 \le k \le m_j \}. \tag{5.9}$$

Moreover, since $J_{n_{jk}}(\lambda_j)$ is similar to $J_{n_{jk}}(\lambda_j, d)$ for $d \neq 0$, there is a nonsingular matrix $S_d \in \mathbb{C}^{n\times n}$ such that

$$S_d^{-1} A S_d = \bigoplus_{j\in\underline{r}} \bigoplus_{k\in\underline{m_j}} J_{n_{jk}}(\lambda_j, d), \qquad d \neq 0. \tag{5.10}$$

Theorem 5.4.2 *Suppose that (5.10) holds. Then we have for $p \in [0,\infty]$ and $\rho > 0$,*

$$\sigma_{\mathbb{C}}^{(p)}(A,\rho) \subseteq \bigcup_{j\in\underline{r}} \mathcal{D}\Big(\lambda_j,\ r_{\nu_j}^{(p,p)}(\,\rho\,\|S_d\|_p\,\|S_d^{-1}\|_p,\, d)\Big).$$

Proof: We have

$$\begin{aligned}
\sigma_{\mathbb{C}}^{(p)}(A,\rho) &= \sigma_{\mathbb{C}}^{(p)}\left(S_d \left(\bigoplus_{j\in\underline{r}} \bigoplus_{k\in\underline{m_j}} J_{n_{jk}}(\lambda_j) \right) S_d^{-1},\ \rho \right) \\
&\subseteq \sigma_{\mathbb{C}}^{(p)}\left(\bigoplus_{j\in\underline{r}} \bigoplus_{k\in\underline{m_j}} J_{n_{jk}}(\lambda_j),\ \rho\,\|S_d\|_p\,\|S_d^{-1}\|_p \right) && \text{(by Proposition 5.2.5)} \\
&= \bigcup_{j\in\underline{r}} \bigcup_{k\in\underline{m_j}} \sigma_{\mathbb{C}}^{(p)}\left(J_{n_{jk}}(\lambda_j),\ \rho\,\|S_d\|_p\,\|S_d^{-1}\|_p \right) && \text{(by Proposition 5.2.4)} \\
&\subseteq \bigcup_{j\in\underline{r}} \bigcup_{k\in\underline{m_j}} \mathcal{D}\left(\lambda_j,\ r_{n_{jk}}^{(p,p)}(\,\rho\,\|S_d\|_p\,\|S_d^{-1}\|_p,\, d)\right) && \text{(by Theorem 5.4.1)} \\
&= \bigcup_{j\in\underline{r}} \mathcal{D}\Big(\lambda_j,\ r_{\nu_j}^{(p,p)}(\,\rho\,\|S_d\|_p\,\|S_d^{-1}\|_p,\, d)\Big) && \text{(by Theorem 5.4.1 and (5.9))}
\end{aligned}$$

□

Corollary 5.4.3 *Let $A \in \mathbb{C}^{n\times n}$. Then to each $\epsilon > 0$ there is a matrix A_ϵ similar to A such that for all $\rho > 0$,*

$$\sigma_{\mathbb{C}}^{(p)}(A_\epsilon,\rho) \subseteq \bigcup_{\lambda\in\sigma(A)} \mathcal{D}(\lambda, \rho+\epsilon).$$

Proof: Set $A_\epsilon = \bigoplus_{j\in\underline{r}} \bigoplus_{k\in\underline{m_j}} J_{n_{jk}}(\lambda_j, \epsilon)$. □

Now, we specialize the results of Section 2.6 on spectral value sets for small perturbations to pseudospectra. We first introduce some notation: Let $\lambda_j \in \mathbb{C}$ be an eigenvalue of algebraic multiplicity m of $A \in \mathbb{C}^{n\times n}$. Recall from Section 2.6 that for $\widetilde{A} \in \mathbb{C}^{n\times n}$,

$$\mathrm{sv}_{(A,\lambda_j)}(\widetilde{A}) = \min\{\, \rho \ge 0 \mid \text{ the disk } \mathcal{D}(\lambda_j,\rho) \text{ contains at least } m \text{ eigenvalues of } \widetilde{A}\ \}.$$

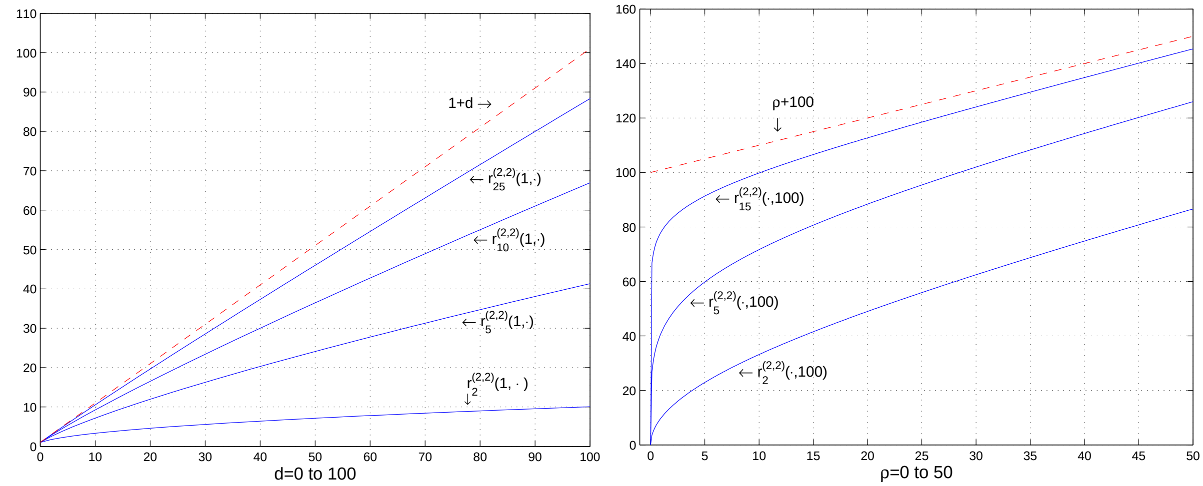


Figure 5.1: Graphs of the functions $0 \le d \mapsto r_n^{(2,2)}(1,d)$ and $\rho \mapsto r_n^{(2,2)}(\rho,100)$.

Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be norms on $\mathbb{C}^n$. The Hölder condition number of order $\omega > 0$ of the eigenvalue λ_j is defined by

$$\mathrm{cond}_{\mathbb{C},\omega}^{(\alpha,\beta)}(A,\lambda_j) \;\;:=\;\; \lim_{\rho\searrow 0} \max_{\substack{0\neq\Delta\in\mathbb{C}^{n\times n}\\ \|\Delta\|_{\alpha,\beta}\le\rho}} \frac{\mathrm{sv}_{(A,\lambda_j)}(A+\Delta)}{\|\Delta\|_{\alpha,\beta}^{\omega}}.$$

The spectral variation of λ_j with respect to the perturbation level ρ is

$$\mathrm{sv}_{\mathbb{C},\omega}^{(\alpha,\beta)}(A,\lambda_j,\rho) \;\;:=\;\; \max_{\substack{\Delta\in\mathbb{C}^{n\times n}\\ \|\Delta\|_{\alpha,\beta}\le\rho}} \mathrm{sv}_{(A,\lambda_j)}(A+\Delta).$$

Note that with the notation in Section 2.6, $\mathrm{cond}_{\mathbb{C},\omega}^{(\alpha,\beta)}(A,\lambda_j) = \mathrm{cond}_{\mathbb{C},\omega}^{\|\cdot\|_{\alpha,\beta}}(A,I_n,I_n,\lambda_j)$ and $\mathrm{sv}_{\mathbb{C},\omega}^{(\alpha,\beta)}(A,\lambda_j,\rho) = \mathrm{sv}_{\mathbb{C},\omega}^{\|\cdot\|_{\alpha,\beta}}(A,I_n,I_n,\lambda_j,\rho)$. Hence, the theorem below is an immediate consequence of Theorem 2.6.6 and Corollary 2.5.4.

Theorem 5.4.4 *Let $A\in\mathbb{C}^{n\times n}$ have the Jordan structure (5.8). Let $\mathcal{K}_j(\rho)$ denote the connected component of $\sigma_{\mathbb{C}}^{(\alpha,\beta)}(A,\rho)$ which contains the eigenvalue $\lambda_j\in\sigma(A)$. Then the following holds.*

Suppose λ_j is a semi-simple eigenvalue of A (i.e. $N_j = 0$). Then

(i) To each $\epsilon > 0$ there is a $\rho_\epsilon > 0$ such that for every $\rho\in(0,\rho_\epsilon]$,

$$\mathcal{D}\,(\lambda_j\,,\,(1-\epsilon)\,\rho\,\|P_j\|_{\beta,\alpha}\,) \subseteq \mathcal{K}_j(\rho) \subseteq \mathcal{D}\,(\lambda_j\,,\,(1+\epsilon)\;\rho\;\|P_j\|_{\beta,\alpha}\,)\,.$$

(ii) We have $\lim_{\rho\searrow 0}\rho^{-1}(\mathcal{K}_j(\rho)-\lambda_j) = \mathcal{D}\,(0\,,\,\|P_j\|_{\beta,\alpha}\,)$, where the limit is taken with respect to the Hausdorff metric.

(iii) The spectral variation satisfies $\mathrm{sv}_{\mathbb{C}}^{(\alpha,\beta)}(A,\lambda_j,\rho) = \|P_j\|_{\beta,\alpha}\,\rho + o(\rho)$.

(iv) We have $\mathrm{cond}_{\mathbb{C},1}^{(\alpha,\beta)}(A,\lambda_j) = \|P_j\|_{\beta,\alpha}$.

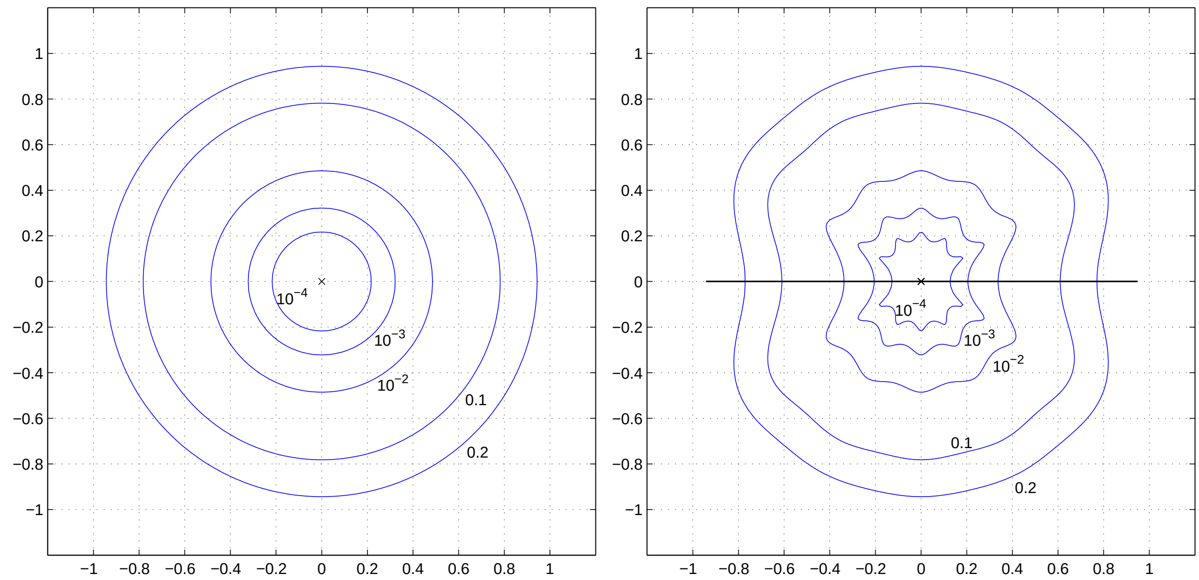


Figure 5.2: The pseudospectra $\sigma_{\mathbb{F}}^{(2)}(J_6(0,1),\rho)$ for $\rho = 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.2$. Left: $\mathbb{F} = \mathbb{C}$. Right: $\mathbb{F} = \mathbb{R}$.

Let λ_j be a defective eigenvalue of A (i.e. $N_j \neq 0$). Then

(v) To each $\epsilon > 0$ there is a $\rho_\epsilon > 0$ such that for every $\rho \in (0, \rho_\epsilon]$,

$$\mathcal{D}\left(\lambda_j\,,\ \left((1-\epsilon)\ \rho\ \|N_j^{\nu_j-1}\|_{\beta,\alpha}\right)^{\frac{1}{\nu_j}}\right) \subseteq \mathcal{K}_j(\rho) \subseteq \mathcal{D}\left(\lambda_j\,,\ \left((1+\epsilon)\ \rho\ \|N_j^{\nu_j-1}\|_{\beta,\alpha}\right)^{\frac{1}{\nu_j}}\right).$$

(vi) We have $\lim_{\rho\searrow 0} \rho^{-\frac{1}{\nu_j}}(\mathcal{K}_j(\rho) - \lambda_j) = \mathcal{D}\left(0\,,\ \|N_j^{\nu_j-1}\|_{\beta,\alpha}^{\frac{1}{\nu_j}}\right)$, where the limit is taken with respect to the Hausdorff metric.

(vii) The spectral variation satisfies $\mathrm{sv}_{\mathbb{C}}^{(\alpha,\beta)}(A,\lambda_j,\rho) = \|N_j^{\nu_j-1}\|_{\beta,\alpha}^{\frac{1}{\nu_j}}\,\rho^{\frac{1}{\nu_j}} + o\left(\rho^{\frac{1}{\nu_j}}\right)$.

(viii) We have $\mathrm{cond}_{\mathbb{C},\frac{1}{\nu_j}}^{(\alpha,\beta)}(A,\lambda_j) = \|N_j^{\nu_j-1}\|_{\beta,\alpha}^{\frac{1}{\nu_j}}$.

For the special case that $\alpha = \beta = 2$ the statements $(iii), (iv), (vii), (viii)$ in the theorem above have been obtained in [19, 37, 73]. Note that if λ_j is a simple eigenvalue such that $Ax = \lambda_j x$, $A^T y = \lambda_j y$, $x, y \in \mathbb{C}^n \setminus \{0\}$, then $P_j = |y^T x|^{-1} x y^T$, and therefore

$$\mathrm{cond}_{\mathbb{C},1}^{(\alpha,\beta)}(A,\lambda_j) = \|P_j\|_{\beta,\alpha} = \frac{\|x\|_\alpha \|y\|_\beta}{|y^T x|}.$$

This basic result can be found in many textbooks on numerical linear algebra, see e.g.[20, 34, 86]. By combining Theorem 5.4.1 and Theorem 5.4.4 we obtain

Corollary 5.4.5 *For every $n \in \mathbb{N}$, $d \in \mathbb{C}$, $p \in [1,\infty]$ we have $r_n^{(p,p)}(\rho, d) = |d|^{\frac{n-1}{n}}\rho^{\frac{1}{n}} + o\left(\rho^{\frac{1}{n}}\right)$.*

5.5 The 2×2-case

In this section we consider the simplest nontrivial case of spectral value sets, namely the pseudospectra of 2×2-matrices with respect to unitarily invariant norms. We start with determining the Schur forms and the departure from normality. Throughout this section $\Sigma_{\lambda_1,\lambda_2}$ denotes the set of $A\in\mathbb{C}^{2\times 2}$ whose eigenvalues are $\lambda_1, \lambda_2 \in\mathbb{C}$. For $A\in\Sigma_{\lambda_1,\lambda_2}$ we set

$$\delta(A) := \sqrt{\operatorname{tr}(A^*A) - \frac{1}{2}\left(|\operatorname{tr}(A)|^2 + |\operatorname{tr}(A)^2 - 4\det(A)|\right)}. \tag{5.11}$$

A short computation yields that $\delta(A) = \sqrt{\|A\|_F^2 - |\lambda_1|^2 - |\lambda_2|^2}$. Thus $\delta(A)$ is the departure from normality of A with respect to the Frobenius norm. However, more is true.

Proposition 5.5.1 *Let $A\in\Sigma_{\lambda_1,\lambda_2}$. Then all upper triangular Schur forms of A are given by*

$$\begin{bmatrix}\lambda_1 & \delta(A)\,z\\ 0 & \lambda_2\end{bmatrix}, \quad \begin{bmatrix}\lambda_2 & \delta(A)\,z\\ 0 & \lambda_1\end{bmatrix} \qquad \text{where } z\in\mathbb{C},\ |z|=1.$$

Thus for all normalized unitarily invariant norms, $\delta(A) = \delta_{\|\cdot\|}(A)$.

Proof: Let $\begin{bmatrix}\lambda_1 & d\\ 0 & \lambda_2\end{bmatrix} = U^*AU$ be a Schur form of A. Then the matrices

$$\begin{bmatrix}\lambda_1 & d\,z\\ 0 & \lambda_2\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & z\end{bmatrix}^* \begin{bmatrix}\lambda_1 & d\\ 0 & \lambda_2\end{bmatrix} \begin{bmatrix}1 & 0\\ 0 & z\end{bmatrix} \qquad z\in\mathbb{C},\ |z|=1$$

are also Schur forms of A. For all normalized unitarily invariant norms we have

$$\left\|\begin{bmatrix}0 & dz\\ 0 & 0\end{bmatrix}\right\|^2 = |d|^2 = \|U^*AU\|_F^2 - |\lambda_1|^2 - |\lambda_2|^2 = \|A\|_F^2 - |\lambda_1|^2 - |\lambda_2|^2 = \delta(A)^2. \qquad \square$$

The quantity $\delta(A)$ is closely related to the distance of A from the set of normal 2×2 matrices with respect to unitarily invariant norms. More precisely, the following statement holds.

Proposition 5.5.2 *Let $A\in\mathbb{C}^{2\times 2}$. Then for all normalized unitarily invariant norms,*

$$\min\{\ \|A-N\|\ |\ N\in\mathbb{C}^{2\times 2},\ N^*N = NN^*\ \} = \frac{\delta(A)}{2}\|I_2\|. \tag{5.12}$$

Proof: We may assume that A is in Schur form, i.e. $A = \begin{bmatrix}\lambda_1 & d\\ 0 & \lambda_2\end{bmatrix}$, $d = \delta(A) \geq 0$. Then, according to [10, Theorem IX.7.6] the minimum on the left hand side of (5.12) is attained for the normal matrix $N_0 = \begin{bmatrix}\lambda_1 & \frac{1}{2}d\\ \frac{1}{2}e^{2i\theta}d & \lambda_2\end{bmatrix}$, where $\theta = \arg(\lambda_1 - \lambda_2)$. The minimum is

$$\|A - N_0\| = \left\|\begin{bmatrix}0 & \frac{1}{2}d\\ -\frac{1}{2}e^{2i\theta}d & 0\end{bmatrix}\right\| = \tfrac{d}{2}\left\|\begin{bmatrix}0 & 1\\ -e^{2i\theta} & 0\end{bmatrix}\right\| = \tfrac{d}{2}\,\|I_2\|. \qquad \square$$

We now give a description of the pseudospectra of a 2×2 matrix in terms of its eigenvalues and its departure from normality.

Proposition 5.5.3 *Let $A \in \Sigma_{\lambda_1,\lambda_2}$, $\rho > 0$. Then for every normalized unitarily invariant norm,*

$$\sigma_{\mathbb{C}}^{\|\cdot\|}(A,\rho) \quad = \quad \mathcal{D}(\lambda_1,\rho) \cup \mathcal{D}(\lambda_2,\rho) \cup \mathcal{M}_{\lambda_1,\lambda_2}(\delta(A),\rho),$$

where

$$\mathcal{M}_{\lambda_1,\lambda_2}(d,\rho) := \{\, s \in \mathbb{C} \mid \quad (|s-\lambda_1|^2-\rho^2)(|s-\lambda_2|^2-\rho^2) \le \rho^2 d^2 \;\}, \quad d \ge 0.$$

Proof: By Theorem 5.2.2 we have that $\mathcal{D}(\lambda_1,\rho) \cup \mathcal{D}(\lambda_2,\rho) \subseteq \sigma_{\mathbb{C}}^{\|\cdot\|}(A,\rho)$. For $s \in \mathbb{C}$ let

$$\begin{aligned}
\theta_s &:= \operatorname{tr}(\,(sI_2-A)^*(sI_2-A)\,) = |s-\lambda_1|^2 + |s-\lambda_2|^2 + \delta(A)^2, \\
\xi_s &:= \det(\,(sI_2-A)^*(sI_2-A)\,) = |s-\lambda_1|^2\,|s-\lambda_2|^2.
\end{aligned}$$

Then a direct computation yields

$$\sigma_{\min}(sI_2-A) = \sqrt{\theta_s/2 - \sqrt{\theta_s/4 - \xi_s}}.$$

If $s \notin \mathcal{D}(\lambda_1,\rho) \cup \mathcal{D}(\lambda_2,\rho)$ then $\frac{\theta_s}{2} - \rho^2 \ge 0$ and the following equivalences hold.

$$\begin{aligned}
s \in \sigma_{\mathbb{C}}^{\|\cdot\|}(A,\rho) \quad &\Leftrightarrow \quad \rho^2 \ge (\sigma_{\min}(sI_2-A))^2 = \frac{\theta_s}{2} - \sqrt{\frac{\theta_s^2}{4} - \xi_s} \\
&\Leftrightarrow \quad \frac{\theta_s}{2} - \rho^2 \le \sqrt{\frac{\theta_s^2}{4} - \xi_s} \\
&\Leftrightarrow \quad \left(\frac{\theta_s}{2} - \rho^2\right)^2 \le \frac{\theta_s^2}{4} - \xi_s \\
&\Leftrightarrow \quad (|s-\lambda_1|^2-\rho^2)(|s-\lambda_2|^2-\rho^2) \le \rho^2\delta(A)^2.
\end{aligned}$$

□

The sets $\mathcal{M}_{\lambda_1,\lambda_2}(d,\rho), \rho > 0$, have the following properties.

- If $0 \le d_1 < d_2$ then $\mathcal{M}_{\lambda_1,\lambda_2}(d_1,\rho) \subset \mathcal{M}_{\lambda_1,\lambda_2}(d_2,\rho)$ and $\partial\mathcal{M}_{\lambda_1,\lambda_2}(d_1,\rho) \subset \operatorname{int}\mathcal{M}_{\lambda_1,\lambda_2}(d_2,\rho)$.
- To each bounded subset $B \subset \mathbb{C}$ there exists a $d > 0$ such that $B \subset \mathcal{M}_{\lambda_1,\lambda_2}(d,\rho)$.
- $\mathcal{M}_{\lambda_1,\lambda_2}(0,\rho) = \partial\mathcal{D}(\lambda_1,\rho) \cup \partial\mathcal{D}(\lambda_2,\rho)$.

Hence we have the following corollary to Proposition 5.5.3.

Corollary 5.5.4 *Let $\|\cdot\|$ be a normalized unitarily invariant norm and let $\rho > 0$. Then*

(a) *if $A_1, A_2 \in \Sigma_{\lambda_1,\lambda_2}$ and $\delta(A_1) < \delta(A_2)$ then $\sigma_{\mathbb{C}}^{\|\cdot\|}(A_1,\rho) \subset \sigma_{\mathbb{C}}^{\|\cdot\|}(A_2,\rho)$;*

(b) *to each bounded subset $B \subset \mathbb{C}$ there exists an $A \in \Sigma_{\lambda_1,\lambda_2}$ such that $B \subset \sigma_{\mathbb{C}}^{\|\cdot\|}(A,\rho)$, whence $\bigcup_{A \in \Sigma_{\lambda_1,\lambda_2}} \sigma_{\mathbb{C}}^{\|\cdot\|}(A,\rho) = \mathbb{C}$.*

(c) *if $A \in \Sigma_{\lambda_1,\lambda_2}$ then $\sigma_{\mathbb{C}}^{\|\cdot\|}(A,\rho) = \mathcal{D}(\lambda_1,\rho) \cup \mathcal{D}(\lambda_2,\rho)$ if and only if A is normal.*

In Section 5.6 we will generalize the results of Corollary 5.5.4 to block triangular matrices . The next proposition gives an upper bound for $\sigma_{\mathbb{C}}^{\|\cdot\|}(A,\rho)$ for the case that ρ is smaller than half of the distance of the eigenvalues of A.

Proposition 5.5.5 *Let $A \in \Sigma_{\lambda_1,\lambda_2}$ and $0 < \rho < \frac{|\lambda_1-\lambda_2|}{2}$. Then*

$$\sigma_{\mathbb{C}}^{\|\cdot\|}(A,\rho) \subseteq \mathcal{D}(\lambda_1, r_{A,\rho}) \cup \mathcal{D}(\lambda_2, r_{A,\rho}), \quad \textit{where} \quad r_{A,\rho} := \rho\sqrt{1+\delta(A)^2\left(\frac{|\lambda_1-\lambda_2|^2}{4}-\rho^2\right)^{-1}}.$$

Proof: For each $s \in \mathbb{C}$ we have that $|s-\lambda_1| \geq \frac{|\lambda_1-\lambda_2|}{2}$ or $|s-\lambda_2| \geq \frac{|\lambda_1-\lambda_2|}{2}$. Let $\rho < \frac{|\lambda_1-\lambda_2|}{2}$. Then

$$|s-\lambda_j|^2 - \rho^2 \geq \frac{|s-\lambda_j|^2}{4} - \rho^2 \geq \frac{|\lambda_1-\lambda_2|^2}{4} - \rho^2 \geq 0 \quad \text{for } j=1 \text{ or } j=2. \tag{5.13}$$

Suppose now that $|s-\lambda_k| > r_{A,\rho}$, $k=1,2$. By defintion of $r_{A,\rho}$ this implies.

$$\left(|s-\lambda_k|^2-\rho^2\right)\left(\frac{|\lambda_1-\lambda_2|^2}{4}-\rho^2\right) \geq \rho^2\delta(A)^2, \quad k=1,2. \tag{5.14}$$

Since $r_{A,\rho} \geq \rho$ we have that $s \notin \mathcal{D}(\lambda_1,\rho) \cup \mathcal{D}(\lambda_1,\rho)$, and by combining the inequalities (5.13) and (5.14) we obtain

$$(|s-\lambda_1|^2-\rho^2)(|s-\lambda_2|^2-\rho^2) > \rho^2\delta(A)^2.$$

Thus $s \notin \sigma_{\mathbb{C}}^{\|\cdot\|}(A,\rho)$ by Proposition 5.5.3. □

Proposition 5.5.5 has an interesting consequence: Consider the following measures of nonnormality [24, 21].

$$\mathfrak{n}_{\|\cdot\|}(A) := \|A^*A - AA^*\|,$$

where $\|\cdot\|$ is a normalized unitarily invariant norm. For $A \in \Sigma_{\lambda_1,\lambda_2}$ we have

$$\mathfrak{n}_{\|\cdot\|}(A) \geq \mathfrak{n}_{\|\cdot\|_2}(A) = \delta(A)\sqrt{\delta(A)^2+|\lambda_1-\lambda_2|^2}.$$

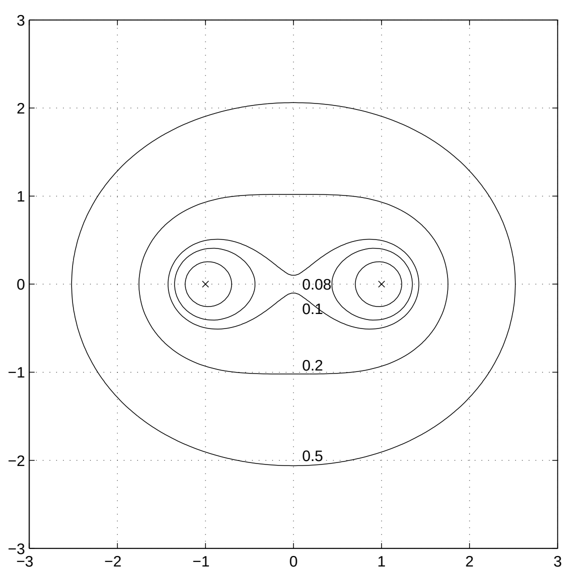


Figure 5.3: the sets $\sigma_{\mathbb{C}}^{(2)}\left(\begin{bmatrix}1 & 10\\ 0 & -1\end{bmatrix}, \rho\right)$ for $\rho = 0.05, 0.08, 0.1, 0.2, 0.5$.

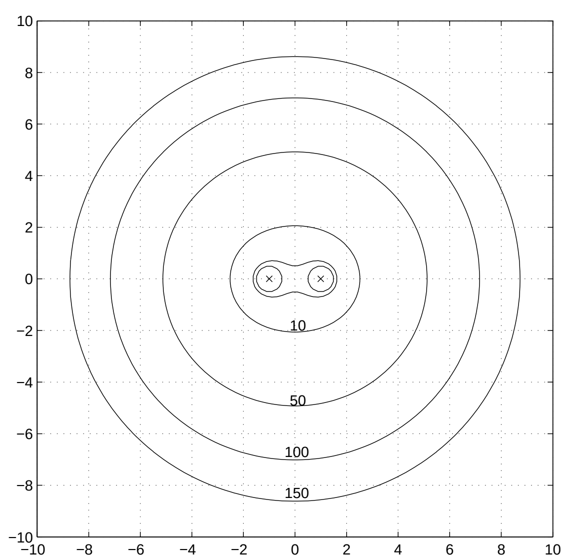


Figure 5.4: the sets $\sigma_{\mathbb{C}}^{(2)}\left(\begin{bmatrix}1 & \delta\\ 0 & -1\end{bmatrix}, 0.5\right)$ for $\delta = 0, 2, 10, 50, 100, 150$.

Let $(A_t)_{t\geq 0}$ be a family of matrices in $\mathbb{C}^{2\times 2}$ such that the distance of the eigenvalues of A_t tends to infinity as $t \to \infty$, and $\delta(A_t) = \text{const} = \delta(A_0) > 0$. Then also $\mathfrak{n}_{\|\cdot\|}(A_t)$ tends to infinity but the radius $r_{A_t,\rho}$ in Proposition 5.5.5 converges to ρ, the smallest possible value. This shows that the quantities $\delta(A)$ and $\mathfrak{n}_{\|\cdot\|}(A)$ fail to fully reflect the sensitivity of the spectrum of A, for if the eigenvalues of A are far away from each other then the pseudospectra of A may be close to the pseudospectra of a normal matrix even if $\delta(A)$ and $\mathfrak{n}_{\|\cdot\|}(A)$ are large. We will investigate this phenomenon further in Section 5.8.

Pseudospectra of 2×2 matrices yield simple lower bounds for pseudospectra of matrices of larger size. We first treat the case that $A \in \mathbb{C}^{n\times n}$ has a defective eigenvalue.

Proposition 5.5.6 *Suppose that $x \in \mathbb{C}^n$ is a generalized eigenvector of $A \in \mathbb{C}^{n\times n}$ such that $(A - \lambda I_n)^{\nu-1}x \neq 0$ and $(A - \lambda I_n)^{\nu}x = 0$, $\lambda \in \mathbb{C}$, $\nu \geq 2$. Then we have for any normalized unitarily invariant norm and any $\rho \geq 0$,*

$$\mathcal{D}\left(\lambda,\, \rho\left(1 + \frac{\delta_x}{\rho}\right)^{\frac{1}{2}}\right) = \sigma_{\mathbb{C}}^{\|\cdot\|}\left(\begin{bmatrix}\lambda & \delta_x\\ 0 & \lambda\end{bmatrix}, \rho\right) \subseteq \sigma_{\mathbb{C}}^{\|\cdot\|}(A, \rho), \tag{5.15}$$

where

$$\delta_x := \left(\frac{\|(A-\lambda I_n)^{\nu-2}x\|_2^2}{\|(A-\lambda I_n)^{\nu-1}x\|_2^2} - \frac{|\,((A-\lambda I_n)^{\nu-2}x)^*(A-\lambda I_n)^{\nu-1}x\,|^2}{\|(A-\lambda I_n)^{\nu-1}x\|_2^4}\right)^{-\frac{1}{2}}.$$

Proof: Let $N := A - \lambda I_n$. By applying Gram-Schmidt orthonormalization to the linearly independent vectors $N^{\nu-1}x, N^{\nu-2}x$ one obtains the vectors

$$v_1 := \|N^{\nu-1}x\|_2^{-1}N^{\nu-1}x, \qquad v_2 := \left(\|N^{\nu-2}x\|_2^2 - |v_1^*N^{\nu-2}x|^2\right)^{-\frac{1}{2}}(N^{\nu-2}x - (v_1^*N^{\nu-2}x)v_1).$$

These vectors form an orthonormal basis of the A-invariant subspace $\operatorname{span}\{N^{\nu-1}x, N^{\nu-2}x\}$. It is straightforward to verify that $Av_1 = \lambda v_1$ and $Av_2 = \delta_x v_1 + \lambda v_2$. Hence, the inclusion in (5.15) follows from Corollary 5.2.8. The equality in (5.15) is an immediate consequence of Proposition 5.5.3. □

Now, we consider the case of different eigenvalues.

Proposition 5.5.7 *Let $x_1, x_2 \in \mathbb{C}^n \setminus \{0\}$ be eigenvectors of $A \in \mathbb{C}^{n\times n}$ to the eigenvalues $\lambda_1, \lambda_2 \in \mathbb{C}$. If $\lambda_1 \neq \lambda_2$ then for any normalized unitarily invariant norm and any $\rho \geq 0$,*

$$\sigma_{\mathbb{C}}^{\|\cdot\|}\left(\begin{bmatrix}\lambda_1 & \delta_{x_1,x_2}\\ 0 & \lambda_2\end{bmatrix}, \rho\right) \subseteq \sigma_{\mathbb{C}}^{\|\cdot\|}(A, \rho),$$

where

$$\delta_{x_1,x_2} := |\lambda_1 - \lambda_2| \cot\varphi, \qquad \varphi := \arccos\frac{|x_1^*x_2|}{\|x_1\|_2\|x_2\|_2}.$$

Proof: The proof is analogous to that of Proposition 5.5.6. Let $v_1 := \|x_1\|_2^{-1}x_1$, $\xi := x_2 - (v_1^*x_2)v_1$ and $v_2 := \|\xi\|_2^{-1}\xi$. Then the vectors v_1, v_2 form an orthonormal basis of the A-invariant subspace $\operatorname{span}\{x_1, x_2\}$. A straightforward computation yields that

$$Av_1 = \lambda_1 v_1, \qquad Av_2 = dv_1 + \lambda_2 v_2, \quad \text{where} \quad d := (\lambda_2 - \lambda_1)\frac{x_1^*x_2}{\sqrt{\|x_1\|_2^2\|x_2\|_2^2 - |x_1^*x_2|^2}}.$$

By multiplying v_2 with a complex number of modulus 1 we can achieve that $d = |d|$. Then

$$d = |\lambda_2 - \lambda_1| \frac{|x_1^* x_2|}{\sqrt{\|x_1\|_2^2 \|x_2\|_2^2 - |x_1^* x_2|^2}} = |\lambda_2 - \lambda_1| \frac{\cos\varphi}{\sqrt{1 - (\cos\varphi)^2}} = |\lambda_2 - \lambda_1| \cot\varphi = \delta_{x_1,x_2}.$$

Now, Corollary 5.2.8 yields the result. □

For a 2×2 matrix A the inclusions in Propositions 5.5.6 and 5.5.7 are equalities. The next proposition gives the relationship between the distance of eigenvalues, the departure from normality, the angle between the eigenspaces and the condition number of the eigenvalues of a 2×2 matrix.

Proposition 5.5.8 *For $A \in \Sigma_{\lambda_1,\lambda_2}$ with $\lambda_1 \neq \lambda_2$ let $\vartheta_A := |\lambda_1 - \lambda_2|$ and $\varphi_A := \arccos \frac{|x_1^* x_2|}{\|x_1\|_2 \|x_2\|_2}$, where $x_k \in \mathbb{C}^2 \setminus \{0\}$ is an eigenvector of A to the eigenvalue λ_k, $k = 1, 2$. Then $\delta(A) = \vartheta_A \cot\varphi_A$ and $\operatorname{cond}_{\mathbb{C},1}^{(2,2)}(A, \lambda_k) = (\sin\varphi_A)^{-1} = \sqrt{1 + (\delta(A)/\vartheta_A)^2}$, $k = 1, 2$.*

Proof: The relation $\delta(A) = \vartheta_A \cot\varphi_A$ follows from Proposition 5.5.7. Hence, $(\sin\varphi_A)^{-1} = \sqrt{1 + (\delta(A)/\vartheta_A)^2}$. In order to show that $\operatorname{cond}_{\mathbb{C},1}^{(2,2)}(A, \lambda_k) = (\sin\varphi_A)^{-1}$ we may assume that A is in Schur form, $A = \begin{bmatrix} \lambda_1 & \delta(A) \\ 0 & \lambda_2 \end{bmatrix}$. Let $x = [1 \quad 0]^T$, $y = [\lambda_1 - \lambda_2 \quad \delta(A)]^T$. Then $Ax = \lambda_1 x$, $A^T y = \lambda_1 y$. Thus $\operatorname{cond}_{\mathbb{C},1}^{(2,2)}(A, \lambda_1) = |y^T x|^{-1}(\|x\|_2 \|y\|_2) = \vartheta_A^{-1} \sqrt{\vartheta_A^2 + \delta(A)^2} = \sqrt{1 + (\delta(A)/\vartheta_A)^2}$. Analogously one obtains $\operatorname{cond}_{\mathbb{C},1}^{(2,2)}(A, \lambda_2) = \sqrt{1 + (\delta(A)/\vartheta_A)^2}$. □

We now investigate the dependence of the pseudospectrum on the parameters ϑ_A and φ_A. To this end we consider the largest radius $r \geq \rho$ such that $\mathcal{D}(\lambda_1, r) \cup \mathcal{D}(\lambda_2, r) \subseteq \sigma_{\mathbb{C}}^{\|\cdot\|}(A, \rho)$.

Proposition 5.5.9 *For all $\rho > 0$, $\vartheta \geq 0$ and all $\varphi \in (0, \frac{\pi}{2}]$ the polynomial*

$$p_{\varphi,\vartheta,\rho}(x) := (\,(x + \vartheta)^2 - \rho^2)(x^2 - \rho^2) - \rho^2 \vartheta^2 (\cot\varphi)^2$$

has exactly one zero $\geq \rho$. Let $\hat{r}(\varphi, \vartheta, \rho)$ denote that zero. Then we have for all $A \in \Sigma_{\lambda_1,\lambda_2}$ with $\lambda_1 \neq \lambda_2$ and all $\rho > 0$,

$$\hat{r}(\varphi_A, \vartheta_A, \rho) = \max\{\, r \geq 0 \mid \mathcal{D}(\lambda_1, r) \cup \mathcal{D}(\lambda_2, r) \subseteq \sigma_{\mathbb{C}}^{\|\cdot\|}(A, \rho)\,\}. \tag{5.16}$$

The function $(\varphi, \vartheta, \rho) \mapsto \hat{r}(\varphi, \vartheta, \rho)$ is continuous. Moreover, the following holds.

(a) *For all $\rho > 0$, $\varphi \in (0, \frac{\pi}{2})$ the function $\vartheta \longmapsto \hat{r}(\varphi, \vartheta, \rho)$, $\vartheta \geq 0$, is strictly increasing with*

$$\hat{r}(\varphi, 0, \rho) = \rho, \qquad \lim_{\vartheta \to \infty} \hat{r}(\varphi, \vartheta, \rho) = \rho(\sin\varphi)^{-1}.$$

(b) *For all $\rho, \vartheta > 0$ the function $\varphi \longmapsto \hat{r}(\varphi, \vartheta, \rho), \varphi \in (0, \frac{\pi}{2}]$, is decreasing with*

$$\lim_{\varphi \to 0} \hat{r}(\varphi, \vartheta, \rho) = \infty, \qquad \hat{r}(\tfrac{\pi}{2}, \vartheta, \rho) = \rho.$$

The above proposition shows the following. For a 2×2 matrix A with different eigenvalues the radius $\hat{r}(\varphi_A, \vartheta_A, \rho)$ always lies between ρ and $\rho \times$ (condition number of the eigenvalues). If we change A such that the angle φ_A between the eigenvectors remains fixed and the distance ϑ_A

between the eigenvalues tends to 0 (resp. ∞) then the radius $\hat{r}(\varphi_A, \vartheta_A, \rho)$ tends to ρ (resp. $\rho \times$ (condition number of the eigenvalues of A)). If we change A such that the distance ϑ_A between the eigenvalues remains fixed and the angle φ_A tends to 0 (resp. $\frac{\pi}{2}$) then the radius $\hat{r}(\varphi_A, \vartheta_A, \rho)$ tends to ∞ (resp. ρ).

Proof of Proposition 5.5.9: On the interval $[\rho, \infty)$ the continuous function $x \mapsto p_{\varphi,\vartheta,\rho}(x)$ is strictly increasing with $p_{\varphi,\vartheta,\rho}(\rho) \le 0$ and $\lim_{x\to\infty} p_{\varphi,\vartheta,\rho}(x) = \infty$. Thus $p_{\varphi,\vartheta,\rho}(\cdot)$ has exactly one zero $\hat{r}(\varphi, \vartheta, \rho)$ in $[\rho, \infty)$ and the implicit function theorem yields that this zero depends continuously on the parameters.

(a). For fixed ρ and φ let $x_\vartheta = \hat{r}(\varphi, \vartheta, \rho)$ and consider the polynomials $q_\vartheta(x) := \vartheta^{-2} p_{\varphi,\vartheta,\rho}(x) = \left(\frac{x^2-\rho^2}{\vartheta^2} + 1 + \frac{2x}{\vartheta}\right)(x^2 - \rho^2) - \rho^2(\cot\varphi)^2$. Let $\vartheta_1 > \vartheta > 0$. Then $x_\vartheta > \rho$, which implies that $q_{\vartheta_1}(x_\vartheta) < q_\vartheta(x_\vartheta) = 0$. Thus $p_{\varphi,\vartheta_1,\rho}(x_\vartheta) < 0$. It follows that $x_{\vartheta_1} > x_\vartheta$. Thus the function $\vartheta \mapsto x_\vartheta$ is strictly increasing and the limit $\hat{x} = \lim_{\vartheta\to\infty} x_\vartheta$ exists. The relation $q_\vartheta(x_\vartheta) = 0$ for all $\vartheta > 0$ implies first that $\hat{x} \ne \infty$, and secondly that $0 = (\hat{x}^2 - \rho^2) - \rho^2(\cot\varphi)^2 = \hat{x}^2 - \rho^2(\sin\varphi)^{-2}$. The statements in ($b$) are obvious.

Next, we prove the relation (5.16). Let $s = \lambda_2 + |z|$ with $\rho \le |z| \le r(\varphi_A, \vartheta_A, \rho)$. Then

$$\begin{aligned}
(|s-\lambda_1|^2 - \rho^2)(|s-\lambda_2|^2 - \rho^2) &= (|z + \lambda_2 - \lambda_1|^2 - \rho^2)(|z|^2 - \rho^2) \\
&\le ((|z| + \vartheta_A)^2 - \rho^2)(|z|^2 - \rho^2) && (5.17) \\
&\le \rho^2 \vartheta_A^2 (\cot\varphi_A)^2 && (5.18) \\
&= \rho^2 \delta(A).
\end{aligned}$$

Thus $s \in \mathcal{M}_{\lambda_1,\lambda_2}(\delta(A), \rho) \subseteq \sigma_{\mathbb{C}}^{\|\cdot\|}(A, \rho)$. By symmetry the same statement holds if we replace $s = \lambda_2 + z$ by $s = \lambda_1 + z$. Therefore $\hat{r}(\varphi_A, d_A, \rho) \le \max\{\, r \ge 0 \mid \mathcal{D}(\lambda_1, r) \cup \mathcal{D}(\lambda_2, r) \subseteq \sigma_{\mathbb{C}}^{\|\cdot\|}(A, \rho) \,\}$. For $z = (\hat{r}(\varphi_A, \vartheta_A, \rho) + \epsilon)(\lambda_2 - \lambda_1)$, $\epsilon > 0$, equality holds in (5.17) and the inequality (5.18) fails. Thus $z \notin \mathcal{M}_{\lambda_1,\lambda_2}(\delta(A), \rho)$. Thus $z \notin \sigma_{\mathbb{C}}^{\|\cdot\|}(A, \rho)$ and (5.16) holds. □

5.6 Diagonal scaling of block triangular matrices

According to Corollary 5.5.4 the pseudospectra of 2×2-matrices have the following monotonicity property for all $z, w, \lambda_k \in \mathbb{C}$, $\rho > 0$:

$$\text{If} \quad |z| \le |w| \quad \text{then} \quad \sigma_{\mathbb{C}}^{(2)}\left(\begin{bmatrix} \lambda_1 & z \\ 0 & \lambda_2 \end{bmatrix}, \rho\right) \subseteq \sigma_{\mathbb{C}}^{(2)}\left(\begin{bmatrix} \lambda_1 & w \\ 0 & \lambda_2 \end{bmatrix}, \rho\right).$$

The inclusion is an equality if $|z| = |w|$. Moreover, we have that to each bounded subset B of $\mathbb{C}$ there exists a $z \in \mathbb{C}$ such that $B \subseteq \sigma_{\mathbb{C}}^{(2)}\left(\begin{bmatrix} \lambda_1 & z \\ 0 & \lambda_2 \end{bmatrix}, \rho\right)$. In the present section we are going to generalize these results to pseudospectra of block triangular matrices depending on several parameters. First, we treat the 1-parameter case.

Proposition 5.6.1 *Let $A_k \in \mathbb{C}^{n_k\times n_k}$, $k = 1, 2$ and let $A_3 \in \mathbb{C}^{n_1\times n_2}\setminus\{0\}$. Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be absolute norms on $\mathbb{C}^{(n_1+n_2)\times(n_1+n_2)}$ and let $\rho > 0$. Then the following holds.*

(a) *For all $z, w \in \mathbb{C}$:*

$$\textit{if} \quad |z| \le |w| \quad \textit{then} \quad \sigma_{\mathbb{C}}^{(\alpha,\beta)}\left(\begin{bmatrix} A_1 & zA_3 \\ 0 & A_2 \end{bmatrix}, \rho\right) \subseteq \sigma_{\mathbb{C}}^{(\alpha,\beta)}\left(\begin{bmatrix} A_1 & wA_3 \\ 0 & A_2 \end{bmatrix}, \rho\right).$$

If $|z| = |w|$ then the inclusion is an equality.

(*b*) *To each bounded set $B \subset \mathbb{C}$ there exists a $z \in \mathbb{C}$ such that $B \subseteq \sigma_{\mathbb{C}}^{(\alpha,\beta)}\left(\begin{bmatrix} A_1 & zA_3 \\ 0 & A_2 \end{bmatrix}, \rho\right)$.*

Note that for $z \neq 0$, $\begin{bmatrix} A_1 & zA_3 \\ 0 & A_2 \end{bmatrix} = \begin{bmatrix} I_{n_1} & 0 \\ 0 & z^{-1}I_{n_2} \end{bmatrix} \begin{bmatrix} A_1 & A_3 \\ 0 & A_2 \end{bmatrix} \begin{bmatrix} I_{n_1} & 0 \\ 0 & zI_{n_2} \end{bmatrix}$. Thus, roughly speaking, part (*b*) of the above Proposition states that the eigenvalues of a 2×2 block triangular matrix can be made arbitrarily sensitive to perturbations by diagonal scaling.

The proof of Proposition 5.6.1 is based on the following lemma.

Lemma 5.6.2 *Let $M_k \in \mathbb{C}^{n_k \times n_k}$ be nonsingular matrices, $k = 1, 2$, and let $M_3 \in \mathbb{C}^{n_1 \times n_2} \setminus \{0\}$. Let $\|\cdot\|$ be an arbitrary norm on $\mathbb{C}^{(n_1+n_2)\times(n_1+n_2)}$. Then the function*

$$f : \mathbb{C} \to (0, \infty) \qquad f(z) := \left\| \begin{bmatrix} M_1 & zM_3 \\ 0 & M_2 \end{bmatrix}^{-1} \right\|$$

is convex. Moreover, $\lim_{|z|\to\infty} f(z) = \infty$. Suppose that $\|\cdot\|$ is an operator norm subordinate to absolute norms on $\mathbb{C}^{(n_1+n_2)}$. Then we have for all $z, w \in \mathbb{C}$,

(*i*) *if $|z| = |w|$ then $f(z) = f(w)$,*

(*ii*) *if $|z| \leq |w|$ then $f(z) \leq f(w)$.*

Proof: Let $F(z) := \begin{bmatrix} M_1 & zM_3 \\ 0 & M_2 \end{bmatrix}^{-1}$, $z \in \mathbb{C}$. Then the following holds.

(*iii*) $F(z) = \begin{bmatrix} M_1^{-1} & -zM_1^{-1}M_3M_2^{-1} \\ 0 & M_2^{-1} \end{bmatrix} = \begin{bmatrix} M_1^{-1} & 0 \\ 0 & M_2^{-1} \end{bmatrix} - z \begin{bmatrix} 0 & M_1^{-1}M_3M_2^{-1} \\ 0 & 0 \end{bmatrix}$.

(*iv*) If $z \neq 0$ then $F(z) = D_z^{-1}F(|z|)D_z$, where $D_z := \operatorname{diag}(I_{n_1}, \frac{z}{|z|}I_{n_2})$.

By (*iii*), F is an affine function. Hence the composition $f = \|\cdot\| \circ F$ is convex. From the inequality $|f(z)| \geq |z| \left\| \begin{bmatrix} 0 & M_1^{-1}M_3M_2^{-1} \\ 0 & 0 \end{bmatrix} \right\| - \left\| \begin{bmatrix} M_1^{-1} & 0 \\ 0 & M_2^{-1} \end{bmatrix} \right\|$ it follows that $\lim_{z\to\infty} f(z) = \infty$. If $\|\cdot\|$ is an operator norm subordinate to absolute norms then (*iv*) yields that $f(z) = f(|z|)$ for all $z \in \mathbb{C}$. Thus claim (*i*) holds. In particular, $f(z) = f(-z)$. From the latter it follows by convexity that $f(tz) \leq f(z)$ for all $t \in [-1, 1]$. This together with (*i*) implies (*ii*). $\square$

Proof of Proposition 5.6.1. For the sake of brevity we set

$$\mathcal{P}_z := \sigma_{\mathbb{C}}^{(\alpha,\beta)}\left(\begin{bmatrix} A_1 & zA_3 \\ 0 & A_2 \end{bmatrix}, \rho\right), \qquad g(s, z) := \left\| \begin{bmatrix} A_1 - sI_{n_1} & zA_3 \\ 0 & A_2 - sI_{n_2} \end{bmatrix}^{-1} \right\|_{\beta,\alpha}.$$

Claim (*a*) of Proposition 5.6.1 is an immediate consequence of Lemma 5.6.2 and the equivalence $s \in \mathcal{P}_z \Leftrightarrow g(s, z) \geq \rho^{-1}$, which holds for all $s \in \mathbb{C} \setminus (\sigma(A_1) \cup \sigma(A_2))$.

(*b*). We may assume that the bounded set $B \subset \mathbb{C}$ is closed. Then $K := B \setminus (\operatorname{int} \mathcal{P}_0)$ is compact and $K \cap (\sigma(A_2) \cup \sigma(A_2)) = \emptyset$. By Lemma 5.6.2 we have that $\lim_{z\to\infty} g(s, z) = \infty$ for all $s \in K$. From the compactness of K and the continuity of g it follows that $\lim_{z\to\infty} \min_{s\in K} g(s, z) = \infty$. Thus to each $\rho > 0$ there is an $r > 0$ such that for all $z \in \mathbb{C}$ with $|z| \geq r$ the inequality

$\rho^{-1} \le \min_{s\in K} g(s,z)$ holds. The latter is equivalent to the statement that $K \subseteq \mathcal{P}_z$. By (a) we also have that $\mathcal{P}_0 \subseteq \mathcal{P}_z$. Hence, $B \subseteq \mathcal{P}_z$. □

In order to discuss block structures of more general type we introduce the following notation.

Notation 5.6.3 *Let A be a block upper triangular matrix of the form*

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1m} \\ & A_{22} & \dots & A_{2m} \\ & & \ddots & \vdots \\ & 0 & & A_{mm} \end{bmatrix} = [A_{jk}]_{j,k\in\underline{m}}, \quad A_{jk} \in \mathbb{C}^{n_j \times n_k}, \quad A_{jk} = 0 \text{ for } j > k,$$

Then we denote by $D_A := \operatorname{diag}(A_{11}, \dots, A_{mm})$ its diagonal part, by $R_A := A - D_A$ its strictly upper triangular part and by $n_A := \sum_{k=1}^m n_k$ its dimension.

In this notation Proposition 5.6.1 in particular states that for $m = 2$ the implication

$$\text{if } |z| \le |w| \text{ then } \sigma_{\mathbb{C}}^{(\alpha,\beta)}(D_A + zR_A, \rho) \subseteq \sigma_{\mathbb{C}}^{(\alpha,\beta)}(D_A + wR_A, \rho)$$

holds for all $z, w \in \mathbb{C}$. This implication is no longer valid in general if $m \ge 3$. We have the following counterexample. Let

$$A = \begin{bmatrix} -3 & 26 & -60 \\ 0 & 0 & 5 \\ 0 & 0 & 3 \end{bmatrix}. \tag{5.19}$$

Then $\sigma_{\min}(A - 5I_3) = 0.4217$ and $\sigma_{\min}(D_A + 2R_A - 5I_3) = 0.5650$. Thus $5 \in \sigma_{\mathbb{C}}^{(2)}(A, 0.5)$ but $5 \notin \sigma_{\mathbb{C}}^{(2)}(D_A + 2R_A, 0.5)$. The pseudospectra are shown in Figure 5.5. However, the following analog to claim (b) of Proposition 5.6.1 holds.

Proposition 5.6.4 *Let $A = [A_{jk}]_{j,k\in\underline{m}}$ be a block upper triangular matrix, $m \ge 2$. Let $\|\cdot\|_\alpha, \|\cdot\|_\beta$ be absolute norms on $\mathbb{C}^{n_A \times n_A}$ and let $\rho > 0$. If $R_A \ne 0$ then to each bounded set $B \subset \mathbb{C}$ there exists an $r > 0$ such that $B \subseteq \sigma_{\mathbb{C}}^{(\alpha,\beta)}(D_A + zR_A, \rho)$ for all $z\in\mathbb{C}$ with $|z| \ge r$.*

The proof of Proposition 5.6.4 will be given at the end of this section.

Next, we extend Proposition 5.6.1 for block triangular matrices depending on several parameters. Let us first introduce a notation.

Notation 5.6.5 *For a block upper triangular matrix $A\in\mathbb{C}^{n_A \times n_A}$ as in notation 5.6.3 and for $z = [z_1, \dots, z_{m-1}]^T \in \mathbb{C}^{m-1}$ let*

$$A_z := \begin{bmatrix} A_{11} & z_1 A_{12} & z_1 z_2 A_{13} & z_1 z_2 z_3 A_{13} & \dots & z_1 z_2 z_3 \dots z_{m-1} A_{1m} \\ & A_{22} & z_2 A_{23} & z_2 z_3 A_{24} & & z_2 z_3 \dots z_{m-1} A_{2m} \\ & & A_{33} & z_3 A_{34} & & z_3 \dots z_{m-1} A_{3m} \\ & & & \ddots & & \vdots \\ & 0 & & & & \\ & & & & A_{m-1,m-1} & z_{m-1} A_{m-1\,m} \\ & & & & & A_{mm} \end{bmatrix}.$$

To be precise, the block in the jth block row and kth block column of A_z is 0 if $j > k$, and $\left(\prod_{i=j}^{k-1} z_i\right) A_{jk}$ otherwise.

If all entries of z are nonzero then we have $A_z = D_z^{-1} A D_z$, where $D_z := \mathrm{diag}(d_1 I_{n_1}, \ldots, d_m I_{n_m})$ with $d_k := \prod_{i=1}^{k-1} z_i$ for all $k \in \underline{m}$.

Theorem 5.6.6 *Let $A = [A_{jk}]_{j,k\in\underline{m}}$ be a block upper triangular matrix, $m \geq 2$. Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be absolute norms on $\mathbb{C}^{n_A}$, and let $\rho > 0$. Then the following holds.*

(a) *For all $z, w \in \mathbb{C}^{m-1}$,*

(i) *if $|z| \leq |w|$ then $\sigma_{\mathbb{C}}^{(\alpha,\beta)}(A_z, \rho) \subseteq \sigma_{\mathbb{C}}^{(\alpha,\beta)}(A_w, \rho)$;*

(ii) *if $|z| = |w|$ then $\sigma_{\mathbb{C}}^{(\alpha,\beta)}(A_z, \rho) = \sigma_{\mathbb{C}}^{(\alpha,\beta)}(A_w, \rho)$.*

(b) *If at least one off-diagonal block of A is nonzero then to each bounded set $B \subset \mathbb{C}$ there exists a $z \in \mathbb{C}^{m-1}$ such that $B \subseteq \sigma_{\mathbb{C}}^{(\alpha,\beta)}(A_z, \rho)$.*

Proof: This follows from Proposition 5.6.1 and the fact that for each index $i \leq m-1$ the matrix A_z can be written in the form $A_z = \begin{bmatrix} \hat{A}_{1i}(z) & z_i \, \hat{A}_{3i}(z) \\ 0 & \hat{A}_{2i}(z) \end{bmatrix}$, where the matrices $\hat{A}_{ji}(z)$ do not depend on the component z_i of z. $\square$

Below we list some consequences of Theorem 5.6.6. The underlying norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ are absolute.

Corollary 5.6.7 *Let $A = [A_{jk}]_{j,k\in\underline{m}}$ be a block upper triangular matrix. Then $\sigma_{\mathbb{C}}^{(\alpha,\beta)}(D_A, \rho) \subseteq \sigma_{\mathbb{C}}^{(\alpha,\beta)}(A, \rho)$. In particular, we have for every $p \in [1, \infty]$, $\bigcup_{j\in\underline{m}} \sigma_{\mathbb{C}}^{(p)}(A_{jj}, \rho) \subseteq \sigma_{\mathbb{C}}^{(p)}(A, \rho)$.*

Corollary 5.6.8 *Let $n \geq 2$, $\rho > 0$. Then the radius $r_n^{(\alpha,\beta)}(\rho, d)$ of the disk $\sigma_{\mathbb{C}}^{(\alpha,\beta)}(J_n(\lambda, d), \rho)$ is an increasing function of $d \geq 0$. Moreover, $\lim_{d\to\infty} r_n^{(\alpha,\beta)}(\rho, d) = \infty$.*

Corollary 5.6.9 *Let $A \in \mathbb{C}^{n\times n}$, $\rho > 0$. If A is not a scalar multiple of I_n then to each bounded subset $B \subset \mathbb{C}$ there exists a nonsingular matrix $S \in \mathbb{C}^{n\times n}$ such that $B \subset \sigma_{\mathbb{C}}^{(\alpha,\beta)}(S^{-1}AS, \rho)$.*

Proof of Corollary 5.6.9: If A is not diagonalizable then the claim follows from Corollary 5.6.8 and (5.10). Suppose that A is diagonalizable but not a scalar multiple of the identity matrix. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A. We may assume that $\lambda_1 \neq \lambda_2$. Then for all $z \in \mathbb{C}$, A is similar to the matrix $A(z) = \mathrm{diag}\left(\begin{bmatrix} \lambda_1 & z \\ 0 & \lambda_2 \end{bmatrix}, \lambda_3, \ldots, \lambda_n \right)$. The claim follows from claim (b) of Proposition 5.6.1. $\square$

We now give the proof of Proposition 5.6.4. To this end we need the following lemma which will also be used in the next two sections.

Lemma 5.6.10 *Let $A = [A_{jk}]_{j,k\in\underline{m}}$ be a block upper triangular matrix. For $s \in \mathbb{C} \setminus \sigma(A)$ let $E_s = R_A(sI_{n_A} - D_A)^{-1}$ and $\nu_s = \min\{\, \ell \in \mathbb{N} \mid E_s^\ell = 0 \,\}$. Then $\nu_s \leq m$ and for all $z \in \mathbb{C}$,*

$$(sI_{n_A} - D_A - zR_A)^{-1} = (sI_{n_A} - D_A)^{-1} \sum_{\ell=0}^{\nu_s - 1} z^\ell E_s^\ell. \tag{5.20}$$

In particular,

$$(sI_{n_A} - A)^{-1} = (sI_{n_A} - D_A)^{-1} \sum_{\ell=0}^{m-1} E_s^\ell. \tag{5.21}$$

Proof: Let $(E_s)_{jk}$ and $(E_s^\ell)_{jk}$ denote the block in the jth block row and kth block column of E_s and E_s^ℓ respectively. We have $(E_s)_{jk} = A_{jk}(sI - A_{kk})^{-1}$ for $j \neq k$ and $(E_s)_{kk} = 0$. Hence, $(E_s)_{jk} = 0$ for $j \geq k$. Thus for $\ell \geq 2$,

$$(E_s^\ell)_{jk} \;=\; \sum_{j_2=1}^{m} \cdots \sum_{j_\ell=1}^{m} (E_s)_{jj_2}(E_s)_{j_2j_3}\ldots(E_s)_{j_\ell k} = \sum_{j<j_2<j_3\ldots j_\ell<k} (E_s)_{jj_2}(E_s)_{j_2j_3}\ldots(E_s)_{j_\ell k}.$$

This implies $E_s^m = 0$. Thus $\nu_s \leq m$ and $(I_{n_A} - zE_s)^{-1} = \sum_{\ell=0}^{\nu_s-1}(zE_s)^\ell$ for all $z\in\mathbb{C}$. Now, the formula (5.20) follows from the relation $(sI_{n_A} - D_A - zR_A)^{-1} = (sI_{n_A} - D_A)^{-1}(I_{n_A} - zE_s)^{-1}$ $\square$

Proof of Proposition 5.6.4. For $s\in\mathbb{C}\backslash\sigma(A)$, $z\in\mathbb{C}$ set $g(s,z) := \|(sI_{n_A} - D_A - zR_A)^{-1}\|_{\beta,\alpha}$. Then (5.20) yields

$$\|g(s,z)\|_{\beta,\alpha} \geq |z|^{\nu_s-1}\|(sI_{n_A} - D_A)^{-1}E_s^{\nu_s-1}\|_{\beta,\alpha} - \sum_{\ell=0}^{\nu_s-2} |z|^\ell \|(sI_{n_A} - D_A)^{-1}E_s^\ell\|_{\beta,\alpha}.$$

Since $\|(sI_{n_A} - D_A)^{-1}E_s^{\nu_s-1}\|_{\beta,\alpha} \neq 0$, we have $\lim_{|z|\to\infty} g(s,z) = \infty$ for all $s \notin \sigma(A)$. Now, the same compactness argument as in the proof of Proposition 5.6.1 yields the result. $\square$

5.7 Henrici's eigenvalue inclusion theorem

In the textbooks [10, Problem VIII.6.4], and [86, Theorem 1.9] one can find the following statement, which is a reformulation of an eigenvalue inclusion theorem by P. Henrici [38].

Theorem 5.7.1 *Let $A, \widetilde{A}\in\mathbb{C}^{n\times n}$ and suppose that $\delta_2(A) \neq 0$. Then for every $\widetilde{\lambda}\in\sigma(\widetilde{A})$ there is a $\lambda\in\sigma(A)$ such that*

$$\frac{\left(\frac{|\lambda-\widetilde{\lambda}|}{\delta_2(A)}\right)^n}{1 + \left(\frac{|\lambda-\widetilde{\lambda}|}{\delta_2(A)}\right) + \ldots + \left(\frac{|\lambda-\widetilde{\lambda}|}{\delta_2(A)}\right)^{n-1}} \leq \frac{\|A - \widetilde{A}\|_2}{\delta_2(A)}.$$

In this section we give a new reformulation of Henrici's theorem in terms of pseudospectra. Moreover, we provide a variant of that theorem for block triangular matrices. To this end we introduce the following functions.

Definition 5.7.2 *For $m \geq 2$, $\rho > 0$ and $d \geq 0$ let $h_m(\rho,d)$ denote the positive zero of the polynomial*

$$\widetilde{h}_{m,\rho,d}(x) := \rho(d^{m-1} + xd^{m-2} + \ldots + x^{m-2}d + x^{m-1}) - x^m = \begin{cases} \rho m d^{m-1} - d^m & \text{if } x = d, \\ \rho\frac{x^m - d^m}{x-d} - x^m & \text{otherwise.} \end{cases}$$

We call the map $h_m : (0,\infty)\times[0,\infty) \to (0,\infty)$ the mth Henrici-function.

We discuss the Henrici-functions in Proposition 5.7.6 below. In particular we will show that these functions are well defined, i.e. that each of polynomials $\widetilde{h}_{m,\rho,d}(x)$ has exactly one positive zero. From the facts that $\lim_{x\to\infty} h_{m,\rho,d}(x) = -\infty$ and that $\widetilde{h}_{m,\rho,d}(x) > 0$ for small enough $x > 0$ one concludes that

$$\widetilde{h}_{m,\rho,d}(x) \begin{cases} > 0 & \text{if } x < h_m(\rho, d), \\ < 0 & \text{if } x > h_m(\rho, d). \end{cases} \tag{5.22}$$

We now give our reformulation of Henrici's theorem.

Theorem 5.7.3 *Let $\rho > 0$ and $A \in \mathbb{C}^{n\times n}$. Then for every normalized unitarily invariant norm,*

$$\sigma_{\mathbb{C}}^{\|\cdot\|}(A, \rho) \subseteq \bigcup_{\lambda\in\sigma(A)} \mathcal{D}(\,\lambda\,,\, h_n(\rho, \delta_2(A))\,).$$

By claim (i) of Proposition 5.7.6 below we have that $h_n(\rho, \delta_2(A)) < \rho + \delta_2(A)$. Thus Henrici's theorem is an improvement of Proposition 5.3.7. However, for large n the improvement is rather weak. This can be seen as follows. Theorem 5.7.3 implies that $h_n(\rho, d) \geq r_n^{(2,2)}(\rho, d)$, where $r_n^{(2,2)}(\rho, d)$ is the radius of the pseudospectrum (to the level ρ) of the Jordan matrix $J_n(\lambda, d)$ defined in Section 5.3. By Proposition 5.3.7, the radius $r_n^{(2,2)}(\rho, d)$ converges to $\rho + d$ as n tends to infinity. Moreover, Figure 5.1 indicates that the convergence is fast. Some graphs of Henrici-functions are shown in the Figure 5.6.

According to claim (i) of Proposition 5.2.5 we have for every block upper triangular matrix A as in Notation 5.6.3 and every norm on $\mathbb{C}^{n_A\times n_A}$,

$$\sigma_{\mathbb{C}}^{\|\cdot\|}(A, \rho) \subseteq \sigma_{\mathbb{C}}^{\|\cdot\|}(D_A,\, \rho + \|R_A\|\,),$$

where D_A and R_A denote the block diagonal part and the strictly block upper triangular part of A, respectively. Henrici's theorem for block triangular matrices improves this upper bound for the case that the underlying norm is induced by an operator norm.

Theorem 5.7.4 *Let $A = [A_{jk}]_{j,k\in\underline{m}}$ be a block upper triangular matrix. Let $\|\cdot\| = \|\cdot\|_{\alpha,\alpha}$ be an operator norm on $\mathbb{C}^{n_A\times n_A}$ subordinate to an arbitrary norm $\|\cdot\|_\alpha$ on $\mathbb{C}^{n_A}$. Then we have for every $\rho > 0$,*

$$\sigma_{\mathbb{C}}^{\|\cdot\|}(A, \rho) \subseteq \sigma_{\mathbb{C}}^{\|\cdot\|}(D_A,\, h_m(\rho, \|R_A\|)\,).$$

Proof of Theorem 5.7.4 and Theorem 5.7.3: By Lemma 5.6.10 we have for any $s \notin \sigma(A)$,

$$(sI_{n_A} - A)^{-1} = (sI_{n_A} - D_A)^{-1} \sum_{\ell=0}^{m-1} (R_A(sI_{n_A} - D_A)^{-1})^\ell.$$

Let $s \in \sigma_{\mathbb{C}}^{\|\cdot\|}(A, \rho) \setminus \sigma(A)$, and set $x := \|(sI_{n_A} - D_A)^{-1}\|^{-1}$, $d := \|R_A\|$. Then

$$\rho^{-1} \leq \|(sI_{n_A} - A)^{-1}\| \leq \sum_{\ell=0}^{m-1} \|(sI_{n_A} - D_A)^{-1}\|^{\ell+1} \|R_A\|^\ell = \sum_{\ell=0}^{m-1} x^{-\ell-1} d^\ell. \tag{5.23}$$

Multiplication of (5.23) with ρx^m and subtraction of x^m yields

$$0 \leq \rho \sum_{\ell=0}^{m-1} x^{m-1-j} d^\ell - x^m = \widetilde{h}_{m,\rho,d}(x).$$

Thus, by (5.22), $x \leq h_m(\rho, d)$. The latter implies that $s \in \sigma_{\mathbb{C}}^{\|\cdot\|}(D_A,\, h_m(\rho, \|R_A\|)\,)$. Thus Theorem 5.7.4 is shown.

Now, the proof of Theorem 5.7.3 is as follows. Let $\Lambda + R$ be a Schur form of $A \in \mathbb{C}^{n\times n}$ such that $\|R\|_2 = \delta_2(A)$. Then for every normalized unitarily invariant norm,

$$\sigma_{\mathbb{C}}^{\|\cdot\|}(A,\rho) = \sigma_{\mathbb{C}}^{(2)}(\Lambda + R, \rho) \subseteq \sigma_{\mathbb{C}}^{(2)}(\Lambda, h_n(\rho, \|R\|_2)) = \bigcup_{\lambda\in\sigma(A)} \mathcal{D}(\,\lambda\,,\, h_n(\rho, \delta_2(A))\,). \qquad \square$$

By combining Theorem 5.7.4, Corollary 5.6.7 and Proposition 5.2.4 we obtain

Corollary 5.7.5 *Let $A = [A_{jk}]_{j,k\in\underline{m}}$ be a block upper triangular matrix. Then we have for any $p\in[1,\infty]$ and any $\rho > 0$,*

$$\bigcup_{j\in\underline{m}} \sigma_{\mathbb{C}}^{(p)}(A_{jj},\rho) \subseteq \sigma_{\mathbb{C}}^{(p)}(A,\rho) \subseteq \bigcup_{j\in\underline{m}} \sigma_{\mathbb{C}}^{(p)}(A_{jj}, h_m(\rho, \|R_A\|_p)\,).$$

In the next proposition we list some basic properties of the Henrici-functions.

Proposition 5.7.6 *The Henrici-functions h_m are well-defined, i.e. each of the polynomials $\widetilde{h}_{m,\rho,d}$ (where $m \geq 2$, $(\rho, d)\in(0,\infty)\times[0,\infty)$) has exactly one zero in $(0,\infty)$, namely*

$$h_m(\rho, d) = \begin{cases} \rho & \text{if } d = 0, \\ \frac{d}{\phi_m^{-1}\left(\frac{d}{\rho}\right)} & \text{if } d > 0, \end{cases} \tag{5.24}$$

where ϕ_m^{-1} denotes the inverse of the strictly increasing function

$$\phi_m : (0,\infty) \to (0,\infty), \qquad \phi_m(x) := \sum_{\ell=1}^{m} x^\ell.$$

The Henrici-functions h_m are real analytic on $(0,\infty)\times(0,\infty)$ and continuous on $(0,\infty)\times[0,\infty)$. They have the following properties.

(*i*) *For all $m \geq 2$ and all $(\rho, d)\in(0,\infty)\times[0,\infty)$,*

$$\rho \leq h_m(\rho, d) < \rho + d.$$

(*ii*) *The function $(m, \rho, d) \mapsto h_m(\rho, d)$ is increasing in all variables. Precisely,*

$$\begin{array}{llll} \text{if} & \rho_1 < \rho_2 & \text{then} & h_m(\rho_1, d) < h_m(\rho_2, d), \\ \text{if} & d_1 < d_2 & \text{then} & h_m(\rho, d_1) < h_m(\rho, d_2), \\ \text{if} & m_1 < m_2 & \text{then} & h_{m_1}(\rho, d) < h_{m_2}(\rho, d) \end{array}$$

for every $m, m_1, m_2 \geq 2$, $\rho, \rho_1, \rho_2 \in(0,\infty)$, $d, d_1, d_2\in[0,\infty)$.

(*iii*) *The functions*

$$\rho \longmapsto h_m(\rho, d) - \rho, \qquad \rho\in(0,\infty)$$

are increasing for all $d > 0$, $m \geq 2$. Moreover, $\lim_{\rho\to\infty}(h_m(\rho, d) - \rho) = d$.

(*iv*) *For all $(\rho, d)\in(0,\infty)\times[0,\infty)$ we have $\lim_{m\to\infty} h_m(\rho, d) = \rho + d$.*

(v) *The functions*

$$d \longmapsto \frac{h_m(\rho, d)}{d} \qquad d \in (0, \infty)$$

are decreasing for all $\rho > 0$, $m \geq 2$. *Moreover,*

$$\lim_{d\to\infty} \frac{h_m(\rho, d)}{d} = \lim_{d\to\infty} \frac{h_m(\rho, d)}{\rho + d} = 0.$$

(vi) *For all* $(d, \rho) \in [0, \infty) \times (0, \infty)$, $m \geq 2$, *we have*

$$d^{1-\frac{1}{m}} \rho^{\frac{1}{m}} < h_m(\rho, d) \tag{5.25}$$

and

$$\text{if} \quad \rho < \frac{d}{m} \quad \text{then} \quad h_m(\rho, d) \leq m^{\frac{1}{m}} d^{1-\frac{1}{m}} \rho^{\frac{1}{m}}. \tag{5.26}$$

Remark 5.7.7 *The functions* ϕ_m *occur in Henrici's paper [38]. The estimate (5.26) can be found in [86, Corollary,1.10]. The estimates (5.25) and (5.26) together imply that the functions* $\rho \mapsto h_m(\rho, d)$ *are of the order* $O(\rho^{1/m})$.

Before we give the (technical) proof of Proposition 5.7.6 we show, how Theorem 5.7.1 follows from Theorem 5.7.3 and the formula (5.24). Let $A, \widetilde{A} \in \mathbb{C}^{n\times n}$ and assume that $\delta_2(A) \neq 0$ and $\widetilde{A} \neq A$. Let $\widetilde{\lambda}$ be an eigenvalue of $\widetilde{A}$ which is not an eigenvalue of A. Set $\rho := \|A - \widetilde{A}\|_2$. Then $\widetilde{\lambda} \in \sigma_{\mathbb{C}}^{(2)}(A, \rho)$. Thus, by Theorem 5.7.3 and formula (5.24) there exists an eigenvalue λ of A such that

$$|\lambda - \widetilde{\lambda}| \leq h_m(\rho, \delta_2(A)) = \frac{\delta_2(A)}{\phi_n^{-1}\left(\frac{\delta_2(A)}{\rho}\right)}.$$

Thus $\phi_n^{-1}\left(\frac{\delta_2(A)}{\rho}\right) \leq \frac{\delta_2(A)}{|\lambda - \widetilde{\lambda}|}$. Since the function ϕ_n is increasing, it follows that

$$\frac{\delta_2(A)}{\rho} = \phi_n\left(\phi_n^{-1}\left(\frac{\delta_2(A)}{\rho}\right)\right) \leq \phi_n\left(\frac{\delta_2(A)}{|\lambda - \widetilde{\lambda}|}\right) = \sum_{\ell=1}^{n}\left(\frac{\delta_2(A)}{|\lambda - \widetilde{\lambda}|}\right)^{\ell}. \tag{5.27}$$

By multiplying (5.27) with the factor $\left(\frac{|\lambda-\widetilde{\lambda}|}{\delta_2(A)}\right)^n$ one obtains $\left(\frac{|\lambda-\widetilde{\lambda}|}{\delta_2(A)}\right)^n \frac{\delta_2(A)}{\rho} \leq \sum_{\ell=0}^{n-1}\left(\frac{|\lambda-\widetilde{\lambda}|}{\delta_2(A)}\right)^{\ell}$. But this is equivalent to the inequality in Theorem 5.7.1.

Proof of Proposition 5.7.6: The functions $\phi_m : (0, \infty) \to (0, \infty)$, $m \geq 2$, are strictly increasing and surjective. Moreover, they are analytic and the derivatives ϕ_m' have no zeros. Therefore, the inverse functions $\phi_m^{-1} : (0, \infty) \to (0, \infty)$ exist and are analytic. Let $x, \rho, d > 0$. Then

$$\begin{aligned}
\widetilde{h}_{m,\rho,d}(x) = 0 \quad &\Leftrightarrow \quad 0 = \frac{d}{x^m}\widetilde{h}_{m,\rho,d}(x) = \frac{d}{x^m}\left(\rho\sum_{\ell=0}^{m-1} d^{\ell} x^{m-1-\ell} - x^m\right) = \rho\sum_{\ell=1}^{m}\left(\frac{d}{x}\right)^{\ell} - d \\
&\Leftrightarrow \quad \frac{d}{\rho} = \sum_{\ell=1}^{m}\left(\frac{d}{x}\right)^{\ell} = \phi_m\left(\frac{d}{x}\right) \\
&\Leftrightarrow \quad \phi_m^{-1}\left(\frac{d}{\rho}\right) = \frac{d}{x} \\
&\Leftrightarrow \quad x = \frac{d}{\phi_m^{-1}\left(\frac{d}{\rho}\right)}.
\end{aligned}$$

Thus $h_m(\rho,d) = \frac{d}{\phi_m^{-1}\left(\frac{d}{\rho}\right)}$ is the only positive zero of $\widetilde{h}_{m,\rho,d}$ and the functions h_m are analytic on $(0,\infty)\times(0,\infty)$. Furthermore we have $h_m(\rho,0) = \rho$ since $\rho > 0$ is the only positive zero of the polynomial $x \mapsto \widetilde{h}_{m,\rho,0}(x) = \rho x^{m-1} - x^m$. In order to prove that h_m is continuous on $(0,\infty)\times[0,\infty)$ it remains to show that $\lim_{(\rho,d)\to(\rho_0,0)} h_m(\rho,d) = \rho_0$ for all $\rho_0 > 0$. However, we have for every $y > 0$ that $y = \sum_{\ell=1}^m \left(\phi_m^{-1}(y)\right)^\ell > \phi_m^{-1}(y) > 0$. Therefore

$$\lim_{y\to 0}\frac{y}{\phi_m^{-1}(y)} = \lim_{y\to 0}\left(1 + \sum_{\ell=1}^{m-1}\left(\phi_m^{-1}(y)\right)^\ell\right) = 1.$$

Thus for every $\rho_0 > 0$,

$$\lim_{(\rho,d)\to(\rho_0,d)} h_m(\rho,d)) = \lim_{(\rho,d)\to(\rho_0,0)} \rho\,\frac{\left(\frac{d}{\rho}\right)}{\phi_m^{-1}\left(\frac{d}{\rho}\right)} = \rho_0.$$

(i). Let $\rho, d > 0$. We have $\widetilde{h}_{m,\rho,d}(0) = \rho d^{m-1} > 0$ and $\lim_{x\to\infty}\widetilde{h}_{m,\rho,d}(x) = -\infty$. Since $h_m(\rho,d)$ is the only positive zero of the continuous function $x \mapsto \widetilde{h}_{m,\rho,d}(x)$ it follows that

$$\begin{array}{ll} \widetilde{h}_{m,\rho,d}(x) > 0 & \text{if } x\in(\,0, h_m(\rho,d)\,), \\ \widetilde{h}_{m,\rho,d}(x) < 0 & \text{if } x\in(\,h_m(\rho,d),\infty\,). \end{array} \tag{5.28}$$

Now (i) follows from the inequalities

$$\begin{aligned} \widetilde{h}_{m,\rho,d}(\rho) &= \rho(d^{m-1} + \rho d^{m-2} + \ldots + \rho^{m-2}d + \rho^{m-1}) - \rho^m > 0, \\ \widetilde{h}_{m,\rho,d}(\rho+d) &= \rho\frac{(\rho+d)^m - d^m}{(\rho+d)-d} - (\rho+d)^m = -d^m < 0. \end{aligned}$$

(ii). For every $x\in(0,\infty)$ the map

$$(\rho,d)\mapsto \widetilde{h}_{m,\rho,d}(x), \qquad (\rho,d)\in(0,\infty)\times[0,\infty)$$

is strictly increasing in the variables ρ and d. Thus, if $x = h_m(\rho_1,d_1)$ and $\rho_1 < \rho_2$, $d_1 < d_2$, then $\widetilde{h}_{m,\rho_2,d_1}(x) > 0$ and $\widetilde{h}_{m,\rho_1,d_2}(x) > 0$. Hence, by (5.28), $x < h_m(\rho_2,d_1)$ and $x < h_m(\rho_1,d_2)$. Let $m_1, m_2 \geq 2$ and $m_1 < m_2$. Then $\phi_{m_1}(x) < \phi_{m_2}(x)$ for all $x > 0$. It follows that $\phi_{m_1}^{-1}(y) > \phi_{m_2}^{-1}(y)$ for all $y > 0$. Thus, for $d > 0$,

$$h_{m_1}(\rho,d) = \frac{d}{\phi_{m_1}^{-1}\left(\frac{d}{\rho}\right)} < \frac{d}{\phi_{m_2}^{-1}\left(\frac{d}{\rho}\right)} = h_{m_2}(\rho,d).$$

(iii). We have for all $\rho, d > 0$ and all $y \geq 0$,

$$
\begin{aligned}
\widetilde{h}_{m,\rho,d}(\rho + y) &= \rho \sum_{\ell=0}^{m-1} d^{m-1-\ell}(\rho+y)^\ell - (\rho+y)^m \\
&= (\rho+y) \sum_{\ell=0}^{m-1} d^{m-1-\ell}(\rho+y)^\ell - (\rho+y)^m - y \sum_{\ell=0}^{m-1} d^{m-1-\ell}(\rho+y)^\ell \\
&= (\rho+y) \sum_{\ell=0}^{m-2} d^{m-1-\ell}(\rho+y)^\ell - y \sum_{\ell=0}^{m-1} d^{m-1-\ell}(\rho+y)^\ell \\
&= d \sum_{\ell=1}^{m-1} d^{m-1-\ell}(\rho+y)^\ell - y \sum_{\ell=0}^{m-1} d^{m-1-\ell}(\rho+y)^\ell \\
&= (d-y) \sum_{\ell=1}^{m-1} d^{m-1-\ell}(\rho+y)^\ell - y d^{m-1} \\
&= d^{m-1} \left((d-y) \sum_{\ell=1}^{m-1} \left(\frac{\rho+y}{d}\right)^\ell - y \right) \\
&= d^{m-1} \left((d-y)\, \phi_{m-1}\left(\frac{\rho+y}{d}\right) - y \right). \qquad (5.29)
\end{aligned}
$$

Now let $\rho_0, d > 0$ and $y := h_m(\rho_0, d) - \rho_0$, i.e. $\widetilde{h}_{m,\rho_0,d}(\rho_0 + y) = 0$. Then $d - y > 0$ by (i), and (5.29) yields that for all $\rho > \rho_0$,

$$\widetilde{h}_{m,\rho,d}(\rho + y) > 0.$$

By (5.28) this is equivalent to

$$\rho + y < h_m(\rho, d),$$

Hence,

$$y = h_m(\rho_0, d) - \rho_0 < h_m(\rho, d) - \rho.$$

So the function $\rho \mapsto h_m(\rho, d) - \rho$ is increasing. From (5.29) we obtain that for all $y \in [0, d)$,

$$\lim_{\rho \to \infty} \widetilde{h}_{m,\rho,d}(\rho + y) = \infty.$$

Hence, for large ρ, $\widetilde{h}_{m,\rho,d}(\rho + y) > 0$, and consequently $h_m(\rho, d) > \rho + y$. Thus

$$\lim_{\rho \to \infty} (h_m(\rho, d) - \rho) = d.$$

(iv). Let $\rho \geq 0$, $d > 0$ and $y \in (\max\{0, d - \rho\}, d)$. Then $d - y > 0$ and $\lim_{m \to \infty} \phi_{m-1}\left(\frac{\rho+y}{d}\right) = \infty$ since $\frac{\rho+y}{d} > 1$. From (5.29) we obtain that for some $m_0 \in \mathbb{N}$,

$$\widetilde{h}_{m,\rho,d}(\rho + y) > 0 \quad \text{for all } m \geq m_0.$$

Equivalently,

$$\rho + y < h_m(\rho, d) \quad \text{for all } m \geq m_0.$$

This and the fact that $h_m(\rho, d) < \rho + d$ yield $\lim_{m \to \infty} h_m(\rho, d) = \rho + d$.
(v) follows from (5.24) and the fact that the functions ϕ_m^{-1} are strictly increasing with $\lim_{y \to \infty} \phi_m^{-1}(y) =$

∞.
(vi). Let $x_1 := d\left(\frac{\rho}{d}\right)^{\frac{1}{m}}$. Then

$$\widetilde{h}_{m,\rho,d}(x_1) = \rho\sum_{\ell=0}^{m-1} x_1^\ell d^{m-1-\ell} - x_1^m = \rho\sum_{\ell=0}^{m-1} x_1^\ell d^{m-1-\ell} - \rho d^{m-1} > 0,$$

so $h_m(\rho,d) > x_1$. Suppose now that $\frac{\rho}{d} < \frac{1}{m}$ and let $x_2 := d\left(\frac{m\rho}{d}\right)^{\frac{1}{m}} < d$. Then

$$\widetilde{h}_{m,\rho,d}(x_2) = \rho\sum_{\ell=0}^{m-1} x_2^\ell d^{m-1-\ell} - x_2^m < \rho m d^{m-1} - x_2^m < 0,$$

so $h_m(\rho,d) < x_2$. □

5.8 A theorem of Ruhe

In [80] Ruhe proved a refinement of Henrici's theorem for the case that the distance of one eigenvalue of A to the other eigenvalues is large compared to $\delta_2(A)$. However, Ruhe's theorem does not seem to be fully correct. At least the proof in [80] is incomplete. In this section we provide a corrected version of Ruhe's theorem in the language of pseudospectra. Moreover, we generalize it to block triangular matrices and discuss its consequences. We need some notation.

Notation 5.8.1 *Let $B\in\mathbb{C}^{n\times n}$ be partitioned as follows.*

$$B = \begin{bmatrix} B_{11} & \dots & B_{1m} \\ \vdots & & \vdots \\ B_{m1} & \dots & B_{mm} \end{bmatrix}, \qquad B_{jk}\in\mathbb{C}^{n_j\times n_k}.$$

Then its norm compression with respect to a Hölder p-norm is given by

$$\langle B\rangle_p := \begin{bmatrix} \|B_{11}\|_p & \dots & \|B_{1m}\|_p \\ \vdots & & \vdots \\ \|B_{m1}\|_p & \dots & \|B_{mm}\|_p \end{bmatrix} \in\mathbb{R}^{m\times m}$$

It is easily verified that $\|B\|_p \le \|\langle B\rangle_p\|_p$ and that $\langle B\rangle_p \le C$ implies $\|\langle B\rangle_p\|_p \le \|C\|_p$. Our proof of Ruhe's theorem is based on the following estimate for the p-norm of a block triangular matrix.

Theorem 5.8.2 *Let $A = [A_{jk}]_{j,k\in\underline{m}}$ be a block upper triangular matrix. Then we have for every $s\in\mathbb{C}\setminus\sigma(A)$ and every $p\in[1,\infty]$,*

$$\left\|(sI_{n_A} - A)^{-1}\right\|_p \le \left\|(sI_{n_A} - D_A)^{-1}\right\|_p + \sum_{\ell=1}^{m-1} c_\ell(s)\,\left\|(\langle R_A\rangle_p)^\ell\right\|_p,$$

where

$$c_\ell(s) = \max_{1\le j_1<j_2\dots<j_\ell<j_{\ell+1}\le m} \prod_{\nu=1}^{\ell+1} \left\|(sI - A_{j_\nu j_\nu})^{-1}\right\|_p.$$

Proof: By Lemma 5.6.10 we have $(sI_{n_A} - A)^{-1} = (sI_{n_A} - D_A)^{-1} \sum_{\ell=0}^{m-1} E_s^\ell$, where $E_s = R_A(sI_{n_A} - D_A)^{-1}$. Thus,

$$\begin{aligned}
\|(sI_{n_A} - A)^{-1}\|_p &\le \|(sI_{n_A} - D_A)^{-1}\|_p + \sum_{\ell=1}^{m-1} \|(sI_{n_A} - D_A)^{-1} E_s^\ell\|_p \\
&\le \|(sI_{n_A} - D_A)^{-1}\|_p + \sum_{\ell=1}^{m-1} \|\langle (sI_{n_A} - D_A)^{-1} E_s^\ell \rangle_p\|_p. \qquad (5.30)
\end{aligned}$$

The block in the jth block row and kth block column of E_s is $(E_s)_{jk} = A_{jk}(sI - A_{kk})^{-1}$ for $j \ne k$ and $(E_s)_{kk} = 0$. Since E_s is strictly block upper triangular the block $(E_s^\ell)_{jk}$ in the jth block row and kth block column of E_s^ℓ satisfies

$$(E_s^\ell)_{jk} = \sum_{j<j_2<j_3\ldots j_\ell<k} (E_s)_{jj_2}(E_s)_{j_2j_3}\ldots(E_s)_{j_\ell k}, \qquad \ell \ge 2.$$

Thus, for $\ell \in \underline{m-1}$ the entry in the jth row and kth column of $\langle (sI_{n_A} - D_A)^{-1} E_s^\ell \rangle_p \in \mathbb{C}^{m\times m}$ is

$$\begin{aligned}
(\langle (sI_{n_A} - D_A)^{-1} E_s^\ell \rangle_p)_{jk} &= \|(sI_{n_j} - A_{jj})^{-1} (E_s^\ell)_{jk}\|_p \\
&\le \sum_{j<j_2<j_3\ldots j_\ell<k} \|(sI_{n_j} - A_{jj})^{-1} (E_s)_{jj_2}(E_s)_{j_2j_3}\ldots(E_s)_{j_\ell,k}\|_p \\
&\le \sum_{j<j_2<j_3\ldots j_\ell<k} g_{jj_1\ldots j_\ell k}\,(\|A_{jj_2}\|_p\,\|A_{j_2j_3}\|_p\ldots\|A_{j_\ell,k}\|_p), \qquad (5.31)
\end{aligned}$$

where

$$g_{jj_2\ldots j_\ell k} := \|(sI_{n_j} - A_{jj})^{-1}\|_p \left(\prod_{\nu=2}^{\ell} \|(sI_{n_{j_\nu}} - A_{j_\nu j_\nu})^{-1}\|_p\right) \|(sI_{n_k} - A_{kk})^{-1}\|_p.$$

We have $g_{jj_2\ldots j_\ell k} \le c_\ell(s)$. The matrix $\langle R_A\rangle_p = \langle A - D_A\rangle_p = \langle A\rangle_p - \langle D_A\rangle_p$ is strictly upper triangular. Hence the entry in the jth row and kth column of $(\langle R_A\rangle_p)^\ell$ is

$$(\,(\langle R_A\rangle_p)^\ell\,)_{jk} = \sum_{j<j_2<j_3\ldots j_\ell<k} (\|A_{jj_2}\|_p\,\|A_{j_2j_3}\|_p\ldots\|A_{j_\ell k}\|_p),$$

Thus (5.31) yields that $\langle (sI_{n_A} - D_A)^{-1} E_s^\ell\rangle_p \le c_\ell(s)\,(\langle R_A\rangle_p)^\ell$. The latter implies

$$\|\langle (sI_{n_A} - D_A)^{-1} E_s^\ell\rangle_p\|_p \le c_\ell(s)\|\,(\langle R_A\rangle_p)^\ell\,\|_p.$$

Now, the Theorem follows from the inequality (5.30). $\square$

Remark 5.8.3 *The proof above is inspired by the proofs in [36]. See also [37].*

By Henrici's Theorem for block triangular matrices (Theorem 5.7.4) we have for every Hölder p-norm

$$\sigma_{\mathbb{C}}^{(p)}(A,\rho) \subseteq \bigcup_{j\in\underline{m}} \sigma_{\mathbb{C}}^{(p)}(A_{jj},\, h_m(\rho, \|R_A\|_p)\,). \qquad (5.32)$$

The next theorem is a refinement of this result. It implies Ruhe's theorem (see Corollary 5.8.6 below). In order to simplify notation we set for $x \ge 0$,

$$\beta_m(x) := \sum_{\ell=0}^{m-1} x^\ell = \begin{cases} m & \text{if } x = 1, \\ \frac{1-x^m}{1-x} & \text{otherwise.} \end{cases}$$

Theorem 5.8.4 *Let $p \in [1, \infty]$ and let $A = [A_{jk}]_{j,k \in \underline{m}}$ be an block upper triangular matrix. Let $\rho > 0$ and $r > h_m(\rho, \|R_A\|_p)$. Let $i \in \underline{m}$ and suppose that*

$$\sigma_{\mathbb{C}}^{(p)}(A_{jj}, r) \cap \sigma_{\mathbb{C}}^{(p)}(A_{ii},\, h_m(\rho, \|R_A\|_p)\,) = \emptyset$$

for all $j \in \underline{m} \setminus \{i\}$. Then

$$\sigma_{\mathbb{C}}^{(p)}(A, \rho) \subseteq \sigma_{\mathbb{C}}^{(p)}\left(A_{ii},\, \beta_m\!\left(\tfrac{\|\langle R_A\rangle_p\|_p}{r}\right)\rho\right) \cup \bigcup_{j \in \underline{m} \setminus \{i\}} \sigma_{\mathbb{C}}^{(p)}(A_{jj},\, h_m(\rho, \|R_A\|_p)\,).$$

Proof: Suppose that $s \in \sigma_{\mathbb{C}}^{(p)}(A, \rho)$ but $s \notin \sigma_{\mathbb{C}}^{(p)}(A_{jj},\, h_m(\rho, \|R_A\|_p)\,)$ for all $j \neq i$. Then (5.32) yields that $s \in \sigma_{\mathbb{C}}^{(p)}(A_{ii},\, h_m(\rho, \|R_A\|_p)\,)$. Therefore $\|(sI - A_{ii})^{-1}\|_p > h_m(\rho, \|R_A\|_p)^{-1} > r^{-1}$. By assumption we have $s \notin \sigma_{\mathbb{C}}^{(p)}(A_{jj},\, r\,)$ for $j \neq i$. This implies that for $j \neq i$,

$$\|(sI - A_{jj})^{-1}\| < r^{-1} < \|(sI - A_{ii})^{-1}\|_p$$

Hence $\left\|(sI_n - D_A)^{-1}\right\|_p = \|(sI - A_{ii})^{-1}\|_p$, and the numbers

$$c_\ell(s) = \max_{1 \le j_1 < j_2 \ldots < j_{\ell+1} \le m} \prod_{\nu=1}^{\ell+1} \left\|(sI - A_{j_\nu})^{-1}\right\|_p$$

satisfy $c_\ell(s) \le r^\ell \|(sI - A_{ii})^{-1}\|_p$. Thus

$$\begin{aligned}
\rho^{-1} &\le \left\|(sI_n - A)^{-1}\right\|_p && (\text{since } s \in \sigma_{\mathbb{C}}^{(p)}(A, \rho)) \\
&\le \left\|(sI_n - D_A)^{-1}\right\|_p + \sum_{\ell=1}^{m-1} c_\ell(s) \left\|(\langle R_A\rangle_p)^\ell\right\|_p && (\text{by Theorem 5.8.2}) \\
&\le \left\|(sI_n - D_A)^{-1}\right\|_p + \sum_{\ell=1}^{m-1} c_\ell(s)\, \|\langle R_A\rangle_p\|_p^\ell \\
&= \|(sI - A_{ii})^{-1}\|_p\, \beta_m\!\left(\tfrac{\|\langle R_A\rangle_p\|_p}{r}\right).
\end{aligned}$$

Thus $\|(sI - A_{ii})^{-1}\|_p \ge \left(\beta_m\!\left(\frac{\|\langle R_A\rangle_p\|_p}{r}\right)\rho\right)^{-1}$. Thus $s \in \sigma_{\mathbb{C}}^{(p)}\left(A_{ii},\, \beta_m\!\left(\frac{\|\langle R_A\rangle_p\|_p}{r}\right)\rho\right)$. □

In order to discuss the just proved Theorem let us combine it with the lower estimate

$$\bigcup_{j \in \underline{m}} \sigma_{\mathbb{C}}^{(p)}(A_{jj}, \rho) \subseteq \sigma_{\mathbb{C}}^{(p)}(A, \rho). \tag{5.33}$$

For $j \in \underline{m}$ and $\rho \ge 0$ set $\mathcal{K}_j(A, \rho) := \sigma_{\mathbb{C}}^{(p)}(A, \rho) \cap \sigma_{\mathbb{C}}^{(p)}(A_{jj},\, h_m(\rho, \|R_A\|_p)\,)$. Then by (5.32) and (5.33),

$$\sigma_{\mathbb{C}}^{(p)}(A, \rho) = \bigcup_{j \in \underline{m}} \mathcal{K}_j(A, \rho)$$

and

$$\sigma_{\mathbb{C}}^{(p)}(A_{jj}, \rho) \subseteq \mathcal{K}_j(A, \rho) \subseteq \sigma_{\mathbb{C}}^{(p)}(A_{jj},\, h_m(\rho, \|R_A\|_p)\,).$$

We thus have the following corollary to Theorem 5.8.4.

Corollary 5.8.5 *Let $p\in[1,\infty]$ and let $A=[A_{jk}]_{j,k\in\underline{m}}$ be an block upper triangular matrix. Let $\rho>0$ and let $r>h_m(\|R_A\|_p,\rho)$. Let $i\in\underline{m}$ and suppose that*

$$\sigma_{\mathbb{C}}^{(p)}(A_{jj},r)\cap\sigma_{\mathbb{C}}^{(p)}(A_{ii},\,h_m(\rho,\|R_A\|_p)\,)=\emptyset$$

for all $j\in\underline{m}\setminus\{i\}$. Then, with the notation above, $\mathcal{K}_i(A,\rho)\cap\bigcup_{j\in\underline{m}\setminus\{i\}}\mathcal{K}_j(A,\rho)=\emptyset$ and

$$\sigma_{\mathbb{C}}^{(p)}(A_{ii},\rho)\subseteq\mathcal{K}_i(A,\rho)\subseteq\sigma_{\mathbb{C}}^{(p)}\Big(A_{ii},\;\beta_m\Big(\tfrac{\|\langle R_A\rangle_p\|_p}{r}\Big)\,\rho\,\Big).$$

If the quotient $\|\langle R_A\rangle_p\|_p/r$ is small then $\beta_m(\|\langle R_A\rangle_p\|_p/r)$ is close to 1. Hence, Corollary 5.8.5 yields that if $\|\langle R_A\rangle_p\|_p$ is small compared with r then the Hausdorff-distance of $\mathcal{K}_i(A,\rho)$ and $\sigma_{\mathbb{C}}^{(p)}(A_{ii},\rho)$ is small, as well. In particular, the following holds: Let A_z be the matrix which is obtained from A by replacing the diagonal block A_{ii} with $A_{ii}-zI_{n_i}$, $z\in\mathbb{C}$. Then the Hausdorff distance of $\mathcal{K}_i(A_z,\rho)$ and $\sigma_{\mathbb{C}}^{(p)}(A_{ii}-zI_{n_i},\rho)$ tends to 0 as z tends to infinity.

We are now in the position to give our version of Ruhe's result. It is a consequence of Corollary 5.8.5 and the trivial fact that for a 1×1 matrix $[\lambda]\in\mathbb{C}^{1\times1}$, $\sigma_{\mathbb{C}}^{(2)}([\lambda],\rho)=\mathcal{D}(\lambda,\rho)$.

Corollary 5.8.6 *Let $\lambda_1,\ldots,\lambda_n\in\mathbb{C}$ be the eigenvalues of $A\in\mathbb{C}^{n\times n}$. Let $\mathcal{K}_j(\rho)$ denote the connected component of $\sigma_{\mathbb{C}}^{(2)}(A,\rho)$ containing the eigenvalue λ_j, $j\in\underline{n}$. Let $R\in\mathbb{C}^{n\times n}$ be a strictly upper triangular matrix occuring in a Schur form of A such that $\|R\|_2=\delta_2(A)$. Let $i\in\underline{n}$ and $r>h_n(\rho,\delta_2(A))$. Suppose that*

$$|\lambda_i-\lambda_j|>r+h_n(\rho,\delta_2(A))\qquad\text{for all } j\in\underline{n}\setminus\{i\}.$$

Then $\mathcal{K}_i(\rho)\cap\bigcup_{j\in\underline{n}\setminus\{i\}}\mathcal{K}_j(\rho)=\emptyset$ and $\mathcal{D}(\lambda_i,\rho)\subseteq\mathcal{K}_i(A,\rho)\subseteq\mathcal{D}\Big(\lambda_i,\;\beta_n\Big(\tfrac{\|\,|R|\,\|_2}{r}\Big)\;\rho\,\Big)$.

In the original theorem of Ruhe the norm $\|\,|R|\,\|_2$ is replaced by $\delta_2(A)=\|R\|_2$. But there is a gap in the original proof in this respect. Note, however, that $\|\,|R|\,\|_2\le\delta_F(A)$, the departure from normality of A with respect to Frobenius norm. Corollary 5.8.6 confirms what we already have seen in the 2×2 case, namely that not only the departure from normality but also the distance of the eigenvalues of A determines the sensitivity of its spectrum: The departure from normality of A may be arbitrarily large. If the mutual distance of the eigenvalues is large compared with the departure from normality, then the pseudospectrum of A to the level ρ is close to the pseudospectrum of a normal matrix to the same level.

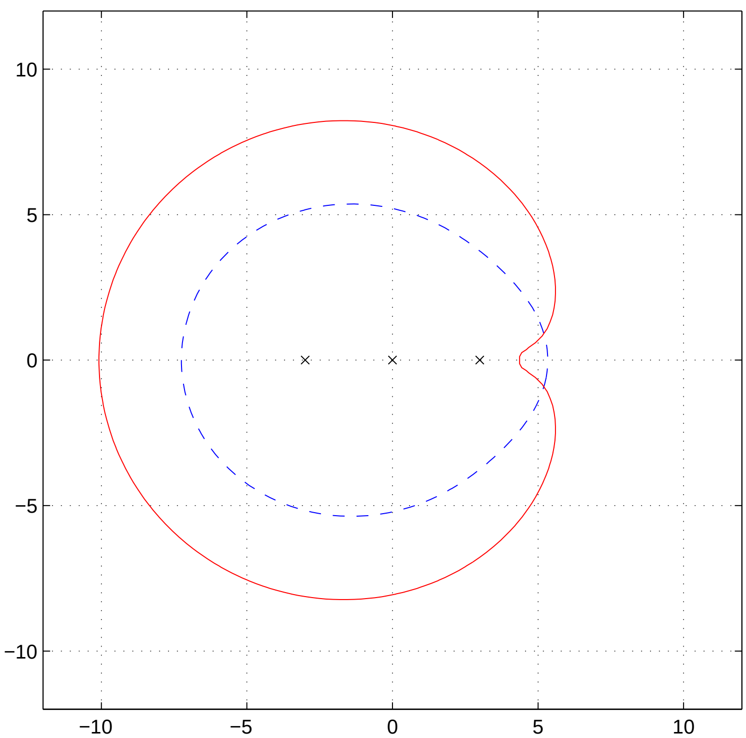


Figure 5.5: The boundaries of the pseudospectra $\sigma_{\mathbb{C}}^{(2)}(A, 0.5)$ (dashed) and $\sigma_{\mathbb{C}}^{(2)}(D_A + 2R_A, 0.5)$ (solid), where A is the matrix defined in (5.19).

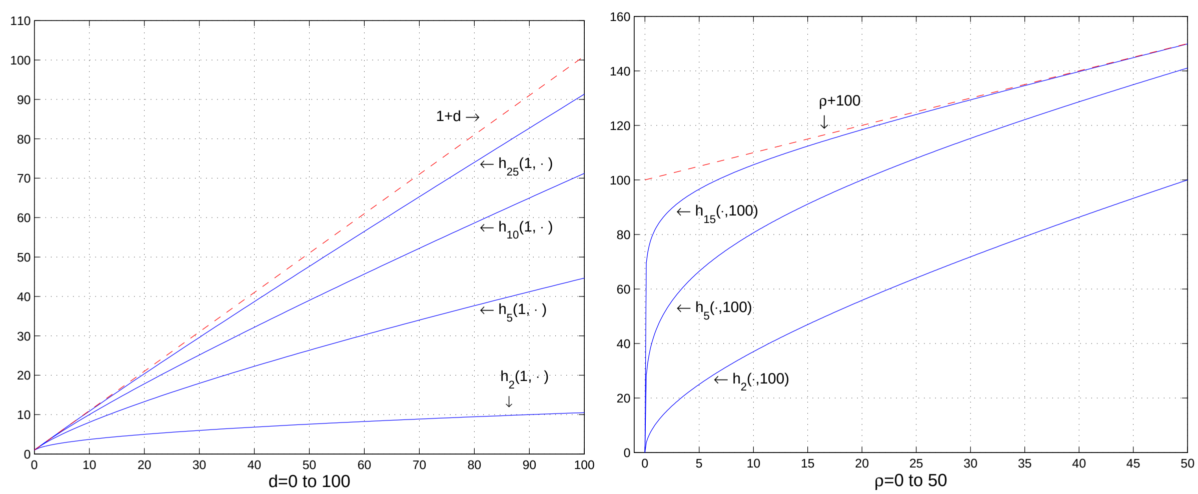


Figure 5.6: The functions $0 \le d \mapsto h_m(1, d)$ and $\rho \mapsto h_m(\rho, 100)$ for several values of m.

Chapter 6

Spectral value sets for real perturbations

In this and the next chapter we study spectral value sets and μ-functions for real matrix perturbations. If the underlying norm is the spectral norm then the corresponding μ-functions are called real perturbation values. This chapter splits into two parts. The first part is basic. Here we display the theory of real perturbation values as it was developed in [7, 8, 9]. Some of the properties of real perturbation values follow from general principles. In order to bring out these principles we introduce higher μ-functions. This gives us the necessary background for our further considerations in the second part of this chapter in which we discuss special properties of real spectral value sets and real μ-functions.

6.1 Real μ-functions and real perturbation values

In this chapter we present the basic theory of real spectral value sets. These are sets of the form.

$$\sigma_{\mathbb{R}}^{\|\cdot\|}(A,B,C,\rho) := \bigcup_{\Delta\in\mathbb{R}^{l\times q},\ \|\Delta\|\le\rho} \sigma(A+B\Delta C),$$

where $(A,B,C)\in L_{n,l,q}$ and $\|\cdot\|$ is a norm on $\mathbb{R}^{l\times q}$. Note that we do not presume that the matrices (A,B,C) are real. This might seem to be unnatural since we consider real perturbations only. However, most of the theory below goes through for arbitrary complex matrix triples (A,B,C). The sets

$$\sigma_{\mathbb{R}}^{\|\cdot\|}(A,\rho) := \sigma_{\mathbb{R}}^{\|\cdot\|}(A,I_n,I_n,\rho), \qquad A\in\mathbb{C}^{n\times n}$$

are called real pseudospectra. By Theorems 2.4.5 and 2.4.9 we have for $\rho > 0$,

$$\begin{aligned}\sigma_{\mathbb{R}}^{\|\cdot\|}(A,B,C,\rho) &= \sigma(A)\cup\{\, s\in\mathbb{C}\backslash\sigma(A) \ \mid\ \mu_{\mathbb{R}}^{\|\cdot\|}(G(s)) \ge \rho^{-1}\,\},\\ \sigma_{\mathbb{R}}^{\|\cdot\|}(A,\rho) &= \{\, s\in\mathbb{C} \ \mid\ \widetilde{\mu}_{\mathbb{R}}^{\|\cdot\|}(sI_n-A)\le\rho\,\},\end{aligned}$$

where $G(s) = C(sI_n-A)^{-1}B$ and

$$\mu_{\mathbb{R}}^{\|\cdot\|}(M) := \Big(\min\{\,\|\Delta\| \ \mid \Delta\in\mathbb{R}^{l\times q},\ 1\in\sigma(\Delta M)\,\}\Big)^{-1}, \qquad M\in\mathbb{C}^{q\times l}, \tag{6.1}$$

$$\widetilde{\mu}_{\mathbb{R}}^{\|\cdot\|}(M) := \min\{\,\|\Delta\| \ \mid \Delta\in\mathbb{R}^{n\times n},\ M-\Delta \text{ is singular }\}, \qquad \text{for } M\in\mathbb{C}^{n\times n}. \tag{6.2}$$

The most important problem in the theory of real spectral value sets is to find computable formulas for the real μ-functions $\mu_{\mathbb{R}}^{\|\cdot\|}(M)$ and $\widetilde{\mu}_{\mathbb{R}}^{\|\cdot\|}(M)$. To the best of our knowledge this problem has been solved only for the following cases.

(a) The matrix M is either a row or a column vector. The underlying norm $\|\cdot\|$ is arbitrary.

(b) The matrix M has arbitrary size. The underlying norm $\|\cdot\|$ is the spectral norm.

The solution for the case (a) has been found by Hinrichsen and Pritchard [48]. We will come back to this in Section 7.7. The long standing problem (b) to find a computable formula for the real μ-function with respect to spectral norm has been solved by Qiu, Bernhardsson, Rantzer, Davison, Young and Doyle in [78]. They have shown the following theorem.

Theorem 6.1.1 *Let $\|\cdot\|$ denote the spectral norm. Then, for every $M\in\mathbb{C}^{q\times l}$,*

$$\mu_{\mathbb{R}}^{\|\cdot\|}(M) = \inf_{\gamma\in(0,1]} \sigma_2\left(\begin{bmatrix} \Re M & -\gamma\,\Im M \\ \gamma^{-1}\,\Im M & \Re M \end{bmatrix}\right), \tag{6.3}$$

where $\sigma_2(\cdot)$ denotes the second largest singular value. Let $M\in\mathbb{C}^{n\times n}$. Then

$$\widetilde{\mu}_{\mathbb{R}}^{\|\cdot\|}(M) = \sup_{\gamma\in(0,1]} \sigma_{2n-1}\left(\begin{bmatrix} \Re M & -\gamma\,\Im M \\ \gamma^{-1}\,\Im M & \Re M \end{bmatrix}\right), \tag{6.4}$$

where $\sigma_{2n-1}(\cdot)$ is the second smallest singular value.

The function to be minimized in (6.3) is always quasiconvex [78]. The function to be maximized in (6.4) is always quasiconcave. Thus the quantities $\mu_{\mathbb{R}}^{\|\cdot\|}(M)$ and $\widetilde{\mu}_{\mathbb{R}}^{\|\cdot\|}(M)$ can be calculated by means of the above formulas using a bisection algorithm. This method has been used to generate some of the pictures in this work. Other pictures of real spectral value sets have been made by applying explicite formulas which can be given for special matrices. We will derive such formulas in Chapter 7. In our opinion the original proof of Theorem 6.1.1 in [78] is dissatisfying because it does not yield insight into the geometry behind the formulas (6.3) and (6.4). However, in [9] Bernhardsson, Qiu and Rantzer provided an alternative and more elegant proof of Theorem 6.1.1. This proof connects the real μ-problem with hermitian-symmetric inequalities which we will be introduced below. The proof in [9] also yields formulas for the computation of the so called real perturbation values. The latter are defined as follows.

Definition 6.1.2 *Let $M\in\mathbb{C}^{q\times l}$. The real perturbation values of M of first kind are*

$$\tau_k(M) \;:=\; \Big(\inf\{\ \|\Delta\|_2 \ \mid\ \Delta\in\mathbb{R}^{l\times q},\ \dim\ker(I_l-\Delta M)\ge k\ \}\Big)^{-1}, \qquad k\in\mathbb{N}. \tag{6.5}$$

The real perturbation values of M of second kind are

$$\widetilde{\tau}_k(M) \;:=\; \inf\{\ \|\Delta\|_2 \ \mid\ \Delta\in\mathbb{R}^{q\times l},\ \operatorname{rank}(M-\Delta)<k\ \}, \qquad k\in\mathbb{N}. \tag{6.6}$$

In the next section we will show that the infimum in (6.5) is always attained, and that the infimum in (6.6) is attained if the set $\{\ \|\Delta\|_2 \ \mid\ \Delta\in\mathbb{R}^{q\times l},\ \operatorname{rank}(M-\Delta)<k\ \}$ is nonempty. If this set is empty then $\widetilde{\tau}(M)=\infty$. Obviously, $k_1<k_2$ implies $\tau_{k_1}(M)\ge\tau_{k_2}(M)$ and $\widetilde{\tau}_{k_1}(M)\ge\widetilde{\tau}_{k_2}(M)$. Moreover, $\tau_k(M)=\widetilde{\tau}_k(M)=0$ if $k>\operatorname{rank}M$. For square matrices $M\in\mathbb{C}^{n\times n}$ we use the notation $\widetilde{\tau}_{\min}(M):=\widetilde{\tau}_n(M)$. Then

$$\widetilde{\tau}_{\min}(M)=\inf\{\ \|\Delta\|_2 \ \mid\ \Delta\in\mathbb{R}^{n\times n},\ \ M-\Delta \text{ is singular }\ \}.$$

With these notations we have for the real μ-functions with respect to spectral norm,

$$\begin{aligned}\tau_1(M) &= \mu_{\mathbb{R}}^{\|\cdot\|}(M) \qquad \text{for } M\in\mathbb{C}^{q\times l},\\ \widetilde{\tau}_{\min}(M) &= \widetilde{\mu}_{\mathbb{R}}^{\|\cdot\|}(M) \qquad \text{for } M\in\mathbb{C}^{n\times n}.\end{aligned}$$

In [9] the following generalization of Theorem 6.1.1 has been shown.

Theorem 6.1.3 *Let $M\in\mathbb{C}^{q\times l}$ and let $k\in\mathbb{N}$. Then*

$$\tau_k(M) = \inf_{\gamma\in(0,1]} \sigma_{2k}\left(\begin{bmatrix}\Re M & -\gamma\,\Im M\\ \gamma^{-1}\,\Im M & \Re M\end{bmatrix}\right), \tag{6.7}$$

$$\widetilde{\tau}_k(M) = \sup_{\gamma\in(0,1]} \sigma_{2k-1}\left(\begin{bmatrix}\Re M & -\gamma\,\Im M\\ \gamma^{-1}\,\Im M & \Re M\end{bmatrix}\right). \tag{6.8}$$

We emphazise that we will not apply Theorem 6.1.3 or its special case, Theorem 6.1.1, in our further considerations concerning real spectral value sets. However, for the sake of completeness we will give the full proof of Theorem 6.1.3, partly at the end of this section and partly in the appendix.

We now give a brief outline of the chapter. In the remainder of the present section we discuss analogies between the real perturbation values and the singular values of a matrix. In doing so we provide a characterization of real perturbation values by hermitian-symmetric inequalities. This characterization yields the basis for the proof of Theorem 6.1.3. Moreover, it will be our main tool in deriving explicite formulas for real spectral value sets of special matrices in the next chapter. In the present section some important statements remain unproved. The missing proofs will be given in Section 6.5 after some preliminary consideration in Sections 6.2-6.4. In these sections we study the real μ-problem from a general point of view. Here we introduce higher μ-functions and discuss them for the case that the underlying norm is unitarily invariant. In Sections 6.6-6.9 we deal with the standard μ again. We give an easily computable lower bound for real μ and derive criteria under which the real and the complex μ coincide. Finally, in Section 6.9 we consider continuity properties of real μ-functions.

We now continue with displaying the theory of real perturbation values. Below we will show that the singular values of $M\in\mathbb{C}^{q\times l}$ satisfy [1]

$$\begin{aligned}\sigma_k(M) &= \Big(\inf\{\ \|\Delta\|_2\ \mid\ \Delta\in\mathbb{C}^{\,l\times q},\ \ \dim\ker(I_l-\Delta M)\geq k\ \}\Big)^{-1} \qquad (6.9)\\ &= \inf\{\ \|\Delta\|_2\ \mid\ \Delta\in\mathbb{C}^{\,q\times l},\ \ \operatorname{rank}(M-\Delta)<k\ \}, \qquad (6.10)\end{aligned}$$

Thus if we replace in the definition of the real perturbation values the sets $\mathbb{R}^{\,l\times q}$ and $\mathbb{R}^{\,q\times l}$ by $\mathbb{C}^{\,l\times q}$ and $\mathbb{C}^{\,q\times l}$ we obtain the real perturbation values. Furthermore, the relations (6.9) and (6.10) imply that

$$\tau_k(M)\leq\sigma_k(M)\leq\widetilde{\tau}_k(M). \tag{6.11}$$

If M is real then the infima in (6.9) and (6.10) are attained by a real matrices Δ. Thus

$$\tau_k(M)=\sigma_k(M)=\widetilde{\tau}_k(M) \quad \text{ if } M\in\mathbb{R}^{q\times l}. \tag{6.12}$$

[1]The relation (6.10) is a special case of the Schmidt-Mirsky Theorem, also called the Eckard-Young Theorem, see [86, Theorem 4.18].

In the sequel we denote by $\mathfrak{G}_{n,k}$ the Grassmann-manifold of k-dimensional complex subspaces of $\mathbb{C}^n$. It is well known [62] that the following max-min-formulas hold for the singular values of $M\in\mathbb{C}^{q\times l}$.

$$\sigma_k(M) \quad = \quad \max_{\mathcal{X}\in\mathfrak{G}_{l,k}} \min_{0\neq x\in\mathcal{X}} \frac{\|Mx\|_2}{\|x\|_2} = \min_{\mathcal{X}\in\mathfrak{G}_{l,l-k+1}} \max_{0\neq x\in\mathcal{X}} \frac{\|Mx\|_2}{\|x\|_2}, \qquad k\in\underline{l}.$$

For the real perturbation values we have the following max-min resp. min-max characterization which will be shown in Section 6.5.

Theorem 6.1.4 *[9] Let $M\in\mathbb{C}^{q\times l}$ and $k\in\underline{l}$. Then*

$$\tau_k(M) \quad = \quad \max_{\mathcal{X}\in\mathfrak{G}_{l,k}} \min_{x\in\mathcal{X}} \frac{\|\Re(Mx)\|_2}{\|\Re x\|_2}, \tag{6.13}$$

$$\widetilde{\tau}_k(M) \quad = \quad \min_{\mathcal{X}\in\mathfrak{G}_{l,l-k+1}} \max_{x\in\mathcal{X}} \frac{\|\Re(Mx)\|_2}{\|\Re x\|_2}, \tag{6.14}$$

where by convention,

$$\frac{\|\Re(Mx)\|_2}{\|\Re x\|_2} := \begin{cases} 0 & \text{if } \|\Re x\|_2 = \|\Re(Mx)\|_2 = 0, \\ \infty & \text{if } \|\Re x\|_2 = 0,\ \|\Re(Mx)\|_2 > 0. \end{cases}$$

In order to state the next analogy between singular values and real perturbation values we need more notation.

Notation 6.1.5 *Let $\mathcal{H}_n := \{\ H\in\mathbb{C}^{n\times n}\ |\ H^* = H\ \}$ be the set of hermitian $n\times n$ matrices. If $H\in\mathcal{H}_n$ then we denote by $\lambda_1(H)\geq\lambda_2(H)\geq\ldots\geq\lambda_n(H)$ its eigenvalues in nonincreasing order. By $i_{\geq}(H)$ we denote the positive semi-definite inertia of H, i.e.*

$$i_{\geq}(H) := \max\{\ k\in\mathbb{N}_0 \mid \exists\mathcal{X}\in\mathfrak{G}_{n,k} : \forall x\in\mathcal{X} : x^*Hx\geq 0\ \}.$$

It is a basic result of linear algebra that for any $H\in\mathcal{H}_n$,

$$i_{\geq}(H) \quad = \quad \begin{cases} 0 & \text{if } \lambda_1(H) < 0, \\ \max\{\ k\in\underline{n} \mid \lambda_k(H)\geq 0\ \} & \text{otherwise.} \end{cases}$$

Using this it is easily seen that the eigenvalues of $H\in\mathcal{H}_n$ can be characterized in the following way.

$$\begin{aligned} \lambda_k(H) &= \max\{\ t\in\mathbb{R} \mid i_{\geq}(H - tI_n)\geq k\ \} \\ &= \min\{\ t\in\mathbb{R} \mid i_{\geq}(tI_n - H)\geq n-k+1\ \}. \end{aligned}$$

Since the singular values of $M\in\mathbb{C}^{q\times l}$ are the square roots of the eigenvalues of the positive semi-definite matrix M^*M, it follows that

$$\sigma_k(M) \quad = \quad \max\{\ t\geq 0 \mid i_{\geq}(M^*M - t^2I_l)\geq k\ \} \tag{6.15}$$

$$\quad = \quad \min\{\ t\geq 0 \mid i_{\geq}(t^2I_l - M^*M)\geq l-k+1\ \}. \tag{6.16}$$

An analogous statement for real perturbation values uses hermitian-symmetric inequalities. These are inequalities of the form

$$x^*Hx \bullet |x^TSx|, \qquad \bullet\in\{>,<,\geq,\leq\}, \quad x\in\mathbb{C}^n,$$

where $H\in\mathbb{C}^{n\times n}$ is a hermitian matrix and $S\in\mathbb{C}^{n\times n}$ is a complex symmetric matrix.

Notation 6.1.6 *By*

$$\mathcal{HS}_n := \{\ (H,S)\in\mathbb{C}^{n\times n}\times\mathbb{C}^{n\times n}\ \mid\ H^*=H,\ S^T=S\ \}$$

we denote the set of $n\times n$ *hermitian-symmetric matrix pairs. For* $(H,S)\in\mathcal{HS}_n$ *we set*

$$i_{\geq}(H,S) := \max\{\ k\in\mathbb{N}_0 \mid \exists\mathcal{X}\in\mathfrak{G}_{n,k} : \forall x\in\mathcal{X} : x^*Hx \geq |x^TSx|\ \}$$

Note that $i_{\geq}(H,0)=i_{\geq}(H)$. Below we will see that the quantity $i_{\geq}(H,S)$ is closely related to the inertia of a hermitian pencil. We now consider the sets

$$\begin{aligned} T_k(M) &:= \{\ t\geq 0 \mid\ i_{\geq}(M^*M-t^2I_l\,,\ M^TM-t^2I_l)\geq k\ \}, \\ \widetilde{T}_k(M) &:= \{\ t\geq 0 \mid\ i_{\geq}(t^2I_l-M^*M\,,\ t^2I_l-M^TM)\geq l-k+1\ \}, \qquad M\in\mathbb{C}^{q\times l}. \end{aligned}$$

In Section 6.5 we will prove the following analog to the formulas (6.15) and (6.16). It relates the real perturbation values to hermitian symmetric inequalities.

Theorem 6.1.7 *Let* $M\in\mathbb{C}^{q\times l}$ *and* $k\in\underline{l}$*. Then* $\tau_k(M)=\max T_k(M)$ *and* $T_k(M)=[0,\tau_k(M)]$*. If* $\widetilde{T}_k(M)=\emptyset$ *then* $\widetilde{\tau}_k(M)=\infty$*. Otherwise we have* $\widetilde{\tau}_k(M)=\min\widetilde{T}_k(M)$ *and* $\widetilde{T}_k(M)=[\widetilde{\tau}_k(M),\infty)$*.*

By specializing Theorem 6.1.7 to $\tau_1(\cdot)$ and $\widetilde{\tau}_{\min}(\cdot)$ we obtain

Corollary 6.1.8 *For any* $M\in\mathbb{C}^{q\times l}$*,*

$$\tau_1(M) = \max\{\ t\geq 0 \mid \exists\, x\in\mathbb{C}^l\setminus\{0\} : x^*(M^*M-t^2I_l)x \geq |x^T(M^TM-t^2I_l)x|\ \}.$$

If $M\in\mathbb{C}^{n\times n}$ *is a square matrix then*

$$\widetilde{\tau}_{\min}(M) = \inf\{\ t\geq 0 \mid \exists\, x\in\mathbb{C}^l\setminus\{0\} : x^*(t^2I_l-M^*M)x \geq |x^T(t^2I_l-M^TM)x|\ \}. \qquad (6.17)$$

If the set $\{\ t\geq 0 \mid \exists\, x\in\mathbb{C}^l\setminus\{0\} : x^*(t^2I_l-M^*M)x \geq |x^T(t^2I_l-M^TM)x|\ \}$ *is nonempty then the infimum in (6.17) is attained. Otherwise* $\widetilde{\tau}_{\min}(M)=\infty$*.*

The proposition below is a supplement to the Theorems 6.1.4 and 6.1.7. It in particular states how to construct a real matrix Δ of minimal spectral norm satisfying $\dim\ker(I_l-\Delta M)\geq k$ resp. $\operatorname{rank}(M-\Delta)<k$ from the solution of a hermitan symmetric inequality. By $A^\dagger$ we denote the Moore-Penrose generalized inverse of A.

Proposition 6.1.9 *Let* $M\in\mathbb{C}^{q\times l}$ *and* $k\in\underline{l}$*.*

(*i*) *For any* k*-dimensional subspace* $\mathcal{X}_0$ *of* $\mathbb{C}^l$ *the following conditions are equivalent.*

(*a*) *The hermitian-symmetric inequality*

$$x^*(M^*M-\tau_k(M)^2\,I_l)x \geq |\,x^T(M^TM-\tau_k(M)^2\,I_l)x\,|$$

holds for all $x\in\mathcal{X}_0$*.*

(*b*) *We have*

$$\min_{x\in\mathcal{X}_0}\frac{\|\Re(Mx)\|_2}{\|\Re x\|_2} = \tau_k(M).$$

If $\tau_k(M) = 0$ then there is no $\Delta \in \mathbb{R}^{l\times q}$ such that $\dim\ker(I_l - \Delta M) \geq k$. *Suppose that $\tau_k(M) > 0$. Let $\mathcal{X}_0$ be any k-dimensional subspace of $\mathbb{C}^l$ satisfying the conditions (a) and (b) above, and let X_0 be any matrix whose columns form a basis of $\mathcal{X}_0$. Set*

$$\Delta_0 := [\,\Re X_0 \;\; \Im X_0\,]\,[\,\Re(MX_0) \;\; \Im(MX_0)\,]^{\dagger} \in \mathbb{R}^{l\times q}.$$

Then

$$\dim\ker(I_l - \Delta_0 M) \geq k \quad \textit{and} \quad \|\Delta_0\|_2 = \tau_k(M)^{-1}.$$

(ii) *If $\widetilde{\tau}_k(M) = \infty$ then there is no $\Delta \in \mathbb{R}^{q\times l}$ such that* $\operatorname{rank}(M - \Delta) < k$. *Suppose that $\widetilde{\tau}_k(M) < \infty$. Then for any $(l-k+1)$-dimensional subspace $\widetilde{\mathcal{X}}_0$ of $\mathbb{C}^l$ the following conditions are equivalent.*

($\widetilde{a}$) *The hermitian-symmetric inequality*

$$x^*(\widetilde{\tau}_k(M)^2\, I_l - M^*M)x \geq |\, x^T(\widetilde{\tau}_k(M)^2\, I_l - M^TM)x\,|$$

holds for all $x \in \widetilde{\mathcal{X}}_0$.

($\widetilde{b}$) *We have*

$$\max_{x\in\widetilde{\mathcal{X}}_0} \frac{\|\Re(Mx)\|_2}{\|\Re x\|_2} = \widetilde{\tau}_k(M).$$

Let $\widetilde{\mathcal{X}}_0$ be any $(l-k+1)$-dimensional subspace of $\mathbb{C}^l$ satisfying the conditions ($\widetilde{a}$) and ($\widetilde{b}$), and let $\widetilde{X}_0 \in \mathbb{C}^{l\times(l-k+1)}$ be any matrix whose columns form a basis of $\widetilde{\mathcal{X}}_0$. Set

$$\widetilde{\Delta}_0 := [\,\Re(M\widetilde{X}_0) \;\; \Im(M\widetilde{X}_0)\,]\,[\,\Re\widetilde{X}_0 \;\; \Im\widetilde{X}_0\,]^{\dagger} \in \mathbb{R}^{l\times q}.$$

Then

$$\operatorname{rank}(M - \widetilde{\Delta}_0) < k \quad \textit{and} \quad \|\widetilde{\Delta}_0\|_2 = \widetilde{\tau}_k(M).$$

We will show Proposition 6.1.9 in Section 6.5. For comparison we give the analogous statement for the singular values. It is based on a singular value decomposition

$$M = \textstyle\sum_{j\in\underline{l}} \sigma_j(M)\, u_j v_j^* \tag{6.18}$$

of $M \in \mathbb{C}^{q\times l}$, where $v_1, \ldots, v_l$ form an orthogonal basis of $\mathbb{C}^l$, $u_1, \ldots, u_q$ are pairwise orthogonal unit vectors in $\mathbb{C}^q$ and $u_j = 0$ if $l > q$ and $j > q$.

Proposition 6.1.10 *Let $M \in \mathbb{C}^{q\times l}$ have a singular value decomposition as in (6.18). Then the following statements hold for any $k \in \underline{l}$.*

(i) *The k-dimensional subspace $\mathcal{X}_0 :=$ span$\{v_1, \ldots, v_k\}$ satisfies*

$$\min_{x\in\mathcal{X}_0\setminus\{0\}} \frac{\|Mx\|_2}{\|x\|_2} = \sigma_k(M).$$

Moreover,

$$x^*(M^*M - \sigma_k(M)^2 I_l)x \geq 0 \qquad \textit{for all } x \in \mathcal{X}_0.$$

If $\sigma_k(M) = 0$ then there is no matrix $\Delta \in \mathbb{C}^{l\times q}$ such that $\dim\ker(I_l - \Delta M) \geq k$. *Suppose $\sigma_k(M) > 0$. Let*

$$\Delta_0 := \sum_{j=1}^{k} \sigma_j(M)^{-1} v_j u_j^*$$

Then $\dim\ker(I_l - \Delta_0 M) \geq k$ *and* $\|\Delta_0\|_2 = \sigma_k(M)^{-1}$.

(ii) The $(l-k+1)$*-dimensional subspace* $\widetilde{\mathcal{X}}_0 = \operatorname{span}\{v_k, \dots, v_l\}$ *satisfies*

$$\max_{x \in \widetilde{\mathcal{X}}_0 \setminus \{0\}} \frac{\|Mx\|_2}{\|x\|_2} = \sigma_k(M)$$

Moreover,

$$x^*(\sigma_k(M)\, I_l - M^*M)x \geq 0 \qquad \text{for all } x \in \widetilde{\mathcal{X}}_0.$$

Let

$$\widetilde{\Delta}_0 := \sum_{j=l}^{l-k+1} \sigma_j(M) u_j v_j^*.$$

Then $\operatorname{rank}(M - \widetilde{\Delta}_0) < k$ *and* $\|\widetilde{\Delta}_0\|_2 = \sigma_k(M)$.

All statements of Proposition 6.1.10 are easily verified. Observe that Proposition 6.1.10 implies that

$$\sigma_k(M) \;\leq\; \Big(\inf\{\; \|\Delta\|_2 \;\mid\; \Delta \in \mathbb{C}^{l \times q},\;\; \dim\ker(I_l - \Delta M) \geq k \;\;\} \Big)^{-1}, \tag{6.19}$$

$$\sigma_k(M) \;\geq\; \inf\{\; \|\Delta\|_2 \;\mid\; \Delta \in \mathbb{C}^{q \times l},\;\; \operatorname{rank}(M - \Delta) < k \;\;\}. \tag{6.20}$$

We claimed in the begining of our considerations that equality holds in (6.19) and (6.20). In order to see this suppose first that $\Delta \in \mathbb{C}^{l \times q}$ is such that $\dim\ker(I_l - \Delta M) \geq k$. Then there is a k-dimensional subspace $\mathcal{X}_0$ such that $\Delta M x = x$ for all $x \in \mathcal{X}_0$. This implies $\|\Delta\|_2 \|Mx\|_2 \geq \|x\|_2$ for $x \in \mathcal{X}_0$ and therefore

$$\|\Delta\|_2^{-1} \leq \min_{0 \neq x \in \mathcal{X}_0} \frac{\|Mx\|_2}{\|x\|_2} \leq \max_{\mathcal{X} \in \mathfrak{G}_{l,k}} \min_{0 \neq x \in \mathcal{X}} \frac{\|Mx\|_2}{\|x\|_2} = \sigma_k(M).$$

Hence, $\sigma_k(M) \geq \big(\inf\{\; \|\Delta\|_2 \;\mid\; \Delta \in \mathbb{C}^{l \times q},\;\; \dim\ker(I_l - \Delta M) \geq k \;\;\} \big)^{-1}$. Now suppose that $\Delta \in \mathbb{C}^{q \times l}$ satisfies $\operatorname{rank}(M - \Delta) < k$. Then $\dim\ker(M - \Delta) \geq l - k + 1$. Thus $\Delta x = Mx$ for all x in an $(l-k+1)$-dimensional subspace $\mathcal{X}_0$. Hence

$$\|\Delta\|_2 \geq \max_{0 \neq x \in \mathcal{X}_0} \frac{\|Mx\|_2}{\|x\|_2} \geq \min_{\mathcal{X} \in \mathfrak{G}_{l,l-k+1}} \max_{0 \neq x \in \mathcal{X}} \frac{\|Mx\|_2}{\|x\|_2} = \sigma_k(M).$$

Therefore, $\sigma_k(M) \leq \inf\{\; \|\Delta\|_2 \;\mid\; \Delta \in \mathbb{C}^{q \times l},\;\; \operatorname{rank}(M - \Delta) < k \;\;\}$. We thus have shown the relations (6.9) and (6.10). These imply the inequalities (6.11). If M is real then the singular vectors in (6.18) can be taken to be real. Hence, the matrices Δ_0 and $\widetilde{\Delta}_0$ defined in Proposition 6.1.10 are real. Thus, the identities in (6.12) hold.

We are now going to show how the formulas (6.7) and (6.8) for the computation of the real perturbation values are obtained using hermitian-symmetric inequalities. To this end we need the result below which relates the quantity $i_{\geq}(H, S)$ to the inertia of the hermitian pencil $\begin{bmatrix} H & 0 \\ 0 & \overline{H} \end{bmatrix} + \alpha \begin{bmatrix} 0 & \overline{S} \\ S & 0 \end{bmatrix}$, $\alpha \in \mathbb{R}$.

Theorem 6.1.11 *(Bernhardsson, Qiu, Rantzer [8, 9]) Let* $(H, S) \in \mathcal{HS}_n$ *and* $k \in \underline{n}$. *The following assertions are equivalent.*

(a) $\quad i_{\geq}(H, S) \geq k$.

(b) $\quad i_{\geq}\left(\begin{bmatrix} H & \alpha \overline{S} \\ \alpha S & \overline{H} \end{bmatrix} \right) \geq 2k \qquad \text{for all} \quad \alpha \in (-1, 0).$

$$(c) \qquad \lambda_{2k}\left(\begin{bmatrix} H & \alpha\,\overline{S} \\ \alpha\,S & \overline{H}\end{bmatrix}\right) \geq 0 \qquad \textit{for all} \quad \alpha\in(-1,0).$$

We will refer to this important and deep result as the BQR-Theorem. The interested reader can find its proof and some additional remarks in the appendix. Note that the equivalence $(b) \Leftrightarrow (c)$ is trivial. The implication $(a) \Rightarrow (b)$ can be easily obtained from Lemma 6.4.2 in the next section. The implication $(b) \Rightarrow (a)$ is the hard part of the Theorem. We close this section by showing how the formulas (6.7) and (6.8) can be derived from Theorem 6.1.7 and the BQR-Theorem by means of some simple congruence transformations. Set

$$D_{n,\alpha} \;:=\; \frac{1}{\sqrt{2}}\begin{bmatrix} \frac{1}{\sqrt{1+\alpha}}\,I_n & \frac{i}{\sqrt{1-\alpha}}\,I_n \\ \frac{1}{\sqrt{1+\alpha}}\,I_n & -\frac{i}{\sqrt{1-\alpha}}\,I_n \end{bmatrix} \qquad n\in\mathbb{N},\;\; \alpha\in(-1,1).$$

These matrices satisfy

$$\begin{aligned} D_{n,\alpha}^{-1} &= \frac{1}{\sqrt{2}}\begin{bmatrix} \sqrt{1+\alpha}\,I_n & \sqrt{1+\alpha}\,I_n \\ -i\sqrt{1-\alpha}\,I_n & i\sqrt{1-\alpha}\,I_n \end{bmatrix}, \\ D_{n,\alpha}^{-*}D_{n,\alpha}^{-1} &= \begin{bmatrix} I_n & \alpha\,I_n \\ \alpha\,I_n & I_n \end{bmatrix} \qquad \text{(where } D_{n,\alpha}^{-*} := \left(D_{n,\alpha}^{-1}\right)^{*}\text{)} \end{aligned}$$

and, for every $A\in\mathbb{C}^{q\times n}$,

$$D_{q,\alpha}^{-1}\begin{bmatrix} A & 0 \\ 0 & \overline{A}\end{bmatrix} D_{n,\alpha} = \begin{bmatrix} \Re A & -\sqrt{\frac{1+\alpha}{1-\alpha}}\,\Im A \\ \sqrt{\frac{1-\alpha}{1+\alpha}}\,\Im A & \Re A \end{bmatrix}.$$

Let $M\in\mathbb{C}^{q\times l}$, $k\in\underline{n}$. Then the following equivalences hold for every $t\geq 0$.

$$
\begin{aligned}
\tau_k(M) \geq t \;\; & \Leftrightarrow \;\; i_{\geq}(M^*M - t^2 I_l \,,\, M^T M - t^2 I_l) \geq k \\
& \qquad \text{(by Theorem 6.1.7)} \\
& \Leftrightarrow \;\; \lambda_{2k}\left(\begin{bmatrix} M^*M - t^2 I_l & \alpha\,\overline{(M^T M - t^2 I_l)} \\ \alpha\,(M^T M - t^2 I_l) & \overline{M^*M - t^2 I_l} \end{bmatrix}\right) \geq 0 \qquad \forall \alpha \in (-1,0) \\
& \qquad \text{(by the BQR-Theorem)} \\
& \Leftrightarrow \;\; \lambda_{2k}\left(\begin{bmatrix} M & 0 \\ 0 & \overline{M} \end{bmatrix}^* \begin{bmatrix} I_q & \alpha\, I_q \\ \alpha\, I_q & I_q \end{bmatrix} \begin{bmatrix} M & 0 \\ 0 & \overline{M} \end{bmatrix} - t^2 \begin{bmatrix} I_l & \alpha\, I_l \\ \alpha\, I_l & I_l \end{bmatrix}\right) \geq 0 \qquad \forall \alpha \in (-1,0) \\
& \Leftrightarrow \;\; \lambda_{2k}\left(\begin{bmatrix} M & 0 \\ 0 & \overline{M} \end{bmatrix}^* D_{q,\alpha}^{-*}\, D_{q,\alpha}^{-1} \begin{bmatrix} M & 0 \\ 0 & \overline{M} \end{bmatrix} - t^2\, D_{l,\alpha}^{-*}\, D_{l,\alpha}^{-1}\right) \geq 0 \qquad \forall \alpha \in (-1,0) \\
& \Leftrightarrow \;\; \lambda_{2k}\left(D_{l,\alpha}^*\left(\begin{bmatrix} M & 0 \\ 0 & \overline{M} \end{bmatrix}^* D_{q,\alpha}^{-*}\, D_{q,\alpha}^{-1} \begin{bmatrix} M & 0 \\ 0 & \overline{M} \end{bmatrix} - t^2 D_{l,\alpha}^{-*}\, D_{l,\alpha}^{-1}\right) D_{l,\alpha}\right) \geq 0 \quad \forall \alpha \in (-1,0) \\
& \qquad \text{(since congruence transformations do not change the signs} \\
& \qquad \text{ of the eigenvalues of a symmetric matrix)} \\
& \Leftrightarrow \;\; \lambda_{2k}\left(\left(D_{q,\alpha}^{-1} \begin{bmatrix} M & 0 \\ 0 & \overline{M} \end{bmatrix} D_{l,\alpha}\right)^* \left(D_{q,\alpha}^{-1} \begin{bmatrix} M & 0 \\ 0 & \overline{M} \end{bmatrix} D_{l,\alpha}\right) - t^2\, I_{2l}\right) \geq 0 \quad \forall \alpha \in (-1,0) \\
& \Leftrightarrow \;\; \lambda_{2k}\left(\left(D_{q,\alpha}^{-1} \begin{bmatrix} M & 0 \\ 0 & \overline{M} \end{bmatrix} D_{l,\alpha}\right)^* \left(D_{q,\alpha}^{-1} \begin{bmatrix} M & 0 \\ 0 & \overline{M} \end{bmatrix} D_{l,\alpha}\right)\right) \geq t^2 \qquad \forall \alpha \in (-1,0) \\
& \Leftrightarrow \;\; \sqrt{\lambda_{2k}\left(\left(D_{q,\alpha}^{-1} \begin{bmatrix} M & 0 \\ 0 & \overline{M} \end{bmatrix} D_{l,\alpha}\right)^* \left(D_{q,\alpha}^{-1} \begin{bmatrix} M & 0 \\ 0 & \overline{M} \end{bmatrix} D_{l,\alpha}\right)\right)} \geq t \qquad \forall \alpha \in (-1,0) \\
& \Leftrightarrow \;\; \sigma_{2\,k}\left(\begin{bmatrix} \Re M & -\sqrt{\frac{1+\alpha}{1-\alpha}}\, \Im M \\ \sqrt{\frac{1-\alpha}{1+\alpha}}\, \Im M & \Re M \end{bmatrix}\right) \geq t \qquad \forall \alpha \in (-1,0) \\
& \Leftrightarrow \;\; \sigma_{2\,k}\left(\begin{bmatrix} \Re M & -\gamma\, \Im M \\ \gamma^{-1}\, \Im M & \Re M \end{bmatrix}\right) \geq t \qquad \forall \gamma \in (0,1) \\
& \Leftrightarrow \;\; \inf_{\gamma \in (0,1)} \sigma_{2\,k}\left(\begin{bmatrix} \Re M & -\gamma\, \Im M \\ \gamma^{-1}\, \Im M & \Re M \end{bmatrix}\right) \geq t
\end{aligned}
$$

Thus

$$
\tau_k(M) \;=\; \inf_{\gamma \in (0,1)} \sigma_{2\,k}\left(\begin{bmatrix} \Re M & -\gamma\, \Im M \\ \gamma^{-1}\, \Im M & \Re M \end{bmatrix}\right) = \inf_{\gamma \in (0,1]} \sigma_{2\,k}\left(\begin{bmatrix} \Re M & -\gamma\, \Im M \\ \gamma^{-1}\, \Im M & \Re M \end{bmatrix}\right).
$$

The right equation holds since the function to be minimized is continuous.

Analogously we have for every $t \geq 0$,

$$\begin{aligned}
\widetilde{\tau}_k(M) \le t \;&\Leftrightarrow\; i_{\geq}(t^2 I_l - M^*M\,,\, t^2 I_l - M^T M) \geq l-k+1 \\
&\qquad \text{(by Theorem 6.1.7)} \\
&\Leftrightarrow\; \lambda_{2(l-k+1)} \begin{bmatrix} t^2 I_l - M^*M & \alpha\,\overline{(t^2 I_l - M^T M)} \\ \alpha\,(t^2 I_l - M^T M) & \overline{t^2 I_l - M^*M} \end{bmatrix} \geq 0 \qquad \forall \alpha \in (-1,0) \\
&\qquad \text{(by the BQR-Theorem)} \\
&\Leftrightarrow\; -\lambda_{2k-1} \begin{bmatrix} M^*M - t^2 I_l & \alpha\,\overline{(M^T M - t^2 I_l)} \\ \alpha\,(M^T M - t^2 I_l) & \overline{M^*M - t^2 I_l} \end{bmatrix} \geq 0 \qquad \forall \alpha \in (-1,0) \\
&\qquad \text{(since } \lambda_j(H) = -\lambda_{n+1-j}(-H) \text{ for } H = H^* \in \mathbb{C}^{n\times n}) \\
&\Leftrightarrow\; \lambda_{2k-1} \begin{bmatrix} M^*M - t^2 I_l & \alpha\,\overline{(M^T M - t^2 I_l)} \\ \alpha\,(M^T M - t^2 I_l) & \overline{M^*M - t^2 I_l} \end{bmatrix} \leq 0 \qquad \forall \alpha \in (-1,0) \\
&\Leftrightarrow\; \lambda_{2k-1} \left(\left(D_{q,\alpha}^{-1} \begin{bmatrix} M & 0 \\ 0 & \overline{M} \end{bmatrix} D_{l,\alpha} \right)^* \left(D_{q,\alpha}^{-1} \begin{bmatrix} M & 0 \\ 0 & \overline{M} \end{bmatrix} D_{l,\alpha} \right) - t^2\, I_{2l} \right) \leq 0 \quad \forall \alpha \in (-1,0) \\
&\Leftrightarrow\; \sigma_{2\,k-1} \begin{bmatrix} \Re M & -\sqrt{\frac{1+\alpha}{1-\alpha}}\,\Im M \\ \sqrt{\frac{1-\alpha}{1+\alpha}}\,\Im M & \Re M \end{bmatrix} \leq t \qquad \forall \alpha \in (-1,0) \\
&\Leftrightarrow\; \sigma_{2\,k-1} \begin{bmatrix} \Re M & -\gamma\,\Im M \\ \gamma^{-1}\,\Im M & \Re M \end{bmatrix} \leq t \qquad \forall \gamma \in (0,1) \\
&\Leftrightarrow\; \sup\nolimits_{\gamma\in(0,1)} \sigma_{2\,k-1} \left(\begin{bmatrix} \Re M & -\gamma\,\Im M \\ \gamma^{-1}\,\Im M & \Re M \end{bmatrix} \right) \leq t
\end{aligned}$$

Thus

$$\widetilde{\tau}_k(M) = \sup_{\gamma\in(0,1)} \sigma_{2\,k-1} \left(\begin{bmatrix} \Re M & -\gamma\,\Im M \\ \gamma^{-1}\,\Im M & \Re M \end{bmatrix} \right) = \sup_{\gamma\in(0,1]} \sigma_{2\,k-1} \left(\begin{bmatrix} \Re M & -\gamma\,\Im M \\ \gamma^{-1}\,\Im M & \Re M \end{bmatrix} \right).$$

6.2 Generalized μ-functions

In this section we define higher μ-functions with respect to arbitrary perturbation structures. Moreover, we introduce the concept of ν-functions and show how μ can be obtained by minimizing ν over a certain set of matrix pairs. We think that real μ-functions and in particular real perturbation values are best understood in this framework. ν-functions with respect to unitarily invariant norms will be discussed in some detail in the next sections. At the end of our considerations we will obtain the proofs of Theorem 6.1.4, Theorem 6.1.7 and Proposition 6.1.9. As a byproduct we get additional information about the real μ which is needed later on.

Definition 6.2.1 *Let $M\in\mathbb{C}^{q\times l}$ and $k\in\underline{n}$. We define generalized μ-functions by*

$$\begin{aligned}
\mu^{\|\cdot\|}_{\mathbf{\Delta},k}(M) \;&:=\; \left(\inf\{\, \|\Delta\| \mid \Delta\in\mathbf{\Delta},\ \dim\ker(I_l - \Delta M) \geq k \,\} \right)^{-1}, && (6.21)\\
\widetilde{\mu}^{\|\cdot\|}_{\mathbf{\Delta},k}(M) \;&:=\; \inf\{\, \|\Delta\| \mid \Delta\in\mathbf{\Delta},\ \operatorname{rank}(M-\Delta) < k \,\}, && (6.22)
\end{aligned}$$

where $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}$ *resp.* $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{q,l}$.

The functions $\mu^{\|\cdot\|}_{\mathbf{\Delta},k}(\cdot)$ and $\widetilde{\mu}^{\|\cdot\|}_{\mathbf{\Delta},k}(\cdot)$ are natural generalizations of both the μ-functions defined in Section 2.4, and the real perturbation values. We have

$$\tau_k(M) = \mu^{\|\cdot\|_2}_{\mathbb{R}^{l\times q},k}(M), \qquad \widetilde{\tau}_k(M) = \widetilde{\mu}^{\|\cdot\|_2}_{\mathbb{R}^{q\times l},k}(M), \qquad \mu^{\|\cdot\|}_{\mathbf{\Delta}}(M) = \mu^{\|\cdot\|}_{\mathbf{\Delta},1}(M),$$

and if $l = q =: n$ and $M \in \mathbb{C}^{n\times n}$,

$$\widetilde{\mu}^{\|\cdot\|}_{\mathbf{\Delta}}(M) = \widetilde{\mu}^{\|\cdot\|}_{\mathbf{\Delta},n}(M).$$

By (6.9) and (6.10) we also have for all $M \in \mathbb{C}^{q\times l}$, $k \in \mathbb{N}$,

$$\sigma_k(M) = \mu^{\|\cdot\|_2}_{\mathbb{C}^{l\times q},k}(M) = \widetilde{\mu}^{\|\cdot\|_2}_{\mathbb{R}^{q\times l},k}(M).$$

Let us discuss some basic properties of the generalized μ-functions.

(*a*) We have $\widetilde{\mu}^{\|\cdot\|}_{\mathbf{\Delta},k}(M) = \infty$ (resp. $\mu^{\|\cdot\|}_{\mathbf{\Delta},k}(M) = 0$) if and only if there is no $\Delta \in \mathbf{\Delta}$ satisfying $\operatorname{rank}(M - \Delta) < k$ (resp. $\dim\ker(I_l - \Delta M) \geq k$). Moreover, $\widetilde{\mu}^{\|\cdot\|}_{\mathbf{\Delta},k}(M) = \mu^{\|\cdot\|}_{\mathbf{\Delta},k}(M) = 0$ if $k > \operatorname{rank} M$.

(*b*) For $1 \leq k \leq \min\{q, l\}$ let $m^{(k)}_{q,l} : \mathbb{C}^{q\times l} \to \mathbb{C}$ denote the function which maps $A = [a_{ij}] \in \mathbb{C}^{q\times l}$ to the sum of the moduli of all its subdeterminants of kth order, i.e.

$$m^{(k)}_{q,l}(A) = \sum_{\substack{1\leq i_1<i_2<\ldots i_r\leq q\\ 1\leq j_1<j_2<\ldots j_s\leq l}} \left|\det([a_{i_r j_s}]_{r,s\in\underline{k}})\right|.$$

From the equivalence

$$\dim\ker A \geq l - k + 1 \quad\Leftrightarrow\quad \operatorname{rank}(A) < k \quad\Leftrightarrow\quad m^{(k)}_{q,l}(A) = 0$$

we obtain

$$\widetilde{\mu}^{\|\cdot\|}_{\mathbf{\Delta},k}(M) \;=\; \inf\{\, \|\Delta\| \mid \Delta \in \mathbf{\Delta},\; m^{(k)}_{q,l}(M - \Delta) = 0 \,\},$$

and analogously,

$$\mu^{\|\cdot\|}_{\mathbf{\Delta},k}(M) \;=\; \Big(\; \inf\{\, \|\Delta\| \mid \Delta \in \mathbf{\Delta},\; m^{(l-k+1)}_{l,l}(I_l - \Delta M) = 0 \,\} \;\Big)^{-1}.$$

The functions $(\Delta, M) \mapsto m^{(l-k+1)}_{l,l}(I_l - \Delta M)$ and $(\Delta, M) \mapsto m^{(k)}_{q,l}(M - \Delta)$ are continuous. Thus Lemma 1.7.1 yields the following.

(*i*) With the convention $\min\emptyset = \infty$ the infima in (6.21) and (6.22) are minima.

(*ii*) The functions $\widetilde{\mu}^{\|\cdot\|}_{\mathbf{\Delta},k} : \mathbb{C}^{l\times q} \to [0, \infty]$ are lower semi-continuous.

(*iii*) The functions $\mu^{\|\cdot\|}_{\mathbf{\Delta},k} : \mathbb{C}^{q\times l} \to [0, \infty)$ are upper semi-continuous.

(*c*) By using the Tarski-Seidenberg Theorem (see Chapter 3) the following statement is easily verified: If the underlying norm $\|\cdot\| : \operatorname{span}_{\mathbb{R}} \mathbf{\Delta} \to \mathbb{R}$ is a semi-algebraic function and $\mathbf{\Delta}$ is a semi-algebraic subset of $\mathbb{C}^{q\times l}$ (rep. $\mathbb{C}^{l\times q}$) then the maps $\widetilde{\mu}^{\|\cdot\|}_{\mathbf{\Delta},k} : \mathbb{C}^{q\times l} \to [0, \infty]$, (resp. the maps $\mu^{\|\cdot\|}_{\mathbf{\Delta},k} : \mathbb{C}^{q\times l} \to [0, \infty)$) are semi-algebraic functions. Hence they are analytic on an open and dense semi-algebraic subset of $\mathbb{C}^{q\times l}$.

(*d*) Suppose $l = q =: n$ and let $M \in \mathbb{C}^{n\times n}$ be nonsingular. Then the relation $M - \Delta = (I_n - \Delta M^{-1})M$ implies that $\operatorname{rank}(M-\Delta) < k$ if and only if $\operatorname{rank}(I_n - \Delta M^{-1}) < k$. The latter holds if and only if $\dim\ker(I_n - \Delta M^{-1}) \geq n+1-k$. Thus we have

$$\widetilde{\mu}^{\|\cdot\|}_{\mathbf{\Delta},k}(M) = \left(\mu^{\|\cdot\|}_{\mathbf{\Delta},n+1-k}(M^{-1})\right)^{-1}.$$

(*e*) Let $\mathcal{E}_\lambda(A)$ denote the eigenspace of $A\in\mathbb{C}^{n\times n}$, $n\in\mathbb{N}$, to the eigenvalue $\lambda\in\mathbb{C}$. Let $M\in\mathbb{C}^{q\times l}$, $\Delta\in\mathbb{C}^{l\times q}$. It is easily seen that the maps

$$\begin{array}{ll} \phi_M : \mathcal{E}_1(\Delta M) \to \mathcal{E}_1(M\Delta), & \phi_M(x) := Mx \\ \psi_\Delta : \mathcal{E}_1(M\Delta) \to \mathcal{E}_1(\Delta M), & \psi_\Delta(x) := \Delta x \end{array}$$

are linear isomorphisms, and $\phi_M^{-1} = \psi_\Delta$. From this we obtain the following further characterizations of $\mu^{\|\cdot\|}_{\mathbf{\Delta},k}(M)$.

$$\begin{aligned} \mu^{\|\cdot\|}_{\mathbf{\Delta},k}(M) &= \left(\ \min\{\ \|\Delta\| \mid \Delta\in\mathbf{\Delta},\ \dim\ker(I_l - \Delta M) \geq k\ \}\ \right)^{-1} \\ &= \left(\ \min\{\ \|\Delta\| \mid \Delta\in\mathbf{\Delta},\ \dim\mathcal{E}_1(\Delta M) \geq k\ \}\ \right)^{-1} \\ &= \left(\ \min\{\ \|\Delta\| \mid \Delta\in\mathbf{\Delta},\ \dim\mathcal{E}_1(M\Delta) \geq k\ \}\ \right)^{-1} \\ &= \left(\ \min\{\ \|\Delta\| \mid \Delta\in\mathbf{\Delta},\ \dim\ker(I_q - M\Delta) \geq k\ \}\ \right)^{-1} \end{aligned}$$

(*f*) Suppose that $\mathbf{\Delta}$ is a real or complex vector space, and let $M\in\mathbf{\Delta}$. Then for $\Delta\in\mathbb{C}^{l\times q}$ we have that $\Delta\in\mathbf{\Delta}$ if and only if $X := M-\Delta\in\mathbf{\Delta}$. Thus we can rewrite Definition (6.22) in the form

$$\widetilde{\mu}^{\|\cdot\|}_{\mathbf{\Delta},k}(M) = \min\{\ \|M-X\| \mid X\in\mathbf{\Delta},\ \operatorname{rank}(X) < k\ \}.$$

Thus $\widetilde{\mu}^{\|\cdot\|}_{\mathbf{\Delta},k}(M)$ is the distance of M to the set $\{\ X\in\mathbf{\Delta} \mid \operatorname{rank}(X) < k\ \}$.

The following quantities are closely related to the generalized μ-functions.

Definition 6.2.2 *Let $(\mathbf{\Delta}, \|\cdot\|)\in\mathcal{P}_{l,q}$ and let $k\in\mathbb{N}$. We set*

$$\nu^{\|\cdot\|}_{\mathbf{\Delta},k}(X,Y) := \inf\{\ \|\Delta\| \mid \Delta\in\mathbf{\Delta},\ \Delta X = Y\ \}, \quad \textit{where } (X,Y)\in\mathbb{C}^{l\times k}\times\mathbb{C}^{q\times k}. \tag{6.23}$$

Here are some basic properties of the functions $\nu^{\|\cdot\|}_{\mathbf{\Delta},k}$.

(*a*) We have $\nu^{\|\cdot\|}_{\mathbf{\Delta},k}(X,Y) = \infty$ if and only if there is no $\Delta\in\mathbf{\Delta}$ satisfying $\Delta X = Y$.

(*b*) The function $(\Delta,(X,Y)) \mapsto \Delta X - Y$ is continuous. Hence, Lemma 1.7.1 yields the following.

  (*i*) With the convention $\min\emptyset = \infty$ the infimum in (6.23) a minimum.

  (*ii*) The functions $\nu^{\|\cdot\|}_{\mathbf{\Delta},k} : \mathbb{C}^{l\times k}\times\mathbb{C}^{q\times k} \to [0,\infty]$ are lower semi-continuous.

(*c*) If the underlying norm $\|\cdot\| : \operatorname{span}_{\mathbb{R}}\mathbf{\Delta} \to \mathbb{R}$ is a semi-algebraic function and $\mathbf{\Delta}$ is a semi-algebraic subset of $\mathbb{C}^{l\times q}$ then the functions $\nu^{\|\cdot\|}_{\mathbf{\Delta},k} : \mathbb{C}^{l\times k}\times\mathbb{C}^{q\times k} \to [0,\infty]$ are semi-algebraic.

(*d*) For every nonsingular matrix $P\in\mathbb{C}^{k\times k}$ we have

$$\nu^{\|\cdot\|}_{\mathbf{\Delta},k}(XP,YP) = \nu^{\|\cdot\|}_{\mathbf{\Delta},k}(X,Y) \tag{6.24}$$

The next proposition connects the functions $\nu_{\mathbf{\Delta},k}^{\|\cdot\|}(\cdot,\cdot)$ with the generalized μ-functions. In order to state it we need more notation. By $\mathcal{G}_{n,k}$ we denote the set of $X\in\mathbb{C}^{n\times k}$ whose columns form a basis of a k-dimensional subspace of $\mathbb{C}^n$, i.e. $\mathcal{G}_{n,k} := \{\, X\in\mathbb{C}^{n\times k} \mid \operatorname{rank}(X)=k \,\}$. By $\mathcal{ST}_{n,k}$ we denote the Stiefel-manifold of all orthogonal k-frames in $\mathbb{C}^n$, i.e. $\mathcal{ST}_{n,k} := \{\, X\in\mathbb{C}^{n\times k} \mid X^*X = I_k \,\}$. Note that $\mathcal{G}_{n,k}$ is open, $\mathcal{ST}_{n,k}$ is compact, and

$$\mathcal{G}_{n,k} = \{\, XP \mid X\in\mathcal{ST}_{n,k},\ P\in\mathcal{G}_{k,k} \,\}. \tag{6.25}$$

Moreover, $\{\, \operatorname{range}(X) \mid X\in\mathcal{G}_{n,k} \,\} = \{\, \operatorname{range}(X) \mid X\in\mathcal{ST}_{n,k} \,\} = \mathfrak{G}_{n,k}$, the Grassmann manifold of k-dimensional subspaces of $\mathbb{C}^n$.

Proposition 6.2.3 *Let* $(\mathbf{\Delta},\|\cdot\|)\in\mathcal{P}_{l,q}$. *Then*

(*i*) *For every* $M\in\mathbb{C}^{q\times l}$, $k\in\underline{l}$,

$$\mu_{\mathbf{\Delta},k}^{\|\cdot\|}(M) = \left(\min_{X\in\mathcal{ST}_{l,k}} \nu_{\mathbf{\Delta},k}^{\|\cdot\|}(MX,X) \right)^{-1} \tag{6.26}$$

$$= \left(\min_{X\in\mathcal{G}_{l,k}} \nu_{\mathbf{\Delta},k}^{\|\cdot\|}(MX,X) \right)^{-1} \tag{6.27}$$

(*ii*) *For every* $M\in\mathbb{C}^{l\times q}$, $k\in\underline{q}$,

$$\widetilde{\mu}_{\mathbf{\Delta},k}^{\|\cdot\|}(M) = \min_{X\in\mathcal{ST}_{q,q+1-k}} \nu_{\mathbf{\Delta},q+1-k}^{\|\cdot\|}(X,MX) \tag{6.28}$$

$$= \min_{X\in\mathcal{G}_{q,q+1-k}} \nu_{\mathbf{\Delta},q+1-k}^{\|\cdot\|}(X,MX). \tag{6.29}$$

Proof: Since the function $\nu_{\mathbf{\Delta},k}^{\|\cdot\|}(\cdot,\cdot)$ is lower semi-continuous and $\mathcal{ST}_{l,k}$ is compact the minimum in (6.26) is attained. By (6.24) and (6.25) the right hand sides of the equations (6.26) and (6.27) coincide. Let $\mathcal{X}$ be a k-dimensional subspace of $\mathbb{C}^l$ and let $X\in\mathcal{G}_{l,k}$ be any matrix whose columns form a basis of $\mathcal{X}$. Then the following equivalences hold.

$$\mathcal{X}\subseteq\ker(I_l-\Delta M) \quad\Leftrightarrow\quad (I_l-\Delta M)X=0 \quad\Leftrightarrow\quad \Delta MX = X.$$

Therefore

$$\begin{aligned}
\min_{X\in\mathcal{G}_{l,k}} \nu_{\mathbf{\Delta},k}^{\|\cdot\|}(MX,X) &= \min_{X\in\mathcal{G}_{l,k}} \min\{\, \|\Delta\| \mid \Delta\in\mathbf{\Delta},\ \Delta MX = X \,\} \\
&= \min\{\, \|\Delta\| \mid \Delta\in\mathbf{\Delta},\ \exists X\in\mathcal{G}_{l,k} \colon \Delta MX = X \,\} \\
&= \min\{\, \|\Delta\| \mid \Delta\in\mathbf{\Delta},\ \dim\ker(I_l-\Delta M)\ge k \,\} \\
&= \mu_{\mathbf{\Delta},k}^{\|\cdot\|}(M)^{-1}.
\end{aligned}$$

Thus (6.26) and (6.27) are shown. The proof of (6.28) and (6.29) is analogous if one uses the equivalence

$$\operatorname{rank}(M-\Delta)<k \quad\Leftrightarrow\quad \dim\ker(M-\Delta)\ge q+1-k. \qquad \square$$

To the best of our knowledge there is no general method available to compute the quantities $\nu_{\mathbf{\Delta},k}^{\|\cdot\|}(X,Y)$ for arbitrary $(\mathbf{\Delta},\|\cdot\|)\in\mathcal{P}_{l,q}$, $k\in\mathbb{N}$ and $(X,Y)\in\mathbb{C}^{l\times k}\times\mathbb{C}^{q\times k}$. However, explicit formulas can be given in the following cases.

(a) $k=1$, $\mathbf{\Delta}=\mathbb{F}^{l\times q}$, $(X,Y)\in\mathbb{F}^l\times\mathbb{F}^q$, and the underlying norm $\|\cdot\|$ is rank-1-compatible with an operator norm.

(b) k is arbitrary, $\mathbf{\Delta}=\mathbb{F}^{l\times q}$, and the norm $\|\cdot\|$ is unitarily invariant.

The case (b) will be discussed in the next two sections. For case (a) we have the following result. We use the notation

$$\nu_{\mathbb{F}}^{\|\cdot\|}(x,y) := \nu_{\mathbb{F}^{q\times l},1}^{\|\cdot\|}(x,y).$$

Proposition 6.2.4 *Let $\|\cdot\|_\alpha$ be a norm on $\mathbb{F}^q$ and $\|\cdot\|_\beta$ be a norm on $\mathbb{F}^l$, $\mathbb{F}\in\{\mathbb{R},\mathbb{C}\}$. Let $\|\cdot\|$ be a norm on $\mathbb{F}^{l\times q}$ which is rank-1-compatible with the operator norm $\|\cdot\|_{\alpha,\beta}$. Let $x\in\mathbb{F}^q\setminus\{0\}$, $y\in\mathbb{F}^l$. Then there is $w\in\mathbb{F}^q$ satisfying $w^Tx=1$ and $\|w\|_\alpha^D=\|x\|_\alpha^{-1}$. Set $\Delta_0=yw^T$. Then $\Delta_0x=y$, and $\nu_{\mathbb{F}}^{\|\cdot\|}(x,y)=\|\Delta_0\|=\|\Delta_0\|_{\alpha,\beta}=\|y\|_\beta/\|x\|_\alpha$.*

Proof: The relation $\Delta x=y$ implies $\|\Delta\|\geq\|\Delta\|_{\alpha,\beta}\geq\|y\|_\beta/\|x\|_\alpha$. The existence of the vector w follows from the Hahn-Banach theorem. $\square$

Let us remark that Theorem 2.5.2 and Corollary 2.5.3 can be derived by combining Propositions 6.25 and 6.2.4. We leave this to the reader as an exercise.

Next, we discuss the connection between ν-functions with respect to real and complex perturbation structures.

Definition 6.2.5 *A perturbation structure $(\mathbf{\Delta},\|\cdot\|)\in\mathcal{P}_{l,q}$ is said to be conjugation invariant if it satisfies the following conditions.*

- *If $\Delta\in\mathbf{\Delta}$ then $\overline{\Delta}\in\mathbf{\Delta}$ and $\Re\Delta\in\mathbf{\Delta}$.*
- *For every $\Delta\in\mathbf{\Delta}$, $\|\overline{\Delta}\|=\|\Delta\|$.*

For a conjugation invariant perturbation structure we set $\mathbf{\Delta}_{\mathbb{R}}:=\mathbf{\Delta}\cap\mathbb{R}^{l\times q}$.

Observe that if $(\mathbf{\Delta},\|\cdot\|)\in\mathcal{P}_{l,q}$ is conjugation invariant then $(\mathbf{\Delta}_{\mathbb{R}},\|\cdot\|_{\mathbb{R}})\in\mathcal{P}_{l,q}$, where $\|\cdot\|_{\mathbb{R}}$ denotes the restriction of the norm $\|\cdot\|:\mathrm{span}_{\mathbb{R}}(\mathbf{\Delta})\to[0,\infty)$ to the vector space $\mathrm{span}_{\mathbb{R}}(\mathbf{\Delta}_{\mathbb{R}})$. In the sequel we will not distinguish these norms in our notation and denote both $\|\cdot\|$ and $\|\cdot\|_{\mathbb{R}}$ by the same symbol $\|\cdot\|$. The simplest examples of conjugation invariant perturbation structures are the pairs $(\mathbb{C}^{l\times q},\|\cdot\|)$, where $\|\cdot\|$ is a unitarily invariant norm or an operator norm subordinate to absolute norms. In the proof of the proposition below we will use the following elementary fact. Let

$$W_k := \begin{bmatrix} \frac{1}{2}I_k & \frac{1}{2i}I_k \\ \frac{1}{2}I_k & -\frac{1}{2i}I_k \end{bmatrix}, \qquad k\in\mathbb{N}. \tag{6.30}$$

Then for every $X\in\mathbb{C}^{n\times k}$,

$$[\,X\ \ \overline{X}\,]W_k = [\,\Re X\ \ \Im X\,] \tag{6.31}$$

Proposition 6.2.6 *Let $(\mathbf{\Delta},\|\cdot\|)\in\mathcal{P}_{l,q}$ be a conjugation invariant perturbation structure. Then for any $X\in\mathbb{C}^{l\times k}$, $Y\in\mathbb{C}^{q\times k}$,*

$$\nu_{\mathbf{\Delta}_{\mathbb{R}},k}^{\|\cdot\|}(X,Y)=\nu_{\mathbf{\Delta},2k}^{\|\cdot\|}(\,[X\ \ \overline{X}],[Y\ \ \overline{Y}]\,)=\nu_{\mathbf{\Delta},2k}^{\|\cdot\|}(\,[\Re X\ \ \Im X],[\Re Y\ \ \Im Y]\,).$$

Proof: Let $\Delta\in\mathbf{\Delta}_{\mathbb{R}}$ be such that $\Delta X=Y$. Then $\Delta\overline{X}=\overline{Y}$, $\Delta\Re X=\Re Y$ and $\Delta\Im X=\Im Y$. Thus $\Delta[X\ \overline{X}]=[Y\ \overline{Y}]$ and $\Delta[\Re X\ \Im X]=[\Re Y\ \Im Y]$. Thus $\nu^{\|\cdot\|}_{\mathbf{\Delta}_{\mathbb{R}},k}(X,Y)\geq\nu^{\|\cdot\|}_{\mathbf{\Delta},2k}([X\ \overline{X}],[Y\ \overline{Y}])$ and $\nu^{\|\cdot\|}_{\mathbf{\Delta}_{\mathbb{R}},k}(X,Y)\geq\nu^{\|\cdot\|}_{\mathbf{\Delta},2k}([\Re X\ \Im X],[\Re Y\ \Im Y])$. However, from (6.24) and (6.31) it follows that

$$\nu^{\|\cdot\|}_{\mathbf{\Delta},2k}([\Re X\ \Im X],[\Re Y\ \Im Y])=\nu^{\|\cdot\|}_{\mathbf{\Delta},2k}([X\ \overline{X}]W_k,\ [Y\ \overline{Y}]W_k)=\nu^{\|\cdot\|}_{\mathbf{\Delta},2k}([X\ \overline{X}],\ [Y\ \overline{Y}]).$$

Suppose there is no $\Delta\in\mathbf{\Delta}_{\mathbb{R}}$ such that $\Delta X=Y$. Then $\nu^{\|\cdot\|}_{\mathbf{\Delta},2k}([X\ \overline{X}],\ [Y\ \overline{Y}])=\nu^{\|\cdot\|}_{\mathbf{\Delta}_{\mathbb{R}},k}(X,Y)=\infty$. Supppose $\nu^{\|\cdot\|}_{\mathbf{\Delta},2k}([X\ \overline{X}],\ [Y\ \overline{Y}])<\infty$. Then there is a $\Delta_0\in\mathbf{\Delta}$ satisfying $\|\Delta_0\|=\nu^{\|\cdot\|}_{\mathbf{\Delta},2k}([X\ \overline{X}],[Y\ \overline{Y}])$ and $\Delta_0[X\ \overline{X}]=[Y\ \overline{Y}]$. The latter implies that $\Delta_0X=Y$ and $\overline{\Delta_0}X=Y$. Hence, $(\Re\Delta_0)X=Y$. We have $\Re\Delta_0\in\mathbf{\Delta}_{\mathbb{R}}$ and $\|\Re\Delta_0\|\leq(1/2)(\|\Delta_0\|+\|\overline{\Delta_0}\|)=\|\Delta_0\|$. Therefore, $\nu^{\|\cdot\|}_{\mathbf{\Delta}_{\mathbb{R}},k}(X,Y)\leq\nu^{\|\cdot\|}_{\mathbf{\Delta},2k}([X\ \overline{X}],[Y\ \overline{Y}])$. In summary, $\nu^{\|\cdot\|}_{\mathbf{\Delta}_{\mathbb{R}},k}(X,Y)=\nu^{\|\cdot\|}_{\mathbf{\Delta},2k}([X\ \overline{X}],[Y\ \overline{Y}])$. $\square$

By combining Proposition 6.2.6 with Proposition 6.2.3 we obtain the following corollary.

Corollary 6.2.7 *Let $(\mathbf{\Delta},\|\cdot\|)\in\mathcal{P}_{l,q}$ be a conjugation invariant perturbation structure and let $k\in\mathbb{N}$. Then*

(i) *For every $M\in\mathbb{C}^{q\times l}$, $k\in\underline{l}$,*

$$\begin{aligned}\mu^{\|\cdot\|}_{\mathbf{\Delta}_{\mathbb{R}},k}(M) &= \left(\min_{X\in\mathcal{G}_{l,k}}\nu^{\|\cdot\|}_{\mathbf{\Delta},2k}(\,[MX\ \overline{MX}],[X\ \overline{X}]\,)\right)^{-1}\\ &= \left(\min_{X\in\mathcal{G}_{l,k}}\nu^{\|\cdot\|}_{\mathbf{\Delta},2k}(\,[\Re(MX)\ \Im(MX)],[\Re X\ \Im X]\,)\right)^{-1}\end{aligned}$$

(ii) *For every $M\in\mathbb{C}^{l\times q}$,$k\in\underline{q}$,*

$$\begin{aligned}\widetilde{\mu}^{\|\cdot\|}_{\mathbf{\Delta}_{\mathbb{R}},k}(M) &= \min_{X\in\mathcal{G}_{q,q+1-k}}\nu^{\|\cdot\|}_{\mathbf{\Delta},2(q+1-k)}(\,[X\ \overline{X}],[MX\ \overline{MX}]\,)\\ &= \min_{X\in\mathcal{G}_{q,q+1-k}}\nu^{\|\cdot\|}_{\mathbf{\Delta},2(q+1-k)}(\,[\Re X\ \Im X],[\Re(MX)\ \Im(MX)]\,).\end{aligned}$$

6.3 The functions $\nu^{\|\cdot\|}_{\mathbb{F},k}(X,Y)$ for unitarily invariant norms

In this section we discuss the functions $\nu^{\|\cdot\|}_{\mathbb{F},k}(X,Y):=\nu^{\|\cdot\|}_{\mathbb{F}^{q\times l},k}(X,Y)$ for unitarily invariant norms. We need the following lemma.

Lemma 6.3.1 *Let $\Delta_0,\Delta_1\in\mathbb{F}^{l\times q}$ and suppose that $(\ker\Delta_1)^{\perp}\subseteq\ker\Delta_0$. Then, for every unitarily invariant norm, $\|\Delta_0\|\leq\|\Delta_0+\Delta_1\|$.*

Proof: Let $U\in\mathbb{F}^{k\times k}$ be the unitary matrix satisfying $U\xi=-\xi$ for all $\xi\in(\ker\Delta_1)^{\perp}$ and $U\xi=\xi$ for all $\xi\in\ker\Delta_1$. Let $\xi_0\in(\ker\Delta_1)^{\perp}\subseteq\ker\Delta_0$ and $\xi_1\in\ker\Delta_1$. Then

$$\begin{aligned}\Delta_0U(\xi_0+\xi_1) &= \Delta_0(-\xi_0+\xi_1)=\Delta_0\xi_1=\Delta_0(\xi_0+\xi_1).\\ \Delta_1U(\xi_0+\xi_1) &= \Delta_1(-\xi_0+\xi_1)=-\Delta_1\xi_0=-\Delta_1(\xi_0+\xi_1).\end{aligned}$$

Thus $\Delta_0U=\Delta_0$ and $\Delta_1U=-\Delta_1$. Therefore, by the unitary invariance of the norm, $\|\Delta_0-\Delta_1\|=\|(\Delta_0+\Delta_1)U\|=\|\Delta_0+\Delta_1\|$. Hence, by convexity, $\|\Delta_0\|\leq\frac{1}{2}\left(\|\Delta_0+\Delta_1\|+\|\Delta_0-\Delta_1\|\right)=\|\Delta_0+\Delta_1\|$. $\square$

Proposition 6.3.2 *Let* $(X,Y)\in\mathbb{F}^{l\times k}\times\mathbb{F}^{q\times k}$ *and let* $\|\cdot\|$ *be a unitarily invariant norm on* $\mathbb{F}^{l\times q}$.

(*i*) *Suppose that* $\ker X\not\subseteq\ker Y$. *Then there exists no matrix* $\Delta\in\mathbb{F}^{l\times k}$ *satisfying* $\Delta X=Y$. *Thus* $\nu^{\|\cdot\|}_{\mathbb{F},k}(X,Y)=\infty$.

(*ii*) *Suppose that* $\ker X\subseteq\ker Y$. *Then there is a unique matrix* $\Delta_0\in\mathbb{F}^{l\times q}$ *satisfying* $\Delta_0X=Y$ *and* $(\operatorname{range}X)^\perp\subseteq\ker\Delta_0$. *We have* $\Delta_0=YX^\dagger$ *and* $\nu^{\|\cdot\|}_{\mathbb{F},k}(X,Y)=\|\Delta_0\|$.

Proof: (*i*). Suppose that $\Delta X=Y$ for some $\Delta\in\mathbb{F}^{l\times q}$. Let $\xi\in\ker X$. Then $0=\Delta X\xi=Y\xi$. Thus $\xi\in\ker Y$. Thus $\ker X\subseteq\ker Y$.

(*ii*). Suppose that $\ker X\subseteq\ker Y$. Set $\Delta_0:=YX^\dagger$. We show that $\Delta_0X=Y$ and $(\operatorname{range}X)^\perp\subseteq\ker\Delta_0$. By definition of the Moore-Penrose generalized inverse we have

$$X^\dagger(X\xi)=\xi \text{ for } \xi\in(\ker X)^\perp,\quad\text{and}\quad X^\dagger x=0 \text{ for } x\in(\operatorname{range}X)^\perp.$$

Thus $\Delta_0x=0$ for $x\in\operatorname{range}(X)^\perp$. Every $\xi\in\mathbb{F}^k$ has a decomposition $\xi=\xi_1+\xi_2$ with $\xi_1\in(\ker X)^\perp$ and $\xi_2\in\ker X$. The assumption kerX $\subseteq$ kerY yields $Y\xi_2=0$. Thus, for every $\xi\in\mathbb{F}^k$,

$$\Delta_0X\xi=YX^\dagger X\xi_1+YX^\dagger X\xi_2=Y\xi_1=Y\xi_1+Y\xi_2=Y\xi.$$

Hence, $\Delta_0X=Y$, and therefore $\nu^{\|\cdot\|}_{\mathbb{F},k}(X,Y)\le\|\Delta_0\|$.

Now, let $\Delta\in\mathbb{F}^{q\times k}$ be any matrix satisfying $\Delta X=Y$. Then we have for the matrix $\Delta_1:=\Delta-\Delta_0$ that $\operatorname{range}X\subseteq\ker\Delta_1$. The latter implies $(\ker\Delta_1)^\perp\subseteq(\operatorname{range}X)^\perp\subseteq\ker\Delta_0$. Thus, Lemma 6.3.1 yields, $\|\Delta_0\|\le\|\Delta_0+\Delta_1\|=\|\Delta\|$. Therefore, $\nu^{\|\cdot\|}_{\mathbb{F},k}(X,Y)=\|\Delta_0\|$.

Finally, suppose that Δ satisfies the additional condition $(\operatorname{range}X)^\perp\subseteq\ker\Delta$. Then we also have $(\operatorname{range}X)^\perp\subseteq\ker\Delta_1$. Hence $\mathbb{F}^k=(\operatorname{range}X)\oplus(\operatorname{range}X)^\perp\subseteq\ker\Delta_1$. Thus $\Delta_1=0$, i.e. $\Delta=\Delta_0$. □

The unitarily invariant norm $\|YX^\dagger\|$ is a function of the singular values of $YX^\dagger$. The next proposition in particular states that the nonzero singular values of $YX^\dagger$ coincide with the square roots of the nonzero eigenvalues of the hermitian pencil $P(\lambda)=Y^*Y-\lambda X^*X$, $\lambda\in\mathbb{C}$. Note that the pencil $P(\lambda)$ is singular (i.e. $\det(P(\lambda))\equiv0$) if and only if $\{0\}\ne\ker X\subseteq\ker Y$.

Proposition 6.3.3 *Let* $(X,Y)\in\mathbb{F}^{l\times k}\times\mathbb{F}^{q\times k}$, $m:=\operatorname{rank}X$ *and* $r:=\operatorname{rank}Y$. *Assume that* $\ker X\subseteq\ker Y$. *Then* $r\le m$. *Moreover, there exist a basis* $\xi_1,\ldots,\xi_m\in\mathbb{F}^l$ *of* $(\ker X)^\perp$ *and nonnegative real numbers* $\lambda_1\ge\lambda_2\ge\ldots\ge\lambda_m\ge0$ *such that* $(Y^*Y)\xi_j=\lambda_j(X^*X)\xi_j$ *for all* $j\in\underline{m}$ *and* $\xi_j^*(X^*X)\xi_k=0$ *for* $j\ne k$. *We have* $\lambda_j\ne0$ *if and only if* $j\in\underline{r}$. *For* $j\in\underline{r}$ *set* $u_j:=\frac{Y\xi_j}{\|Y\xi_j\|_2}$, $v_j:=\frac{X\xi_j}{\|X\xi_j\|_2}$. *Then*

$$YX^\dagger=\sum_{j\in\underline{r}}\sqrt{\lambda_j}\,u_jv_j^*$$

is a singular value decomposition of $YX^\dagger$. *Thus* $\sigma_j(YX^\dagger)=\begin{cases}\sqrt{\lambda_j} & \text{if } j\in\underline{r}\\ 0 & \text{otherwise.}\end{cases}$

Proof: For any matrix $X\in\mathbb{F}^{l\times k}$ we have $(\ker X)^\perp=(\ker X^*X)^\perp=\operatorname{range}X^*X$. Hence, $(\ker X)^\perp$ is an invariant subspace of X^*X and $\ker(X^*X)^\dagger=(\ker X)^\perp$. By definition of the Moore-Penrose inverse the linear endomorphism $(\ker X)^\perp\ni\xi\mapsto(X^*X)^\dagger\xi$ is the inverse to the linear endomorphism $(\ker X)^\perp\ni\xi\mapsto(X^*X)\xi$. The assumption $\ker X\subseteq\ker Y$ implies that

$(\ker X)^\perp \supseteq (\ker Y)^\perp = (\ker Y^*Y)^\perp = \operatorname{range} Y^*Y$. Thus $(X^*X)(X^*X)^\dagger(Y^*Y) = Y^*Y$. Using these facts it is straightforward to verify that the linear operator

$$f : (\ker X)^\perp \to (\ker X)^\perp, \qquad f(\xi) := (X^*X)^\dagger(Y^*Y)\xi$$

is selfadjoint and positive semi-definite with respect to the inner product

$$\langle\cdot|\cdot\rangle : (\ker X)^\perp \times (\ker X)^\perp \to \mathbb{F}, \qquad \langle\xi_1|\xi_2\rangle := \xi_1^*(X^*X)\xi_2.$$

Thus there is a basis $\xi_1,\ldots,\xi_m$ of $(\ker X)^\perp$ such that $f(\xi_j) = \lambda_j\xi_j$, $\lambda_1 \geq \ldots \geq \lambda_m \geq 0$, and

$$\langle\xi_j|\xi_k\rangle = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{otherwise.} \end{cases}$$

The latter implies that the vectors $v_j = X\xi_j/\|X\xi_j\|_2$, $j \in \underline{m}$, form an orthonormal basis of $\operatorname{range} X$ with respect to the canonical inner product of $\mathbb{F}^l$. We have

$$(Y^*Y)\xi_j \;=\; (X^*X)(X^*X)^\dagger(Y^*Y)\xi_j = (X^*X)f(\xi_j) = \lambda_j(X^*X)\xi_j, \qquad j\in\underline{m}.$$

This implies $\xi_k^*(Y^*Y)\xi_j = \lambda_j\langle\xi_k|\xi_j\rangle = 0$ for $k \neq j$ and $\|Y\xi_j\|_2 = \sqrt{\lambda_j}\|X\xi_j\|_2$ for all $j\in\underline{m}$. Let $r' := \max\{\, j\in\underline{m} \mid \lambda_j > 0\}$ and $\Delta_0 := \sum_{k\in\underline{r'}} \sqrt{\lambda_k}\, u_k v_k^*$, where $u_k = Y\xi_k/\|Y\xi_k\|_2$. Then

$$u_k^*u_j = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{otherwise} \end{cases}$$

Thus, $\sqrt{\lambda_1},\ldots,\sqrt{\lambda_{r'}}$ are the singular values of Δ_0 and the v_k, u_k are corresponding singular vectors. We show that $\Delta_0 = YX^\dagger$ and $r' = \operatorname{rank} Y$: For any $\xi\in(\ker X)^\perp$ we have $X^\dagger X\xi = \xi$. Thus the vectors v_j, $j\in\underline{m}$ satisfy

$$YX^\dagger v_j = YX^\dagger\frac{X\xi_j}{\|X\xi_j\|} = \frac{Y\xi_j}{\|X\xi_j\|_2} = \begin{cases} \lambda_j u_j = \Delta_0 v_j & \text{if } j \leq r', \\ 0 = \Delta_0 v_j & \text{otherwise.} \end{cases}$$

Thus $\Delta_0 x = YX^\dagger x$ for all $x\in\operatorname{range} X$. Moreover, we have for $x\in(\operatorname{range} X)^\perp$ that $\Delta_0 x = 0 = YX^\dagger x$. Thus $\Delta_0 = YX^\dagger$ and $r' = \operatorname{rank}\Delta_0 = \operatorname{rank}(YX^\dagger) \leq \operatorname{rank} Y$. However, the linear map $\xi \mapsto Y\xi$, where $\xi\in(\ker Y)^\perp \supseteq (\ker X)^\perp = \operatorname{range}(X^\dagger)$ is injective. Thus $r' = \operatorname{rank} Y$. □

We now specialize the results of the latter two propositions to the case that the matrix X has full column rank.

Corollary 6.3.4 *Let $(X,Y)\in\mathbb{F}^{l\times k}\times\mathbb{F}^{q\times k}$ and suppose that* $\ker X = \{0\}$. *Let $\Delta_0 = YX^\dagger$. Then*

(i) We have $\Delta_0 = Y(X^*X)^{-1}X^*$ and $\Delta_0 X = Y$.

(ii) For all unitarily invariant norms, $\nu^{\|\cdot\|}_{\mathbb{F},k}(X,Y) = \|\Delta_0\|$.

(iii) The eigenvalues of $(X^*X)^{-1}(Y^*Y)$ are nonnegative.

(iv) Let $\lambda_j((X^*X)^{-1}(Y^*Y))$, $j\in\underline{k}$ denote the eigenvalues of $(X^*X)^{-1}(Y^*Y)$ in nonincreasing order. Then

$$\sigma_j(\Delta_0) = \sqrt{\lambda_j((X^*X)^{-1}(Y^*Y))}, \quad j\in\underline{k}.$$

Proof: The first claim holds since $X^\dagger = (X^*X)^{-1}X^*$ if X has full column rank. Claim (ii) is a special case of Proposition 6.3.2. The claims (iii) and (iv) follow from Proposition 6.3.3. □

The next corollary deals with the case when Y^*Y is a scalar multiple of X^*X.

Corollary 6.3.5 *Let* $(X, Y) \in \mathbb{F}^{l\times k} \times \mathbb{F}^{q\times k}$ *and suppose that* $Y^*Y = t^2X^*X$, *where* $t > 0$. *Let* $r = \operatorname{rank} X$ *and* $\Delta_0 = YX^\dagger$. *Then*

(i) *We have* $\ker X = \ker Y$ *and* $\Delta_0 X = Y$.

(ii) *The singular values of* Δ_0 *satisfy* $\sigma_j(\Delta_0) = \begin{cases} t & \text{if } j \in \underline{r}, \\ 0 & \text{otherwise.} \end{cases}$

Proof: The first relation in (i) holds since $\ker X = \ker X^*X$ and $\ker Y = \ker Y^*Y$. The second relation in (i) follows from the first and Proposition 6.3.2. (ii). Let $\lambda_j \geq 0$ and $\xi_j \in (\ker X)^\perp$ be as in Proposition 6.3.3. The relations $(Y^*Y)\xi_j = \lambda_j(X^*X)\xi_j$ and $Y^*Y = t^2\, X^*X$ yield that $0 = (\lambda_j - t^2)(X^*X)\xi_j$. Thus $t = \sqrt{\lambda_j} = \sigma_j(\Delta_0)$. □

So far we dealt with $\nu^{\|\cdot\|}_{\mathbb{R},k}(X, Y)$ only for the special case that X and Y are real matrices. We now drop this restriction.

Proposition 6.3.6 *Let* $X \in \mathbb{C}^{l\times k}$, $Y \in \mathbb{C}^{q\times k}$ *and* $\Delta_0 := [Y\ \overline{Y}]\,[X\ \overline{X}]^\dagger$. *Then*

(i) *We have* $\Delta_0 = [\Re Y\ \Im Y]\,[\Re X\ \Im X]^\dagger \in \mathbb{R}^{2q\times 2l}$.

(ii) *The following statements are equivalent.*

- (a) $\ker[X\ \overline{X}] \subseteq \ker[Y\ \overline{Y}]$.
- (b) $\ker[\Re X\ \Im X] \subseteq \ker[\Re Y\ \Im Y]$.
- (c) *There exists* $\Delta \in \mathbb{R}^{q\times l}$ *such that* $\Delta X = Y$.
- (d) $\Delta_0 X = Y$.
- (e) $\nu^{\|\cdot\|}_{\mathbb{R},k}(X, Y) \neq \infty$ *for any norm* $\|\cdot\|$ *on* $\mathbb{R}^{q\times l}$.

(iii) *For every unitarily invariant norm we have*

$$\begin{aligned}
\nu^{\|\cdot\|}_{\mathbb{R},k}(X, Y) &= \begin{cases} \|\Delta_0\| & \text{if } \ker[X\ \overline{X}] \subseteq \ker[Y\ \overline{Y}], \\ \infty & \text{otherwise.} \end{cases} \\
&= \nu^{\|\cdot\|}_{\mathbb{C},2k}([X\ \overline{X}], [Y\ \overline{Y}]) \\
&= \nu^{\|\cdot\|}_{\mathbb{C},2k}([\Re X\ \Im X], [\Re Y\ \Im Y]).
\end{aligned}$$

Proof: (i). Let $A := [\Re X\ \Im X]\,W_k$, where W_k is the matrix defined in (6.30). Using the fact that $\sqrt{2}W_k$ is unitary it is easily checked that the matrix $Z := W_k^{-1}[\Re X\ \Im X]^\dagger$ satisfies the Penrose conditions $AZA = A$, $ZAZ = Z$, $(AZ)^* = AZ$, $(ZA)^* = ZA$. Thus $W_k^{-1}[\Re X\ \Im X]^\dagger = ([\Re X\ \Im X]\,W_k)^\dagger$. Hence,

$$\begin{aligned}
[Y\ \overline{Y}]\,[X\ \overline{X}]^\dagger &= ([\Re Y\ \Im Y]\,W_k)([\Re X\ \Im X]\,W_k)^\dagger && (6.32) \\
&= ([\Re Y\ \Im Y]\,W_k)W_k^{-1}[\Re X\ \Im X]^\dagger && (6.33) \\
&= [\Re Y\ \Im Y][\Re X\ \Im X]^\dagger. && (6.34)
\end{aligned}$$

(ii). The equivalence $(c) \Leftrightarrow (e)$ is obvious, and the equivalence $(a) \Leftrightarrow (b)$ follows from (6.31). The remaining equivalences are immediate from Proposition 6.3.2 and the fact that conditions $(c), (d)$ can be rewritten in the form

(c') There exists $\Delta \in \mathbb{R}^{q\times l}$ such that $\Delta[\Re X \;\; \Im X] = [\Re Y \;\; \Im Y]$.

(d') $\Delta_0[\Re X \;\; \Im X] = [\Re Y \;\; \Im Y]$.

Claim (iii) follows by combining Proposition 6.2.6 with Proposition 6.3.2. □

6.4 The functions $\nu^{\|\cdot\|}_{\mathbb{F},k}(X,Y)$ for spectral norm

We now give useful characterizations of $\nu^{\|\cdot\|}_{\mathbb{F},k}(X,Y)$ with respect to spectral norm. In doing so we use the conventions

$$\tfrac{a}{0} = \infty \text{ for } a > 0, \qquad \tfrac{0}{0} = 0, \qquad \min\emptyset = \infty. \tag{6.35}$$

and the notation

$$\nu^{(2)}_{\mathbb{F},k}(X,Y) := \nu^{\|\cdot\|_2}_{\mathbb{F},k}(X,Y), \qquad \mathbb{F} = \mathbb{C} \text{ or } \mathbb{R}.$$

Proposition 6.4.1 *Let $(X,Y)\in\mathbb{F}^{l\times k}\times\mathbb{F}^{q\times k}$. Then*

$$\begin{aligned} \nu^{(2)}_{\mathbb{F},k}(X,Y) &= \max\left\{ \frac{\|Y\xi\|_2}{\|X\xi\|_2} \;\middle|\; \xi\in\mathbb{F}^k \right\} \\ &= \min\{\, \theta \geq 0 \mid \textit{the matrix } \theta^2 X^*X - Y^*Y \textit{ is positive semi-definite} \,\}. \end{aligned}$$

Proof: Suppose first that $\ker X \not\subseteq \ker Y$. Then there is $\xi_0\in\mathbb{F}^k$ such that $X\xi_0 = 0$ and $Y\xi_0 \neq 0$. Thus $\|Y\xi_0\|_2/\|X\xi_0\|_2 = \infty$, and we have $\xi_0^*(\theta^2X^*X - Y^*Y)\xi_0 < 0$ for all $\theta \geq 0$. Hence, the proposition holds in this case. Suppose now, that $\ker X \subseteq \ker Y$. Then, for every $\xi\in\mathbb{F}^k$, $X\xi = 0$ implies $Y\xi = 0$. Therefore,

$$\max_{\xi\in\mathbb{F}^k}\frac{\|Y\xi\|_2}{\|X\xi\|_2} = \max_{\xi\in\ker X^\perp}\frac{\|Y\xi\|_2}{\|X\xi\|_2} =: c.$$

We obviously have,

$$\begin{aligned} c &= \max\{\, \theta\geq 0 \mid \forall\,\xi\in\mathbb{F}^k : \theta^2\|X\xi\|_2^2 - \|Y\xi\|_2^2 \geq 0 \,\} \\ &= \max\{\, \theta\geq 0 \mid \forall\,\xi\in\mathbb{F}^k : \xi^*(\theta^2X^*X - Y^*Y)\xi \geq 0 \,\}. \end{aligned}$$

Let $\Delta_0 = YX^\dagger$. Then by Proposition 6.3.2, $\Delta_0 X = Y$. Each vector $x\in\mathbb{F}^l$ has a decomposition $x = X\xi + \widetilde{x}$, where $\xi\in\mathbb{F}^k$ and $\widetilde{x}\in(\operatorname{range} X)^\perp$. We have

$$\frac{\|\Delta_0 x\|_2}{\|x\|_2} = \frac{\|\Delta_0(X\xi+\widetilde{x})\|_2}{\|X\xi+\widetilde{x}\|_2} = \frac{\|Y\xi\|_2}{\sqrt{\|X\xi\|_2^2 + \|\widetilde{x}\|_2^2}} \leq \frac{\|Y\xi\|_2}{\|X\xi\|_2}.$$

Thus $\|\Delta_0\|_2 \leq c$. On the other hand, $\Delta_0 X\xi = Y\xi$ implies $\|\Delta_0\|_2\,\|X\xi\|_2 \geq \|Y\xi\|_2$ for all ξ. Therefore, $\|\Delta_0\| \geq c$. □

Now, we consider $\nu^{(2)}_{\mathbb{R},k}(X,Y)$ for not necessarily real matrices X, Y. To this end we need the lemma below which brings hermitian-symmetric inequalities into play. It is a slight variant of a lemma by Fitzgerald and Horn [29].

Lemma 6.4.2 *Let $(H,S)\in\mathcal{HS}_n$ and let $\mathcal{X}$ be a subspace of $\mathbb{C}^n$. The following assertions are equivalent.*

(a) $\quad x^*Hx \geq |x^TSx| \quad$ *for all* $\quad x\in\mathcal{X}$.

(b) $\quad \begin{bmatrix} x \\ \overline{y} \end{bmatrix}^* \begin{bmatrix} H & \overline{S} \\ S & \overline{H} \end{bmatrix} \begin{bmatrix} x \\ \overline{y} \end{bmatrix} \geq 0 \quad$ *for all* $\quad \begin{bmatrix} x \\ \overline{y} \end{bmatrix} \in \mathcal{X}\times\overline{\mathcal{X}}$.

Proof: Note first that for every $x,y\in\mathbb{C}^n$,

$$\begin{bmatrix} x \\ \overline{y} \end{bmatrix}^* \begin{bmatrix} H & \overline{S} \\ S & \overline{H} \end{bmatrix} \begin{bmatrix} x \\ \overline{y} \end{bmatrix} = x^*Hx + y^*Hy + 2\,\Re(x^TSy). \tag{6.36}$$

Let $z\in\mathbb{C}$ be such that $x^TS(zx) = -|x^TSx|$. Then

$$\begin{bmatrix} x \\ \overline{zx} \end{bmatrix}^* \begin{bmatrix} H & \overline{S} \\ S & \overline{H} \end{bmatrix} \begin{bmatrix} x \\ \overline{zx} \end{bmatrix} = 2(x^*Hx - |x^TSx|).$$

Thus (b) implies (a). Suppose now that condition (a) holds. Then

$$\begin{aligned} (x+y)^*H(x+y) &\geq |(x+y)^TS(x+y)| \quad \forall x,y\in\mathcal{X}, \\ (x-y)^*H(x-y) &\geq |(x-y)^TS(x-y)| \quad \forall x,y\in\mathcal{X}. \end{aligned}$$

Addition of these inequalities yields

$$\begin{aligned} 2(x^*Hx + y^*Hy) &\geq |(x+y)^TS(x+y)| + |(x-y)^TS(x-y)| \\ &\geq |(x+y)^TS(x+y) - (x-y)^TS(x-y)| \\ &= 4\,|\Re(x^TSy)| \\ &\geq -4\Re(x^TSy). \end{aligned}$$

Now, (b) follows from the identity (6.36). □

The next proposition relates the quantity $\nu^{(2)}_{\mathbb{R},k}(X,Y)$ to the pencils

$$\begin{aligned} R_\theta(X,Y) &:= \theta^2[\Re X \;\; \Im X]^T\,[\Re X \;\; \Im X] - [\Re Y \;\; \Im Y]^T\,[\Re Y \;\; \Im Y] \\ &= \begin{bmatrix} \theta^2(\Re X)^T(\Re X) - (\Re Y)^T(\Re Y) & \theta^2(\Re X)^T(\Im X) - (\Re Y)^T(\Im Y) \\ \theta^2(\Re X)^T(\Im X) - (\Re Y)^T(\Im Y) & \theta^2(\Im X)^T(\Im X) - (\Im Y)^T(\Im Y) \end{bmatrix} \end{aligned}$$

and

$$C_\theta(X,Y) := \theta^2[X\;\;\overline{X}]^*\,[X\;\;\overline{X}] - [Y\;\;\overline{Y}]^*\,[Y\;\;\overline{Y}] = \begin{bmatrix} \theta^2X^*X - Y^*Y & \overline{\theta^2X^TX - Y^TY} \\ \theta^2X^TX - X^TX & \overline{\theta^2X^*X - Y^*Y} \end{bmatrix}.$$

Note that these pencils are congruent since $R_\theta(X,Y) = W_k^*C_\theta(X,Y)W_k$, where W_k is the matrix in (6.30).

Proposition 6.4.3 *Let $(X,Y)\in\mathbb{C}^{l\times k}\times\mathbb{C}^{q\times k}$. Then the following identitites hold.*

$$\begin{aligned} \nu^{(2)}_{\mathbb{R},k}(X,Y) &= \min\{\,\theta\geq 0 \mid R_\theta(X,Y) \text{ is positive semi-definite }\} \\ &= \min\{\,\theta\geq 0 \mid C_\theta(X,Y) \text{ is positive semi-definite }\} \\ &= \min\left\{\,\theta\geq 0 \;\middle|\; \forall\,\xi\in\mathbb{C}^k:\ \xi^*(\theta^2X^*X - Y^*Y)\xi \geq |\,\xi^T(\theta^2X^TX - Y^TY)\xi\,|\,\right\} \\ &= \max_{\xi\in\mathbb{C}^k}\frac{\|\Re(Y\xi)\|_2}{\|\Re(X\xi)\|_2}. \end{aligned}$$

Proof: According to Proposition 6.3.6 we have $\nu^{(2)}_{\mathbb{R},k}(X,Y) = \nu^{(2)}_{\mathbb{C},2k}([\Re X\ \ \Im X],[\Re Y\ \ \Im Y]) = \nu^{(2)}_{\mathbb{C},2k}([X\ \ \overline{X}],[Y\ \ \overline{Y}])$. Thus the first two identities follow by applying Proposition 6.4.1 to the matrix pairs $([\Re X\ \ \Im X],[\Re Y\ \ \Im Y])$ resp. $([X\ \ \overline{X}],[Y\ \ \overline{Y}])$. Then, the third identity follows from Lemma 6.4.2. We prove the fourth. Since $R_\theta(X,Y)$ is a real matrix, it is positive semi-definite if and only if $\widetilde{\xi}^T R_\theta(X,Y)\widetilde{\xi} \geq 0$ for all *real* vectors $\widetilde{\xi}\in\mathbb{R}^{2k}$. Write $\widetilde{\xi}$ in the form $\widetilde{\xi} = [\xi_1^T,\xi_2^T]^T$ with $\xi_1,\xi_2\in\mathbb{R}^k$. Set $\xi := \xi_1 - i\xi_2$. Then a direct computation yields

$$\widetilde{\xi}^T R_\theta(X,Y)\widetilde{\xi} = \theta^2\|\Re(X\xi)\|_2^2 - \|\Re(Y\xi)\|_2^2.$$

This implies the last identity of the proposition. □

6.5 Proof of Theorems 6.1.4 and 6.1.7 and Proposition 6.1.9

We are now in the position to give the proof of Theorems 6.1.4 and and 6.1.7 and Proposition 6.1.9: From Proposition 6.4.3 it follows, that for all $(X,Y)\in\mathbb{C}^{l\times k}\times\mathbb{C}^{q\times k}$,

$$(\nu^{(2)}_{\mathbb{R},k}(X,Y))^{-1} = \left(\max_{\xi\in\mathbb{C}^k}\frac{\|\Re(Y\xi)\|_2}{\|\Re(X\xi)\|_2}\right)^{-1} = \min_{\xi\in\mathbb{C}^k}\frac{\|\Re(X\xi)\|_2}{\|\Re(Y\xi)\|_2}, \tag{6.37}$$

and

$$\begin{aligned}(\nu^{(2)}_{\mathbb{R},k}(X,Y))^{-1} &= \left(\min\left\{\theta\geq 0 \;\middle|\; \forall\,\xi\in\mathbb{C}^k:\ \xi^*(\theta^2X^*X - Y^*Y)\xi \geq |\,\xi^T(\theta^2X^TX - Y^TY)\xi\,|\ \right\}\right)^{-1}\\ &= \max\left\{\frac{1}{\theta}\;\middle|\;\theta\geq 0,\ \forall\,\xi\in\mathbb{C}^k:\ \xi^*(\theta^2X^*X - Y^*Y)\xi \geq |\,\xi^T(\theta^2X^TX - Y^TY)\xi\,|\right\}\\ &= \max\left\{\frac{1}{\theta}\;\middle|\;\theta\geq 0,\ \forall\,\xi\in\mathbb{C}^k:\ \xi^*(X^*X - \frac{1}{\theta^2}Y^*Y)\xi \geq |\,\xi^T(X^TX - \frac{1}{\theta^2}Y^TY)\xi\,|\right\}\\ &= \max\left\{t\geq 0\;\middle|\;\forall\,\xi\in\mathbb{C}^k:\ \xi^*(X^*X - t^2Y^*Y)\xi \geq |\,\xi^T(X^TX - t^2Y^TY)\xi\,|\right\}.\end{aligned} \tag{6.38}$$

Let $M\in\mathbb{C}^{q\times l}$, $X\in\mathbb{C}^{l\times k}$ and $\mathcal{X} = \mathrm{range}(X)$. On replacing in (6.37) and (6.38) the pair (X,Y) by the pair (MX,X) we obtain

$$(\nu^{(2)}_{\mathbb{R},k}(MX,X))^{-1} = \min_{x\in\mathcal{X}}\frac{\|\Re(Mx)\|_2}{\|\Re x\|_2} \tag{6.39}$$

$$= \max\left\{\ t\geq 0\ \middle|\ \forall\,x\in\mathcal{X}:\ x^*(M^*M - t^2I_l)x \geq |\,x^T(M^TM - t^2I_l)x\,|\ \right\}. \tag{6.40}$$

From Proposition 6.3.6 it follows that

$$(\nu^{(2)}_{\mathbb{R},k}(MX,X))^{-1} = \begin{cases}\|\,[\Re X\ \Im X]\,[\Re(MX)\ \Im(MX)]^\dagger\,\|_2^{-1} & \text{if } \ker[\Re(MX)\ \Im(MX)] \\ & \qquad\subseteq \ker[\Re X\ \Im X],\\ 0 & \text{otherwise.}\end{cases} \tag{6.41}$$

According to Proposition 6.2.3 we have $\tau_k(M) = \max_{X\in\mathcal{G}_{l,k}}(\nu^{(2)}_{\mathbb{R},k}(MX,X))^{-1}$. Combining the latter with (6.39) resp. (6.40) we obtain the identities

$$\tau_k(M) \;=\; \max_{\mathcal{X}\in\mathfrak{G}_{l,k}} \min_{x\in\mathcal{X}} \frac{\|\Re(Mx)\|_2}{\|\Re x\|_2}, \tag{6.42}$$

and

$$\begin{aligned}
\tau_k(M) &= \max_{\mathcal{X}\in\mathfrak{G}_{l,k}} \max\left\{\, t\geq 0 \;\middle|\; \forall\, x\in\mathcal{X}:\, x^*(M^*M-t^2I_l)x \geq |\,x^T(M^TM-t^2I_l)x\,|\,\right\}\\
&= \max\left\{\, t\geq 0 \;\middle|\; \exists\mathcal{X}\in\mathfrak{G}_{l,k}: \forall\, x\in\mathcal{X}:\, x^*(M^*M-t^2I_l)x \geq |\,x^T(M^TM-t^2I_l)x\,|\,\right\}\\
&= \max\{\, t\geq 0 \mid\, i_{\geq}(M^*M-t^2I_l\,,\, M^TM-t^2I_l)\geq k\,\}\\
&= \max\, T_k(M),
\end{aligned} \tag{6.43}$$

where $T_k(M) = \{\, t\geq 0 \mid\, i_{\geq}(M^*M-t^2I_l\,,\, M^TM-t^2I_l)\geq k\,\}$. Claim (i) of Proposition 6.1.9 then follows from (6.39), (6.40) and (6.41).

Analogously we obtain from Proposition 6.3.6 and 6.4.3 the relations

$$\nu^{(2)}_{\mathbb{R},k}(X,MX) \;=\; \max_{x\in\mathcal{X}} \frac{\|\Re(Mx)\|_2}{\|\Re x\|_2} \tag{6.44}$$

$$= \max\left\{\, t\geq 0 \;\middle|\; \forall\, x\in\mathcal{X}:\, x^*(t^2I_l-M^*M)x \geq |\,x^T(t^2I_l-M^TM)x\,|\,\right\} \tag{6.45}$$

$$= \begin{cases} \|\,[\Re(MX)\;\Im(MX)]\,[\Re X\;\Im X]^{\dagger}\,\|_2 & \text{if } \ker\,[\Re X\;\Im X] \subseteq \ker\,[\Re(MX)\;\Im(MX)],\\ \infty & \text{otherwise,}\end{cases} \tag{6.46}$$

where $k\in\underline{l}$, $X\in\mathbb{C}^{l\times(l-k+1)}$, $\mathcal{X} = \text{range}\, X$. By Proposition 6.2.3 we have

$$\widetilde{\tau}_k(M) = \min_{X\in\mathcal{G}_{l,l-k+1}} \nu^{(2)}_{\mathbb{R},k}(X,MX).$$

Combining this with (6.44),(6.45),(6.46) we obtain Claim (ii) of Proposition 6.1.9 and

$$\widetilde{\tau}_k(M) = \max_{\mathcal{X}\in\mathfrak{G}_{l,l-k+1}} \min_{x\in\mathcal{X}} \frac{\|\Re(Mx)\|_2}{\|\Re x\|_2} \qquad\text{and}\qquad \widetilde{\tau}_k(M) = \min\, \widetilde{T}_k(M),$$

where $\widetilde{T}_k(M) = \{\, t\geq 0 \mid\, i_{\geq}(t^2I_l-M^*M\,,\, t^2I_l-M^TM)\geq l-k+1\,\}$. In order to complete the proof of Theorem 6.1.7 it remains to show that the following statements hold for every $k\in\underline{l}$.

(a) $T_k(M)$ is an interval and $0\in T_k(M)$.

(b) If $\widetilde{T}_k(M)\neq 0$ then $\widetilde{T}_k(M)$ is an interval and $\sup \widetilde{T}_k(M) = \infty$.

Let us prove claim (a). Observe that by the Cauchy-Schwarz-inequality we have for any $y\in\mathbb{C}^n$,

$$y^*y = \|\overline{y}\|_2\|y\|_2 \geq |\overline{y}^*y| = |y^Ty| \tag{6.47}$$

This implies $x^*M^*Mx \geq |x^TM^TMx|$ for all $x\in\mathbb{C}^l$. Thus $0\in T_k(M)$ for all $k\in\underline{l}$. Assume now that $t_0\in T_k(M)$ for some $k\in\underline{l}$. Let $t\in[0,t_0]$ and let $\mathcal{X}$ be a k-dimensional subspace of $\mathbb{C}^l$ such that

$$x^*(M^*M-t_0^2I_l)x \geq |\,x^T(M^TM-t_0^2I_l)x\,| \qquad \text{for all } x\in\mathcal{X}.$$

By (6.47) we have

$$(t_0^2 - t^2)x^*x \geq (t_0^2 - t^2)|x^Tx| \qquad \text{for all } x\in\mathbb{C}^l.$$

Addition of the latter two inequalities yields that

$$\begin{aligned} x^*(M^*M - t^2I_l)x &\geq |\, x^T(M^TM - t_0^2I_l)x\,| + (t_0^2 - t^2)|x^Tx| \\ &\geq |\, x^T(M^TM - t_0^2I_l)x + (t_0^2 - t^2)x^Tx\,| \\ &= |\, x^T(M^TM - t^2I_l)x\,| \qquad\qquad \text{for all } x\in\mathcal{X}. \end{aligned}$$

Thus $t\in T_k(M)$, and the proof of (a) is complete.
(b). Analogously one shows that $t_0\in\widetilde{T}_k(M)$ implies that $t\in\widetilde{T}_k(M)$ for any $t > t_0$. The latter implication is equivalent to (b). The details are left to the reader.

6.6 The function $\nu_{\mathbb{R}}^{\|\cdot\|}(x,y)$

In Section 6.3 we dealt with $\nu_{\mathbb{R},k}^{\|\cdot\|}(X,Y)$ for unitarily invariant norms and arbitrary k. We now discuss the case $k=1$ in detail. Throughout this section,

$$\nu_{\mathbb{F}}^{\|\cdot\|}(x,y) := \nu_{\mathbb{F},1}^{\|\cdot\|}(x,y) = \min\{\|\Delta\| \;\mid\; \Delta\in\mathbb{F}^{q\times l},\ \Delta x = y\,\}, \qquad (x,y)\in\mathbb{C}^l\times\mathbb{C}^q.$$

According to Proposition 6.2.6 we have for any norm on $\mathbb{C}^{l\times q}$ satisfying $\|\Delta\| = \|\overline{\Delta}\|$ that

$$\nu_{\mathbb{R}}^{\|\cdot\|}(x,y) = \nu_{\mathbb{C},2}^{\|\cdot\|}(\,[x\ \overline{x}], [y\ \overline{y}]) = \nu_{\mathbb{C},2}^{\|\cdot\|}(\,[\Re x\ \Im x], [\Re y\ \Im y]).$$

By Proposition 6.2.3 we have for any $M\in\mathbb{C}^{q\times l}$ and any norm on $\mathbb{R}^{l\times q}$,

$$\mu_{\mathbb{R}}^{\|\cdot\|}(M) \;=\; \left(\min_{x\in\mathbb{C}^l\backslash\{0\}} \nu_{\mathbb{R}}^{\|\cdot\|}(Mx,x)\right)^{-1}, \tag{6.48}$$

and, if $q = l =: n$,

$$\widetilde{\mu}_{\mathbb{R}}^{\|\cdot\|}(M) \;=\; \min_{x\in\mathbb{C}^n\backslash\{0\}} \nu_{\mathbb{R}}^{\|\cdot\|}(x,Mx), \tag{6.49}$$

where $\mu_{\mathbb{R}}^{\|\cdot\|}(\cdot)$ and $\widetilde{\mu}_{\mathbb{R}}^{\|\cdot\|}(M)$ are the real μ-functions defined in (6.1) resp.(6.2). In this section we consider only unitarily invariant norms. Our aim is to give explicit formulas for $\nu_{\mathbb{R}}^{\|\cdot\|}(x,y)$ and the matrix $\Delta_0\in\mathbb{R}^{l\times q}$ of minimal norm satisfying $\Delta_0x = y$. We start with a lemma on the Gramians

$$[x\ \overline{x}]^*[x\ \overline{x}] = \begin{bmatrix} \|x\|_2^2 & \overline{x^Tx} \\ x^Tx & \|x\|_2^2 \end{bmatrix}, \qquad [\Re x\ \Im x]^T[\Re x\ \Im x] = \begin{bmatrix} \|\Re x\|_2^2 & (\Re x)^T(\Im x) \\ (\Im x)^T(\Re x) & \|\Im x\|_2^2 \end{bmatrix}.$$

Note that $[\Re x\ \Im x]^T[\Re\ \Im x] = W_1^*[x\ \overline{x}]^*[x\ \overline{x}]W_1$ where W_1 is the matrix defined in (6.30).

Lemma 6.6.1 *Let $x\in\mathbb{C}^l\backslash\{0\}$. Then the following conditions are equivalent.*

(i) The vectors x and $\overline{x}$ are linearly dependent.

(ii) The vectors $\Re x$ and $\Im x$ are linearly dependent over $\mathbb{R}$.

(iii) There exists $z\in\mathbb{C}\setminus\{0\}$ such that $zx\in\mathbb{R}^l$.

(*vi*) $\|x\|_2^2 = |x^T x|$.

(*v*) $\operatorname{rank}\left(\begin{bmatrix} \|x\|_2^2 & \overline{x^T x} \\ x^T x & \|x\|_2^2 \end{bmatrix}\right) = 1$.

(*vi*) $\operatorname{rank}\begin{bmatrix} \|\Re x\|_2^2 & (\Re x)^T(\Im x) \\ (\Im x)^T(\Re x) & \|\Im x\|_2^2 \end{bmatrix} = 1$.

Proof: We have

$$\begin{aligned} \det\left(\begin{bmatrix} \|\Re x\|_2^2 & (\Re x)^T(\Im x) \\ (\Re x)^T(\Im x) & \|\Im x\|_2^2 \end{bmatrix}\right) &= \|\Re x\|_2^2\|\Im x\|_2^2 - |(\Re x)^T(\Im x)|^2, \\ \det\left(\begin{bmatrix} \|x\|_2^2 & \overline{x^T x} \\ x^T x & \|x\|_2^2 \end{bmatrix}\right) &= \|x\|_2^4 - |x^T x|^2 = \|x\|_2^2\|\overline{x}\|_2^2 - |\overline{x}^* x|^4. \end{aligned}$$

Thus the equivalences $(i) \Leftrightarrow (iv) \Leftrightarrow (v)$ and $(ii) \Leftrightarrow (vi)$ follow from the Cauchy-Schwarz inequality.

$(i) \Leftrightarrow (iii)$. Suppose x and $\overline{x}$ are linearly dependent. Then there is a $z \in \mathbb{C}$ such that $z^2 x = \overline{x}$ and $|z| = 1$. This implies $zx = z^{-1}\overline{x} = \overline{z}\,\overline{x} = \overline{zx}$. Thus $zx \in \mathbb{R}^l$. On the other hand, if $zx \in \mathbb{R}^l$, $z \neq 0$, then $zx - \overline{zx} = 0$. Thus x and $\overline{x}$ are linearly dependent. $(ii) \Leftrightarrow (iii)$. Let $\lambda_1, \lambda_2 \in \mathbb{R}$ and $z = \lambda_1 + i\lambda_2$. Then

$$\Im(zx) = \lambda_2(\Re x) + \lambda_1(\Im x).$$

The equivalence $(ii) \Leftrightarrow (iii)$ follows from this relation. □

In the proposition below we treat $\nu_{\mathbb{R}}^{\|\cdot\|}(x, y)$ for the case when x and $\overline{x}$ are linearly dependent.

Proposition 6.6.2 *Let $(x, y) \in (\mathbb{C}^l \setminus \{0\}) \times \mathbb{C}^q$. Set $\Delta_0 := \|x\|_2^{-2} y x^*$. Then $\Delta_0 x = y$. Suppose that x and $\overline{x}$ are linearly dependent. Then the following assertions are equivalent.*

(*i*) *There is a $z \in \mathbb{C}\setminus\{0\}$ such that $zx \in \mathbb{R}^l$ and $zy \in \mathbb{R}^q$.*

(*ii*) *The matrix Δ_0 is real.*

(*iii*) *There is a real matrix $\Delta \in \mathbb{R}^{q\times l}$ such that $\Delta x = y$.*

(*iv*) *We have $\Delta_0 = [y\ \overline{y}][x\ \overline{x}]^\dagger$.*

(*v*) *We have* $\begin{bmatrix} \|y\|_2^2 & \overline{y^T y} \\ y^T y & \|y\|_2^2 \end{bmatrix} = t^2 \begin{bmatrix} \|x\|_2^2 & \overline{x^T x} \\ x^T x & \|x\|_2^2 \end{bmatrix}$, *where $t = \|y\|_2/\|x\|_2$.*

(*vi*) *For any norm on $\mathbb{R}^{q\times l}$: $\nu_{\mathbb{R}}^{\|\cdot\|}(x, y) \neq \infty$.*

(*vii*) *For any normalized unitarily invariant norm: $\nu_{\mathbb{R}}^{\|\cdot\|}(x, y) = \|\Delta_0\|_2 = \frac{\|y\|_2}{\|x\|_2}$.*

Proof: By Lemma 6.6.1 the linear dependence of x and $\overline{x}$ implies that $\operatorname{rank}\left(\begin{bmatrix} \|x\|_2^2 & \overline{x^T x} \\ x^T x & \|x\|_2^2 \end{bmatrix}\right) = 1$ and $z\,x \in \mathbb{R}^l$ for some $z \in \mathbb{C} \setminus \{0\}$. The latter yields the implications $(iii) \Rightarrow (i) \Rightarrow (ii) \Rightarrow (iii)$.

The equivalence $(iii) \Leftrightarrow (vi)$ and the implication $(i) \Rightarrow (v)$ are trivial. If $y = 0$ then the implication $(v) \Rightarrow (i)$ is trivial, as well. Suppose now, that $y \neq 0$ and (v) holds. Then

$$\operatorname{rank}\left(\begin{bmatrix} \|y\|_2^2 & \overline{y^T y} \\ y^T y & \|y\|_2^2 \end{bmatrix}\right) = \operatorname{rank}\left(\begin{bmatrix} \|x\|_2^2 & \overline{x^T x} \\ x^T x & \|x\|_2^2 \end{bmatrix}\right) = 1.$$

Thus, by Lemma 6.6.1, $wy \in \mathbb{R}^q$ for some $w \in \mathbb{C} \setminus \{0\}$. Moreover, (v) implies $\|x\|_2^2\, y^T y = \|y\|_2^2\, x^T x$. Hence, $\|x\|_2^2 (w\,y)^T (w\,y) = \|y\|_2^2 (zx)^T (zx) \left(\frac{w}{z}\right)^2$ and therefore $\left(\frac{w}{z}\right)^2 > 0$. Thus $\frac{z}{w} \in \mathbb{R}$, and $z\,y = \frac{z}{w} wy \in \mathbb{R}$. Thus (v) implies (i). The first of the implications $(iv) \Rightarrow (vii) \Rightarrow (vi) \Rightarrow (i)$ follows from Proposition 6.3.6, the second is trivial and the third has already been shown. It remains to prove that (i) implies (iv). However, if $z\,x \neq 0$ is a real vector then $\ker\,[x\;\overline{x}] = \mathbb{C}\begin{bmatrix} z \\ -\overline{z} \end{bmatrix}$ and $(\ker\,[x\;\overline{x}])^{\perp} = \mathbb{C}\begin{bmatrix} z \\ \overline{z} \end{bmatrix}$. Moreover, $\operatorname{range}\,[x\;\overline{x}] = \mathbb{C}(z\,x)$. The matrix $P := \frac{1}{2}\begin{bmatrix} z \\ \overline{z} \end{bmatrix} \frac{(z\,x)^*}{\|z\,x\|_2^2}$ satisfies $P\,([x\;\overline{x}]u) = u$ for all $u \in (\ker\,[x\;\overline{x}])^{\perp}$ and $Pv = 0$ for all $v \in (\operatorname{range}\,[x\;\overline{x}])^{\perp}$. Thus $P = [x\;\overline{x}]^{\dagger}$. If the vector $z\,y$ is real then $[y\;\overline{y}]P = \Delta_0$. Thus (i) implies (iv). The proof is complete. $\square$

In the next proposition we give explicit formulas for $\nu_{\mathbb{R}}^{\|\cdot\|}(x,y)$ when x and $\overline{x}$ are linearly independent.

Proposition 6.6.3 *Let $\|\cdot\|$ be a unitarily invariant norm on $\mathbb{C}^{l\times q}$ and let be $\|\!|\cdot|\!\|$ the absolute and symmetric norm on $\mathbb{R}^{\min\{l,q\}}$ such that $\|\Delta\| = \|\!|\ [\sigma_1(\Delta),\ldots,\sigma_{\min\{l,q\}}(\Delta)]^T\ |\!\|$ for all $\Delta \in \mathbb{C}^{l\times q}$. Let $(x,y) \in (\mathbb{C}^l \setminus \{0\}) \times \mathbb{C}^q$. Suppose the vectors x and $\overline{x}$ are linearly independent. Let $\Delta_0 = [y\;\overline{y}][x\;\overline{x}]^{\dagger}$. Then*

(i) *The matrix Δ_0 satisfies $\Delta_0 x = y$. Moreover, Δ_0 has the following representations.*

$$\begin{aligned}
\Delta_0 &= [y\;\overline{y}] \begin{bmatrix} \|x\|_2^2 & \overline{x^T x} \\ x^T x & \|x\|_2^2 \end{bmatrix}^{-1} \begin{bmatrix} x^* \\ x^T \end{bmatrix} \\
&= (\|x\|_2^4 - |x^T x|^2)^{-1}\ [y\;\overline{y}] \begin{bmatrix} \|x\|_2^2 & -\overline{x^T x} \\ -x^T x & \|x\|_2^2 \end{bmatrix} \begin{bmatrix} x^* \\ x^T \end{bmatrix} \\
&= [\Re y\ \Im y][\Re x\ \Im x]^{\dagger} \\
&= [\Re y\ \Im y] \begin{bmatrix} \|\Re x\|_2^2 & (\Re x)^T (\Im x) \\ (\Re x)^T (\Im x) & \|\Im x\|_2^2 \end{bmatrix}^{-1} \begin{bmatrix} (\Re x)^T \\ (\Im x)^T \end{bmatrix} \\
&= h^{-1}\ [\Re y\ \Im y] \begin{bmatrix} \|\Im x\|_2^2 & -(\Re x)^T (\Im x) \\ -(\Re x)^T (\Im x) & \|\Re x\|_2^2 \end{bmatrix} \begin{bmatrix} (\Re x)^T \\ (\Im x)^T \end{bmatrix},
\end{aligned}$$

where $h = \|\Re x\|_2^2 \|\Im x\|_2^2 - ((\Re x)^T (\Im x))^2$.

(*ii*) *We have* $\nu_{\mathbb{R}}^{\|\cdot\|}(x,y) = \|\Delta_0\| = \|\!|\ [\sigma_1(\Delta_0), \sigma_2(\Delta_0), 0\ldots,0]^T\ \|\!|$ *and*

$$\begin{aligned}
\sigma_j(\Delta_0)^2 &= \lambda_j\left(\begin{bmatrix}\|x\|_2^2 & \overline{x^Tx}\\ x^Tx & \|x\|_2^2\end{bmatrix}^{-1}\begin{bmatrix}\|y\|_2^2 & \overline{y^Ty}\\ y^Ty & \|y\|_2^2\end{bmatrix}\right) \qquad\qquad j=1,2\\
&= \frac{\|x\|_2^2\|y\|_2^2 - 2\Re(x^Tx\ \overline{y^Ty}) + (-1)^{j+1}d_1}{\|x\|_2^4 - |x^Tx|^2}\\
&= \lambda_j\left(\begin{bmatrix}\|\Re x\|_2^2 & (\Re x)^T(\Im x)\\ (\Re x)^T(\Im x) & \|\Im x\|_2^2\end{bmatrix}^{-1}\begin{bmatrix}\|\Re y\|_2^2 & (\Re y)^T(\Im y)\\ (\Re y)^T(\Im x) & \|\Im y\|_2^2\end{bmatrix}\right)\\
&= \frac{(1/2)(\|\Im x\|_2^2\|\Re y\|_2^2 + \|\Im y\|_2^2\|\Re x\|_2^2) - ((\Re x)^T(\Im x))^2 + (-1)^{j+1}d_2}{\|\Re x\|_2^2\|\Im x\|_2^2 - ((\Re x)^T(\Im x))^2},
\end{aligned}$$

where $\lambda_1(\cdot)$, $\lambda_2(\cdot)$ *are the eigenvalues of the matrix in question and*

$$\begin{aligned}
d_1 &:= (\ \|x\|_2^2\|y\|_2^2 - \Re(x^Tx\ \overline{y^Ty})\ )^2 - (\|x\|_2^4 - |x^Tx|^2)(\|y\|_2^4 - |y^Ty|^2)\\
d_2 &:= (\ (1/2)(\|\Im x\|_2^2\|\Re y\|_2^2 + \|\Im y\|_2^2\|\Re x\|_2^2) - ((\Re x)^T(\Im x))^2\ )^2\\
&\quad -(\|\Im x\|_2^2\|\Re x\|_2^2 - ((\Re x)^T(\Im x))^2)(\|\Im y\|_2^2\|\Re y\|_2^2 - ((\Re y)^T(\Im y))^2).
\end{aligned}$$

(*iii*) *Assume that*

$$\begin{bmatrix}\|y\|_2^2 & \overline{y^Ty}\\ y^Ty & \|y\|_2^2\end{bmatrix} = t^2\begin{bmatrix}\|x\|_2^2 & \overline{x^Tx}\\ x^Tx & \|x\|_2^2\end{bmatrix}$$

for some $t \geq 0$. *Then* $\sigma_1(\Delta_0) = \sigma_2(\Delta_0) = t$ *and* $\nu_{\mathbb{R}}^{\|\cdot\|}(x,y) = \|\Delta_0\| = t\|\!|[1,1,0,\ldots,0]^T\|\!|$.

Proof: *This follows by applying Corollaries 6.3.4 and 6.3.5 to the matrix pairs* $(X,Y) = (\,[x\ \overline{x}],\ [y\ \overline{y}]\,)$ *resp.* $(X,Y) = (\,[\Re x\ \Im x],\ [\Re y\ \Im y]\,)$. □

6.7 A lower bound for $\mu_{\mathbb{R}}^{\|\cdot\|}(M)$

The aim of this section is to establish the following Proposition which gives an easily computable lower bound for $\mu_{\mathbb{R}}^{\|\cdot\|}(M)$ provided that the underlying norm is unitarily invariant.

Proposition 6.7.1 *Let* $\|\cdot\|$ *be a unitarily invariant norm on* $\mathbb{C}^{l\times q}$, $\min\{l,q\} \geq 2$, *and let* $\|\!|\cdot\|\!|$ *be the absolute and symmetric norm on* $\mathbb{R}^{\min\{l,q\}}$ *such that* $\|\Delta\| = \|\!|\ [\sigma_1(\Delta),\ldots,\sigma_{\min\{l,q\}}(\Delta)]^T\ \|\!|$ *for all* $\Delta\in\mathbb{C}^{l\times q}$. *Then for any* $M\in\mathbb{C}^{q\times l}$,

$$\mu_{\mathbb{R}}^{\|\cdot\|}(M) \geq \sigma_2(M)/\|\!|[1,1,0,\ldots,0]^T\|\!|.$$

In the corollary below $\mu_{\mathbb{R}}^F(\cdot)$ denotes the real μ-function with respect to Frobenius norm, $\|\cdot\|_F$.

Corollary 6.7.2 *For every* $M\in\mathbb{C}^{q\times l}$ *we have* $\tau_1(M) \geq \sigma_2(M)$ *and* $\mu_{\mathbb{R}}^F(M) \geq \sigma_2(M)/\sqrt{2}$.

Proof: This follows from Proposition 6.7.1 and the fact that

$$\|\Delta\|_2 = \|[\,\sigma_1(\Delta),\ldots,\sigma_{\min\{l,q\}}(\Delta)\,]^T\|_\infty, \qquad \|\Delta\|_F = \|[\,\sigma_1(\Delta),\ldots,\sigma_{\min\{l,q\}}(\Delta)\,]^T\|_2. \qquad □$$

The first statement of the corollary can be generalized: One always has $\tau_k(M) \geq \sigma_{2k}(M)$. The proof is given in the appendix. In order to show Proposition 6.7.1 we need the following Lemma.

Lemma 6.7.3 *Let $M\in\mathbb{C}^{q\times l}$ and let $\mathcal{X}$ be a 2-dimensional subspace of $\mathbb{C}^l$. Let $a=\min_{x\in\mathcal{X}\setminus\{0\}}\frac{\|Mx\|_2}{\|x\|_2}$ and $b=\max_{x\in\mathcal{X}\setminus\{0\}}\frac{\|Mx\|_2}{\|x\|_2}$. Then there exists a $t_0\in[a,b]$ and an $x_0\in\mathcal{X}\setminus\{0\}$ such that*

$$\begin{bmatrix}\|Mx_0\|_2^2 & \overline{(Mx_0)^T(Mx_0)}\\ (Mx_0)^T(Mx_0) & \|Mx_0\|_2^2\end{bmatrix}=t_0^2\begin{bmatrix}\|x_0\|_2^2 & \overline{x_0^Tx_0}\\ x_0^Tx_0 & \|x_0\|_2^2\end{bmatrix}. \tag{6.50}$$

Proof: Let v_1,v_2 be a basis of $\mathcal{X}$. Consider the quadratic forms

$$\begin{aligned}q_t(z_1,z_2) &:= (z_1v_1+z_2v_2)^T(M^TM-t^2I_l)(z_1v_1+z_2v_2) \qquad z_1,z_2\in\mathbb{C},\ \ t\geq 0\\ &= \alpha(t)z_1^2+2\beta(t)z_1z_2+\gamma(t)z_2^2,\end{aligned}$$

where

$$\alpha(t)=v_1^T(M^TM-t^2I_l)v_1,\quad \beta(t)=v_1^T(M^TM-t^2I_l)v_2,\quad \gamma(t)=v_2^T(M^TM-t^2I_l)v_2.$$

Let $x_t:=\gamma(t)v_1-(\beta(t)+s(t))v_2$, where $s(t)$ denotes the square root of $\beta(t)^2-\alpha(t)\gamma(t)$ with nonnegative imaginary part. Then x_t is a continuous function of $t\geq 0$ and we have

$$(Mx_t)^T(Mx_t)-t^2x_t^Tx_t=x_t^T(M^TM-t^2I_l)x_t=q_t(\gamma(t),-(\beta(t)+s(t))\,)=0. \tag{6.51}$$

Assume that $\alpha(t)^2+\beta(t)^2+\gamma(t)^2\neq 0$ for all $t\in[a,b]$. This implies $x_t\neq 0$ for all $t\in[a,b]$. Thus, the function

$$f:[a,b]\to\mathbb{R},\qquad f(t):=\frac{\|Mx_t\|_2}{\|x_t\|_2}-t$$

is well defined. Moreover, f is continuous and satisfies $f(a)\geq 0$, $f(b)\leq 0$. Hence, there is a $t_0\in[a,b]$ such that $f(t_0)=0$. Set $x_0:=x_{t_0}$. Then we have $t_0=\|Mx_0\|_2/\|x_0\|_2$ and, by (6.51), $(Mx_0)^T(Mx_0)=t_0^2\,x_0^Tx_0$. The latter two relations are equivalent to (6.50).

Suppose now, there is a $t_0\in[a,b]$ such that $\alpha(t_0)^2+\beta(t_0)^2+\gamma(t_0)^2=0$. Then $q_{t_0}\equiv 0$. By continuity of the function $x\mapsto\|Mx\|_2/\|x\|_2$, $x\in\mathcal{X}\setminus\{0\}$ there is an $x_0=z_{01}\,v_1+z_{02}\,v_2\in\mathcal{X}\setminus\{0\}$ satisfying $\|Mx_0\|_2/\|x_0\|_2=t_0$. We have $(Mx_0)^T(Mx_0)-t_0^2\,x_0^Tx_0=q_{t_0}(z_{01},z_{02})=0$. Thus t_0 and x_0 satisfy (6.50). $\square$

Proof of Proposition 6.7.1: Let $v_1,v_2\in\mathbb{C}^l\setminus\{0\}$ be singular vectors to the largest singular values of M, i.e. $(M^*M)v_k=\sigma_k(M)^2v_k$, $k=1,2$ and $v_1^*v_2=0$. Then we have for the 2-dimensional subspace $\mathcal{X}:=\mathrm{span}\{v_1,v_2\}$ that

$$\min_{x\in\mathcal{X}\setminus\{0\}}\frac{\|Mx\|_2}{\|x\|_2}=\sigma_2(M)\quad\text{ and }\quad\max_{x\in\mathcal{X}\setminus\{0\}}\frac{\|Mx\|_2}{\|x\|_2}=\sigma_1(M).$$

By Lemma 6.7.3 there exist $x_0\in\mathcal{X}\setminus\{0\}$ and $t_0\in[\sigma_2(M),\sigma_1(M)]$ such that (6.50) holds. By Proposition 6.6.2 and claim (*iii*) of Proposition 6.6.3 the matrix $\Delta_0:=[x\ \overline{x}]\,[Mx\ \overline{Mx}]^\dagger$ satisfies

$$\sigma_1(\Delta_0)=1/t_0\qquad\sigma_2(\Delta_0)=\begin{cases}1/t_0 & \text{if } Mx \text{ and } \overline{Mx} \text{ are linearly independent,}\\ 0 & \text{otherwise.}\end{cases}$$

Thus,

$$1/\mu_{\mathbb{R}}^{\|\cdot\|}(M)\leq\nu_{\mathbb{R}}^{\|\cdot\|}(Mx_0,x_0)=\|\Delta_0\|\leq\tfrac{1}{t_0}\,|\!|\!|[1,1,0,\ldots,0]^T|\!|\!|\leq\tfrac{1}{\sigma_2(M)}\,|\!|\!|[1,1,0,\ldots,0]^T|\!|\!| \qquad\square$$

6.8 Equality of $\mu_{\mathbb{R}}^{\|\cdot\|}(M)$ and $\mu_{\mathbb{C}}^{\|\cdot\|}(M)$.

In this section we address the question in which cases the equality $\mu_{\mathbb{R}}^{\|\cdot\|}(M) = \mu_{\mathbb{C}}^{\|\cdot\|}(M)$ holds. We are motivated by the following fact.

Proposition 6.8.1 *Let $\|\cdot\|$ be an arbitrary norm on $\mathbb{C}^{l\times q}$. Let $(A,B,C) \in L_{n,l,q}$, $G(s) = C(sI_n - A)^{-1}B$ and $\mathcal{E}_G := \{\ s\in\mathbb{C}\backslash\sigma(A) \mid\ \mu_{\mathbb{R}}^{\|\cdot\|}(G(s)) = \mu_{\mathbb{C}}^{\|\cdot\|}(G(s))\ \}$. Then, for all $\rho \geq 0$,*

$$\sigma_{\mathbb{C}}^{\|\cdot\|}(A,B,C,\rho) \cap \mathcal{E}_G \subseteq \sigma_{\mathbb{R}}^{\|\cdot\|}(A,B,C,\rho).$$

Proof: Obvious. □

Our first result is for the case that M resp. $G(s)$ is a real matrix.

Proposition 6.8.2 *Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be norms on $\mathbb{C}^q$ resp. $\mathbb{C}^l$, and let $\|\cdot\|$ be a norm on $\mathbb{C}^{l\times q}$ which is rank-1-compatible with the operator norm $\|\cdot\|_{\alpha,\beta}$. Suppose that one of the following conditions holds.*

- (a) $l = 1$.
- (b) *The norm $\|\cdot\|_\beta$ is absolute and $q = 1$.*
- (c) *The norm $\|\cdot\|_\beta$ is absolute and $\|\cdot\|_\alpha$ is a weighted maximum norm, i.e. there are weights $r_1,\dots,r_m > 0$ such that $\|y\|_\alpha = \max_{j\in\underline{m}} r_j|y_j|$ for all $y\in\mathbb{C}^q$.*
- (d) $\alpha = \beta =: p\in[1,\infty]$.

Then

- (i) *For every $M\in\mathbb{R}^{q\times l}$: $\mu_{\mathbb{R}}^{\|\cdot\|}(M) = \mu_{\mathbb{C}}^{\|\cdot\|}(M) = \|M\|_{\beta,\alpha}$.*
- (ii) *For every $(A,B,C)\in L_{n,l,q}$ and every $\rho \geq 0$,*

$$\sigma_{\mathbb{C}}^{\|\cdot\|}(A,B,C,\rho) \cap \Re_G \subseteq \sigma_{\mathbb{R}}^{\|\cdot\|}(A,B,C,\rho),$$

where

$$\Re_G := \{\ s\in\mathbb{C}\backslash\sigma(A) \mid\ G(s)\in\mathbb{R}^{q\times l}\ \}$$

is the realness locus of $G(s) = C(sI_n - A)^{-1}B$.

Proof: (i) By Theorem 2.5.2 we have for a real matrix M, $\mu_{\mathbb{F}}^{\|\cdot\|}(M) = \max_{x\in\mathbb{F}^l} \frac{\|Mx\|_\alpha}{\|x\|_\beta}$, $\mathbb{F}\in\{\mathbb{R},\mathbb{C}\}$. But, by Proposition 1.6.1 the two maxima coincide if one of the conditions $(a)-(d)$ is satisfied. Claim (ii) is a consequence of (i) and Proposition 6.8.1. □

Note that the realness locus $\Re_G$ is the intersection of the sets $\Re_{g_{jk}} = \{\ s\in\mathbb{C}\backslash\sigma(A) \mid\ \Im(g_{jk}(s)) = 0\ \}$, where $g_{jk}(s)$ is the entry in the jth row and kth column of $G(s)$. The sets $\Re_{g_{jk}}$ are easily seen to be real algebraic curves (with finitely many points removed). If the matrices A, B, C are real then $\Re_G$ contains the real axis.

Next we consider the case when the underlying norm is the spectral norm and the matrix M is not necessarily real. The Proposition below is a refinement of a result by Lewkowicz [71].

Proposition 6.8.3 *For any $M\in\mathbb{C}^{q\times l}$ the following holds.*

(*i*) *If $\sigma_1(M)=\sigma_2(M)$ then $\tau_1(M)=\sigma_1(M)$.*

(*ii*) *Suppose $\sigma_1(M)\neq\sigma_2(M)$ and let $(u,v)\in\mathbb{C}^q\times\mathbb{C}^l$ be a pair of left and right singular vectors of M to the largest singular value (i.e. $Mv=\sigma_1(M)u$, $u^*M=\sigma_1(M)v^*$, $\|v\|_2=\|u\|_2=1$). Then $\tau_1(M)=\sigma_1(M)$ if and only if $u^Tu=v^Tv$.*

Assume that the latter condition is fulfilled. Then the real matrix $\Delta_0:=[Mv\ \overline{Mv}]^\dagger[v\ \overline{v}]=\sigma_1(M)^{-1}[u\ \overline{u}]^\dagger[v\ \overline{v}]$ satisfies $\Delta_0 Mv=v$. The singular values of Δ_0 are $\sigma_1(\Delta_0)=\sigma_1(M)^{-1}$ and $\sigma_2(\Delta_0)=\begin{cases}\sigma_1(M)^{-1} & \text{if } u \text{ and } \overline{u} \text{ are linearly independent}\\ 0 & \text{otherwise}\end{cases}$

Proof: Claim (*i*) follows from Corollary 6.7.2 and the trivial fact that $\sigma_1(M)\geq\tau_1(M)$.
(*ii*). By Corollary 6.1.8 we have $\tau_1(M)\geq\sigma_1(M)$ if and only if there is an $x\neq 0$ such that

$$x^*(M^*M-\sigma_1(M)^2I_l)x\geq|x^T(M^TM-\sigma_1(M)^2I_l)x| \tag{6.52}$$

The hermitian matrix $M^*M-\sigma_1(M)^2I_l$ is negative semi-definite and we have $x^*(M^*M-\sigma_1(M)^2I_l)x=0$ if and only if $M^*Mx=\sigma_1(M)^2x$. The latter implies that $x=zv$ for some $z\in\mathbb{C}\setminus\{0\}$, since $\sigma_1(M)>\sigma_2(M)$. Thus (6.52) holds if and only if there is a $z\in\mathbb{C}\setminus\{0\}$ such that $x=zv$ and $0=x^T(M^TM-\sigma_1(M)^2I_l)x=z^2\sigma_1(M)^2(u^Tu-v^Tv)$. This proves the first statement of Claim (*ii*). Suppose now that $u^Tu=v^Tv$. Then $(Mx)^T(Mx)=\sigma_1(M)^2\,x^Tx$, and therefore

$$\begin{bmatrix}\|Mv\|_2^2 & \overline{(Mv)^T(Mv)}\\ (Mv)^T(Mv) & \|Mv\|_2^2\end{bmatrix}=\sigma_1(M)^2\begin{bmatrix}\|v\|_2^2 & \overline{v^Tv}\\ v^Tv & \|v\|_2^2\end{bmatrix}.$$

Thus, the second statement of Claim (*ii*) follows from Proposition 6.6.2 and Claim (*iii*) of Proposition 6.6.3. □

We now give an application of Proposition 6.8.3 to spectral value sets. For $l,q\in\mathbb{N}$ let $\mathcal{M}_{l,q}$ denote the set of matrices of the form $\begin{bmatrix}M_1 & -M_2\\ M_2 & M_1\end{bmatrix}$, where $M_1,M_2\in\mathbb{C}^{q\times l}$. The following facts are easily verified.

(*a*) $\mathcal{M}_{l,q}$ is a complex vector space.

(*b*) If $M\in\mathcal{M}_{l,q}$, $N\in\mathcal{M}_{p,l}$ then $MN\in\mathcal{M}_{p,q}$. In particular, $\mathcal{M}_{n,n}$ is an algebra for every $n\in\mathbb{N}$.

(*c*) If $M\in\mathcal{M}_{n,n}$ is invertible, then $M^{-1}\in\mathcal{M}_{n,n}$ [2]

(*d*) Let $M_1,M_2\in\mathbb{C}^{q\times l}$, $x_1,x_2\in\mathbb{C}^l$ and $y_1,y_2\in\mathbb{C}^q$. Then the following equivalence holds.

$$\begin{bmatrix}M_1 & -M_2\\ M_2 & M_1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}y_1\\ y_2\end{bmatrix}\quad\Leftrightarrow\quad\begin{bmatrix}M_1 & -M_2\\ M_2 & M_1\end{bmatrix}\begin{bmatrix}-x_2\\ x_1\end{bmatrix}=\begin{bmatrix}-y_2\\ y_1\end{bmatrix}.$$

[2] To see this, recall the following argument: Let $\lambda^n+\sum_{k=0}^{n-1}a_k\lambda^k$ be the characteristic polynomial of $M\in\mathcal{M}_{n,n}$. Then $a_0=(-1)^n\det(M)$. By the Cayley-Hamilton theorem, $M^n+\sum_{k=0}^{n-1}a_kM^k=0$. Hence, if M is invertible, then $-\frac{1}{a_0}\left(\sum_{k=1}^{n}a_kM^{k-1}\right)M=I$. Thus, the matrix $M^{-1}=-\frac{1}{a_0}\left(\sum_{k=1}^{n}a_kM^{k-1}\right)$, being a polynomial in M is an element of the algebra $\mathcal{M}_{n,n}$.

(e) Let $M_1, M_2 \in \mathbb{C}^{q\times l}$, $x_1, x_2 \in \mathbb{C}^l$ and $y_1, y_2 \in \mathbb{C}^q$. Suppose the vectors $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and $\begin{bmatrix} -x_2 \\ x_1 \end{bmatrix}$ are linearly dependent. Then $x_2 = zx_1$ with $z = i$ or $z = -i$. If, in addition, $\begin{bmatrix} M_1 & -M_2 \\ M_2 & M_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$ then $y_2 = zy_1$, as well.

Proposition 6.8.4 *If $M \in \mathcal{M}_{l,q}$ then $\tau_1(M) = \sigma_1(M)$.*

Proof: Let $v = [v_1^T \; v_2^T]^T \in \mathbb{C}^{2l}$ and $u = [u_1^T \; u_2^T]^T \in \mathbb{C}^{2q}$ be a pair of normed left and right singular vectors to the largest singular values of M, i.e.

$$\begin{bmatrix} M_1 & -M_2 \\ M_2 & M_1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \sigma_1(M) \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \qquad \left\| \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \right\|_2 = \left\| \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \right\|_2 = 1.$$

This implies

$$\begin{bmatrix} M_1 & -M_2 \\ M_2 & M_1 \end{bmatrix} \begin{bmatrix} -v_2 \\ v_1 \end{bmatrix} = \sigma_1(M) \begin{bmatrix} -u_2 \\ u_1 \end{bmatrix}, \qquad \left\| \begin{bmatrix} -v_2 \\ v_1 \end{bmatrix} \right\|_2 = \left\| \begin{bmatrix} -u_2 \\ u_1 \end{bmatrix} \right\|_2 = 1.$$

Thus, the vectors $\widetilde{v} = [-v_2^T \; v_1^T]^T$ and $\widetilde{u} = [-u_2^T \; u_1^T]^T$ also form a pair of normed left and right singular vectors to the largest singular value of M. If v and $\widetilde{v}$ are linearly independent, then $\sigma_1(M) = \sigma_2(M)$, and therefore $\tau_1(M) = \sigma_1(M)$. If v and $\widetilde{v}$ are linearly dependent then, by the remark (e) above, $v = \begin{bmatrix} v_1 \\ zv_1 \end{bmatrix}$ and $u = \begin{bmatrix} u_1 \\ zu_1 \end{bmatrix}$ for $z = i$ or $z = -i$. Consequently, $v^T v = u^T u = 0$. Thus $\tau_1(M) = \sigma_1(M)$. □

Corollary 6.8.5 *Let $(A_k, B_k, C_k) \in L_{n,l,q}$, $k = 1, 2$, and let $\rho \geq 0$. Then*

$$\sigma_{\mathbb{R}}^{(2)}\left(\begin{bmatrix} A_1 & -A_2 \\ A_2 & A_1 \end{bmatrix}, \begin{bmatrix} B_1 & -B_2 \\ B_2 & B_1 \end{bmatrix}, \begin{bmatrix} C_1 & -C_2 \\ C_2 & C_1 \end{bmatrix}, \rho \right) = \sigma_{\mathbb{C}}^{(2)}\left(\begin{bmatrix} A_1 & -A_2 \\ A_2 & A_1 \end{bmatrix}, \begin{bmatrix} B_1 & -B_2 \\ B_2 & B_1 \end{bmatrix}, \begin{bmatrix} C_1 & -C_2 \\ C_2 & C_1 \end{bmatrix}, \rho \right).$$

Proof: From the remarks (a), (b), (c) above it follows that the rational matrix function

$$G(s) = \begin{bmatrix} C_1 & -C_2 \\ C_2 & C_1 \end{bmatrix} \left(sI - \begin{bmatrix} A_1 & -A_2 \\ A_2 & A_1 \end{bmatrix} \right)^{-1} \begin{bmatrix} B_1 & -B_2 \\ B_2 & B_1 \end{bmatrix}$$

satisfies $G(s) \in \mathcal{M}_{l,q}$. Hence, $\sigma_1(G(s)) \equiv \tau_1(G(s))$. □

6.9 Continuity of real μ-functions

In [9] the following theorem has been shown for the real perturbation values.

Theorem 6.9.1 *Let $M_0 \in \mathbb{C}^{q\times l}$. If $\operatorname{rank}(\Im M_0) \geq 2k - 1$ then the kth real perturbation value of first kind, $\tau_k : \mathbb{C}^{q\times l} \to \mathbb{R}$, is continuous at M_0.*

In particular Theorem 6.9.1 states that if $M_0 \in \mathbb{C}^{q\times l}$ is nonreal then the real μ-function $\mu_{\mathbb{R}}^{(2)} = \tau_1 : \mathbb{C}^{q\times l} \to \mathbb{R}$ is continuous at M_0. The main goal of this section is to show the following generalization of this result.

Theorem 6.9.2 *Let $\|\cdot\| : \mathbb{R}^{l\times q} \to \mathbb{R}$ be a norm with the following property.*

(g) *For every pair $(u,v)\in(\mathbb{R}^q\setminus\{0\}) \times (\mathbb{R}^l\setminus\{0\})$ there exists a rank-1-matrix $\Delta_0\in\mathbb{R}^{l\times q}$ such that $\Delta_0 u = v$ and $\|\Delta_0\| = \min\left\{ \|\Delta\| \;\middle|\; \Delta\in\mathbb{R}^{l\times q},\ \Delta u = v \right\}$.*

Then the real μ-function

$$\mu_{\mathbb{R}}^{\|\cdot\|} : \mathbb{C}^{q\times l} \to \mathbb{R}, \qquad \mu_{\mathbb{R}}^{\|\cdot\|}(M) = \Big(\min\{ \|\Delta\| \mid \Delta\in\mathbb{R}^{l\times q},\ 1\in\sigma(\Delta M) \}\Big)^{-1}$$

is continuous at all $M_0\in\mathbb{C}^{q\times l}\setminus\mathbb{R}^{q\times l}$.

Note that by Proposition 6.2.4 all norms which are rank-1-compatible with an operator norm have property (g). Before we prove Theorem 6.9.2 we state a consequence for real spectral value sets and give a more specific result for real perturbation values.
Recall that by Proposition 2.2.4 and Theorem 2.4.5 for any $(A,B,C)\in L_{n,l,q}$, $\rho>0$, the set $\sigma_{\mathbb{R}}^{\|\cdot\|}(A,B,C,\rho)$ is the topological closure of

$$\overset{\circ}{\sigma}{}_{\mathbb{R}}^{\|\cdot\|}(A,B,C,\rho) = \sigma(A)\cup\{\ s\in\mathbb{C}\setminus\sigma(A) \ \mid\ \mu_{\mathbb{R}}^{\|\cdot\|}(G(s)) > \rho^{-1}\ \}, \qquad G(s) = C(sI_n - A)^{-1}B.$$

Corollary 6.9.3 *Suppose that the underlying norm $\|\cdot\| : \mathbb{R}^{l\times q} \to \mathbb{R}$ has property (g). Then for any $(A,B,C)\in L_{n,l,q}$, $\rho>0$, the set $\overset{\circ}{\sigma}{}_{\mathbb{R}}^{\|\cdot\|}(A,B,C,\rho)\setminus(\sigma(A)\cup\mathcal{R}_G)$ is open, where $\mathcal{R}_G$ is the realness locus of $G(s)$.*

Proof: This is immediate from Theorem 6.9.2. □

Proposition 6.9.4 *The maximum real perturbation value $\tau_1 : \mathbb{C}^{q\times l} \to \mathbb{R}$ is discontinuous at $M_0\in\mathbb{C}^{q\times l}$ if and only if $M_0\in\mathbb{R}^{q\times l}$ and $\sigma_1(M_0)\neq\sigma_2(M_0)$.*

Proof: Since the spectral norm has property (g), the function $\tau_1(\cdot)$ is continuous at all nonreal matrices by Theorem 6.9.2. In the sequel let $M_0\in\mathbb{R}^{q\times l}$ be real. Then $\tau_1(M_0) = \sigma_1(M_0)$. For all $M\in\mathbb{C}^{q\times l}$ we have $\sigma_1(M)\geq\tau_1(M)\geq\sigma_2(M)$, see Corollary 6.7.2 for the second inequality. Since the functions $\mathbb{C}^{q\times l}\ni M\mapsto\sigma_k(M)$ are continuous it follows that τ_1 is continuous at M_0 if $\sigma_1(M_0)=\sigma_2(M_0)$. In the next chapter (see Proposition 7.2.4) we will show that for $z\in\mathbb{C}$,

$$\tau_1(zM_0) = \begin{cases} |z|\,\sigma_1(M_0) & \text{if } z\in\mathbb{R}, \\ |z|\sqrt{\sigma_1(M_0)\,\sigma_2(M_0)} & \text{otherwise.} \end{cases}$$

Thus, the function $M\mapsto\tau_1(M)$ is discontinuous at M_0 if $\sigma_1(M_0)\neq\sigma_2(M_0)$. □

We are now going to prove Theorem 6.9.2. Since μ-functions are always upper semi-continuous (see Section 2.4 and Section 6.2) we only have to show that $\mu_{\mathbb{R}}^{\|\cdot\|}(\cdot)$ is lower semi-continuous at $M_0\in\mathbb{C}^{q\times l}\setminus\mathbb{R}^{q\times l}$. We distinguish 3 cases, labeled by (Ck), $k=1,2,3$. The first case is

$(C1)$ $\quad\mu_{\mathbb{R}}^{\|\cdot\|}(M_0) = 0.$

If $\mu_{\mathbb{R}}^{\|\cdot\|}(M_0)\neq 0$ then there is $\Delta_0\in\mathbb{R}^{l\times q}$ and $x_0\in\mathbb{C}^l\setminus\{0\}$ such that

$$\|\Delta_0\|^{-1} = \mu_{\mathbb{R}}^{\|\cdot\|}(M_0) \quad\text{and}\quad \Delta_0 M_0 x_0 = x_0. \tag{6.53}$$

Note that the identity $\Delta_0 M_0 x_0 = x_0$ is equivalent to the pure real equation

$$\Delta_0 \left[\Re(M_0x_0)\ \Im(M_0x_0)\right] = \left[\Re(x_0)\ \Im(x_0)\right]. \tag{6.54}$$

We distinguish the cases

($C2$) The vectors $\Re(M_0x_0)$ and $\Im(M_0x_0)$ are linearly independent.

($C3$) The vectors $\Re(M_0x_0)$ and $\Im(M_0x_0)$ are linearly dependent.

The assumption that the underlying norm has property (g) is only needed in the case ($C3$).

In the case ($C1$) the continuity at M_0 of the upper semi-continuous function $\mu_{\mathbb{R}}^{\|\cdot\|} : \mathbb{C}^{q\times l} \to [0,\infty]$ is consequence of the following fact. We leave its easy proof to the reader.

Proposition 6.9.5 *Let $\mathcal{T}$ be a topological space and let $f : \mathcal{T} \to \mathbb{R}$ be an upper semi-continuous function. Suppose that f attaines a local minimum at $\theta_0 \in \mathcal{T}$. Then f is continuous at θ_0.*

Now, we treat the case ($C2$). We show that in this case there is a neighborhood $\mathcal{V} \subset \mathbb{C}^{q\times l}$ of M_0 and a continuous function $\Psi : \mathcal{V} \to \mathbb{R}^{l\times q}$, $M \overset{\Psi}{\longmapsto} \Delta_M$, satisfying $\Delta_{M_0} = \Delta_0$ and $\Delta_M(Mx_0) = x_0$ for every $M \in \mathcal{V}$. The latter implies that on $\mathcal{V}$ the function $\mu_{\mathbb{R}}^{\|\cdot\|}(\cdot)$ is bounded from below by the continuous function $M \mapsto \|\Delta_M\|^{-1}$. Since $\|\Delta_{M_0}\|^{-1} = \|\Delta_0\|^{-1} = \mu_{\mathbb{R}}^{\|\cdot\|}(M_0)$ this proves the lower semi-continuity of $\mu_{\mathbb{R}}^{\|\cdot\|}(\cdot)$ at M_0.

The construction of the function Ψ is based on the following lemma.

Lemma 6.9.6 *Suppose that $\Delta_0 X_0 = Y_0$, where $(X_0, Y_0, \Delta_0) \in \mathbb{R}^{q\times k} \times \mathbb{R}^{l\times k} \times \mathbb{R}^{l\times q}$ and $Y_0 \neq 0$. If the columns of X_0 are linearly independent then there is an open neighborhood $\mathcal{U} \subset \mathbb{R}^{q\times k}$ of X_0 and a continuous function*

$$\Phi : \mathcal{U} \times \mathbb{R}^{\ell\times k} \to \mathbb{R}^{l\times q}, \qquad (X, Y) \overset{\Phi}{\longmapsto} \Delta_{X,Y}$$

such that $\Delta_{X_0,Y_0} = \Delta_0$, and $\Delta_{X,Y} X = Y$ for every $(X, Y) \in \mathcal{U} \times \mathbb{R}^{l\times k}$.

Proof: Consider the bilinear map

$$\beta : \mathbb{R}^{l\times q} \times \mathbb{R}^{q\times k} \to \mathbb{R}^{l\times k} \qquad \beta(\Delta, X) := \Delta X.$$

Since the columns of X_0 are linearly independent, the matrix $X_0^T X_0$ is nonsingular. We have $\beta(Y(X_0^T X_0)^{-1} X_0,\ X_0) = Y$ for every $Y \in \mathbb{R}^{l\times k}$. Thus, the linear map $\beta(\cdot, X_0) : \mathbb{R}^{l\times q} \to \mathbb{R}^{l\times k}$ is onto. Since $Y_0 \neq 0$ and $\beta(\Delta_0, X_0) = \Delta_0 X_0 = Y_0$, the matrix Δ_0 is not contained in the kernel of $\beta(\cdot, X_0)$. Hence we can find a subspace S of $\mathbb{R}^{l\times q}$ such that S contains Δ_0 and the restriction $\beta(\cdot, X_0)|_S$ is bijective. After choosing bases in both spaces S and $\mathbb{R}^{l\times k}$ the linear maps $\beta(\cdot, X)|_S : S \to \mathbb{R}^{l\times k}$, $X \in \mathbb{R}^{q\times l}$, are represented by matrices A_X whose coefficients depend analytically on the entries of X. We have $\det(A_{X_0}) \neq 0$, and therefore $\det(A_X) \neq 0$ for all X in an open neighborhood $\mathcal{U} \subset \mathbb{R}^{q\times k}$ of X. Thus $\beta(\cdot, X)|_S$ is bijective for all $X \in \mathcal{U}$. For $X \in \mathcal{U}$ and $Y \in \mathbb{R}^{\ell\times k}$ set $\Phi(X, Y) := \Delta_{X,Y} := (\beta(\cdot, X)|_S)^{-1} Y$. Then Φ has the required properties. $\square$

We formulate the construction of the function Ψ, mentioned above, as a corollary.

Corollary 6.9.7 *Let $x_0 \in \mathbb{C}^l$, $y_0 \in \mathbb{C}^q$, $M_0 \in \mathbb{C}^{q\times l}$ and $\Delta_0 \in \mathbb{R}^{l\times q}$. Suppose that the vectors $\Re(M_0x_0)$ and $\Im(M_0x_0)$ are linearly independent. Then there is a an open neighborhood $\mathcal{V} \subset \mathbb{C}^{q\times l}$ of M_0 and a continuous function $\mathcal{V} \ni M \mapsto \Delta_M \in \mathbb{R}^{l\times q}$ satisfying $\Delta_{M_0} = \Delta_0$ and $\Delta_M(Mx_0) = x_0$ for every $M \in \mathcal{V}$.*

Proof: Set $X_0 = [\Re(M_0x_0)\ \Im(M_0x_0)]$ and $Y_0 = [\Re x_0\ \Im x_0]$. By Lemma 6.9.6 there is an open neighborhood $\mathcal{U} \subset \mathbb{R}^{q\times 2}$ of X_0 and a continuous function $\mathcal{U} \times \mathbb{R}^{l\times 2} \ni (X,Y) \mapsto \Delta_{X,Y}$ satisfying $\Delta_{X,Y}X = Y$ and $\Delta_{X_0,Y_0} = \Delta_0$. Let $\mathcal{V} \subset \mathbb{C}^{q\times l}$ be an open neighborhood of M_0 such that $[\Re(Mx_0)\ \Im(Mx_0)] \in \mathcal{U}$ for every $M \in \mathcal{V}$. Set $\Delta_M := \Delta_{[\Re(Mx_0)\ \Im(Mx_0)],\,[\Re x_0\ \Im x_0]}$. Then the map $\mathcal{V} \ni M \mapsto \Delta_M \in \mathbb{R}^{l\times q}$ has the required properties. $\square$

Thus, the proof for the case $(C2)$ is complete. Next, we treat the case $(C3)$. We need a lemma.

Lemma 6.9.8 *Let $u\in\mathbb{C}^q$, $v\in\mathbb{C}^q$ and $\Delta\in\mathbb{R}^{l\times q}$ be such that $\Delta u = v$. Suppose that $\Re u$ and $\Im u$ are linearly dependent. Then*

(*i*) *There is $z\in\mathbb{C}\setminus\{0\}$ such that the vectors zu and zv are both real.*

(*ii*) *If v is a real vector and $v \neq 0$ then u is a real vector as well.*

Proof: Claim (*i*) is part of Proposition 6.6.2. However, let us recall the argument. Since $\Re u$ and $\Im u$ are linearly dependent there is $(\lambda_1,\lambda_2) \in \mathbb{R}^2\setminus\{0\}$ such that $\lambda_2 u_1 + \lambda_1 u_2 = 0$. Set $z := \lambda_1 + i\lambda_2$. Then $\Im(zu) = \lambda_2 u_1 + \lambda_1 u_2 = 0$. Thus $zu \in \mathbb{R}^q$. Now, the relation $\Delta u = v$ and the realness of Δ imply $zv = \Delta(zu) \in \mathbb{R}^l$.

(*ii*). If v is real then $\Delta u = v$ implies $\Delta(\Im u) = 0$ and $\Delta(\Re u) = v$. Now, $v \neq 0$ yields that $\Re u \neq 0$. Since $\Re u$ and $\Im u$ are linearly dependent we have $\Im u = \lambda(\Re u)$ for some $\lambda \in \mathbb{R}$. The relation $\lambda v = \Delta(\lambda \Re u) = \Delta(\Im u) = 0$ yields $\lambda = 0$. Thus u is real. $\square$

In case $(C3)$ we have that $\Delta_0 M_0 x_0 = x_0 \neq 0$, where $\Re(M_0x_0)$ and $\Im(M_0x_0)$ are linearly dependent. Thus, according to the first claim of Lemma 6.9.8, we may assume that the vectors $x_0 \neq 0$ and $M_0x_0 \neq 0$ are both real. The relation $\|\Delta_0\|^{-1} = \mu_{\mathbb{R}}^{\|\cdot\|}(M_0)$ implies that

$$\|\Delta_0\| = \min\left\{ \|\Delta\| \;\middle|\; \Delta\in\mathbb{R}^{l\times q},\ \Delta M_0 x_0 = x_0 \right\}.$$

Since the norm $\|\cdot\|$ has property (g), we may assume that Δ_0 has rank 1. Then Δ_0 can be written in the form $\Delta_0 = x_0 w^T$ where $w\in\mathbb{R}^q$ and $w^T M_0 x_0 = 1$. It follows that

$$\Delta_0^T M_0^T w = w(x_0^T M_0^T w) = w. \tag{6.55}$$

Now, we have again two cases.

(a) The vectors $\Re(M_0^T w)$ and $\Im(M_0^T w)$ are linearly independent.

Let us define a 'dual' norm $\|\cdot\|_d : \mathbb{R}^{q\times l} \to \mathbb{R}$ by $\|\Delta^T\|_d := \|\Delta\|$ for every $\Delta \in \mathbb{R}^{l\times q}$. The equivalence

$$1\in\sigma(\Delta M) \quad\Leftrightarrow\quad 1\in\sigma(M\Delta) = \sigma((M\Delta)^T) = \sigma(\Delta^T M^T)$$

implies that $\mu_{\mathbb{R}}^{\|\cdot\|_d}(M^T) = \mu_{\mathbb{R}}^{\|\cdot\|}(M)$ for all $M\in\mathbb{C}^{q\times l}$. Thus, the function $\mu_{\mathbb{R}}^{\|\cdot\|}(\cdot)$ is continuous at M_0 if and only if the function $\mu_{\mathbb{R}}^{\|\cdot\|_d}(\cdot)$ is continuous at M_0^T. However, if $\Re(M_0^T w)$ and $\Im(M_0^T w)$ are linearly independent, then we have case $(C2)$ for the dual norm $\|\cdot\|^d$. Thus $\mu_{\mathbb{R}}^{\|\cdot\|}(\cdot)$ is continuous at M_0.

(b) The vectors $\Re(M_0^T w)$ and $\Im(M_0^T w)$ are linearly dependent.

In this case equation (6.55) and the second claim of Lemma 6.9.8 yield that $M_0^T w$ is real. Thus, $w^T M_0 x \in \mathbb{R} \setminus \{0\}$ for every x in an open neighborhood $\mathcal{W} \subset \mathbb{R}^l$ of x_0. Set $\Delta_x := (w^T M_0 x)^{-1} x w^T$, $x \in \mathcal{W}$. Then Δ_x is a real matrix and $\Delta_x M_0 x = x$ for every $x \in \mathcal{W}$. The function $\mathcal{W} \ni x \mapsto \Delta_x \in \mathbb{R}^{l \times q}$ is continuous. Moreover, $\Delta_{x_0} = \Delta_0$, and consequently, $\|\Delta_{x_0}\|^{-1} = \|\Delta_0\|^{-1} = \mu_{\mathbb{R}}^{\|\cdot\|}(M_0)$.

Now, let $\epsilon > 0$ be given. By assumption, M_0 is nonreal. Thus arbitrarily close to x_0 we can find an $x_1 \in \mathcal{W}$ satisfying $\Im(M_0 x_1) \neq 0$. By continuity, we can choose x_1 such that

$$\|\Delta_{x_1}\|^{-1} \geq \|\Delta_{x_0}\|^{-1} - \frac{\epsilon}{2} = \mu_{\mathbb{R}}^{\|\cdot\|}(M_0) - \frac{\epsilon}{2}. \tag{6.56}$$

We have $\Im(M_0 x_1) \neq 0$ and $\Delta_{x_1} M_0 x_1 = x_1 \in \mathbb{R}^l$. Thus the vectors $\Re(M_0 x_1)$ and $\Im(M_0 x_1)$ are linearly independent by claim (ii) of Lemma 6.9.8. Hence, by Corollary 6.9.7 (with x_0 replaced by x_1) there is an open neighborhood $\mathcal{V} \subset \mathbb{C}^{q \times l}$ of M_0 and a continuous function $\mathcal{V} \ni M \mapsto \Delta_M$ satisfying $\Delta_{M_0} = \Delta_{x_1}$ and $\Delta_M(M x_1) = x_1$ for every $M \in \mathcal{V}$. The latter implies that $\|\Delta_M\|^{-1} \leq \mu_{\mathbb{R}}^{\|\cdot\|}(M)$ for $M \in \mathcal{V}$. By continuity there is a neighborhood $\mathcal{V}'$ of M_0 such that $\|\Delta_{x_1}\|^{-1} - \frac{\epsilon}{2} = \|\Delta_{M_0}\|^{-1} - \frac{\epsilon}{2} \leq \|\Delta_M\|^{-1}$ for every $M \in \mathcal{V}'$. By combining this with (6.56) we obtain

$$\mu_{\mathbb{R}}^{\|\cdot\|}(M_0) - \epsilon \leq \|\Delta_{x_1}\|^{-1} - \frac{\epsilon}{2} \leq \|\Delta_M\|^{-1} \leq \mu_{\mathbb{R}}^{\|\cdot\|}(M), \qquad \forall\, M \in \mathcal{V}'.$$

Thus, we have shown that the function $\mu_{\mathbb{R}}^{\|\cdot\|}(\cdot)$ is lower semi-continuous at M_0, and the proof of Theorem 6.9.2 is complete. $\square$

Chapter 7

Real spectral value sets for special matrices

This chapter deals with real spectral value sets of special matrices. In the first section we give a formula for the largest real perturbation value of block diagonal matrices. The proof of this formula is based on the characterization of real perturbation values by hermitian-symmetric inequalities. As an application we determine in Sections 7.2 and 7.3 the real spectral value sets of diagonal matrices and real pseudospectra of normal matrices. Section 7.4 is about real pseudospectra of 2×2 matrices with respect to arbitrary unitarily invariant norms. Finally, in the last section we consider root sets of uncertain polynomials.

7.1 A formula for block diagonal matrices

For any direct sum of matrices $M_j \in \mathbb{C}^{q_j \times l_j}$ the largest singular value satisfies

$$\sigma_1\left(\bigoplus_{j\in\underline{m}} M_j\right) = \max_{j\in\underline{m}} \sigma_1(M_j). \tag{7.1}$$

From this fact it follows that for any family of matrix triples $(A_j, B_j, C_j) \in L_{n_j,l_j,q_j}$, $j \in \underline{m}$,

$$\sigma_{\mathbb{C}}^{(2)}\left(\bigoplus_{j\in\underline{m}} A_j,\, \bigoplus_{j\in\underline{m}} B_j,\, \bigoplus_{j\in\underline{m}} C_j,\, \rho\right) = \bigcup_{j\in\underline{m}} \sigma_{\mathbb{C}}^{(2)}(A_j, B_j, C_j\ , \rho)\,.$$

Unfortunately, (7.1) fails in most cases if one replaces σ_1 by τ_1. Here is a simple counterexample: There is no real number $\Delta \in \mathbb{R}$ such that $\Delta i = 1$. Thus the 1×1-matrix $[i]$ satisfies $\tau_1([i]) = 0$. However, Proposition 6.8.3 yields $\tau_1([i] \oplus [i]) = \tau_1\left(\begin{bmatrix} i & 0 \\ 0 & i \end{bmatrix}\right) = \sigma_1\left(\begin{bmatrix} i & 0 \\ 0 & i \end{bmatrix}\right) = 1$. Observe that for the largest real perturbation value of direct sums we always have

$$\tau_1\left(\bigoplus_{j\in\underline{m}} M_j\right) \geq \max_{j\in\underline{m}} \tau_1(M_j), \qquad M_j \in \mathbb{C}^{q_j \times l_j}. \tag{7.2}$$

To see this, assume without loss of generality that $\max_{j\in\underline{m}} \tau_1(M_j) = \tau_1(M_1) > 0$, and let $\Delta_0 \in \mathbb{R}^{l_1 \times q_1}$ be such that $\|\Delta_0\|_2^{-1} = \tau_1(M_1)$ and $1 \in \sigma(\Delta_0 M_1)$. Set $\Delta_1 := \Delta \oplus O$, where O is a zero matrix of appropriate dimension such that the product $\Delta_1 \left(\bigoplus_{j\in\underline{m}} M_j\right)$ is well defined. Then $\|\Delta_1\|_2 = \|\Delta_0\|_2$ and $1 \in \sigma\left(\Delta_1 \left(\bigoplus_{j\in\underline{m}} M_j\right)\right)$, whence $\tau_1\left(\bigoplus_{j\in\underline{m}} M_j\right) \geq \tau_1(M_1)$.

The aim of this section is to prove an analog to (7.1) which is more precise then the inequality (7.2). We need the following fact.

Fact: Let $\mathcal{HS}_n^0$ denote the set of $(H,S)\in\mathcal{HS}_n$ with the property that for every $x\in\mathbb{C}^n\setminus\{0\}$, $x^TSx=0$ implies $x^*Hx<0$. For $(H,S)\in\mathcal{HS}_n^0$ we set $\frac{x^*Hx}{|x^TSx|}:=-\infty$ if $x^TSx=0$, $x\neq 0$. With this convention the map

$$\mathcal{HS}_n^0\times(\mathbb{C}^n\setminus\{0\})\ni((H,S),x)\longmapsto\frac{x^*Hx}{|x^TSx|}$$

becomes continuous. Therefore, the map

$$\mathcal{HS}_n^0\ni(H,S)\longmapsto\max_{\substack{x\in\mathbb{C}^n\\ \|x\|_2=1}}\frac{x^*Hx}{|x^TSx|}=\max_{x\in\mathbb{C}^n\setminus\{0\}}\frac{x^*Hx}{|x^TSx|}$$

is well defined and continuous.□

By Corollary 6.1.8 we have for any $M\in\mathbb{C}^{q\times l}$,

$$\tau_1(M)=\max\{\,t\geq 0\,|\ \exists x\in\mathbb{C}\setminus\{0\}:x^*(M^*M-t^2I_l)x\geq|x^T(M^TM-t^2I_l)x|\ \}.$$

Thus $t>\tau_1(M)$ implies $(M^*M-t^2I_l,M^TM-t^2I_l)\in\mathcal{HS}_l^0$. Therefore, the function

$$\tau_1(M)<t\ \longmapsto\ \phi_M(t):=\max_{x\in\mathbb{C}^l\setminus\{0\}}\frac{x^*(M^*M-t^2I_l)x}{|x^T(M^TM-t^2I_l)x|}\tag{7.3}$$

is well defined and continuous. If $t>\sigma_1(M)$ then $\phi_M(t)<0$. For matrix pairs $M_1\in\mathbb{C}^{q_1\times l_1}$, $M_2\in\mathbb{C}^{q_2\times l_2}$ we define

$$T(M_1,M_2)\ \ :=\ \ \{\,t>\max\{\tau_1(M_1),\tau_1(M_2)\}\ \mid\ \phi_{M_1}(t)+\phi_{M_2}(t)\geq 0\,\}.$$

The set $T(M_1,M_2)$ is contained in the intervall $[0,\max\{\sigma_1(M_1),\sigma_1(M_2)\}]$. If $T(M_1,M_2)\neq\emptyset$ then by continuity $\max T(M_1,M_2)$ exists. We now state the main result of this section.

Theorem 7.1.1 *Let $M_j\in\mathbb{C}^{q_j\times l_j}$, $j\in\underline{m}$, $m\geq 2$. Then*

$$\tau_1\left(\bigoplus\nolimits_{j\in\underline{m}}M_j\right)\ =\ \max_{1\leq j<k\leq m}\tau_1\left(M_j\oplus M_k\right).\tag{7.4}$$

Moreover,

$$\tau_1(M_1\oplus M_2)=\begin{cases}\max\{\tau_1(M_1),\tau_1(M_2)\} & \text{if } T(M_1,M_2)=\emptyset,\\ \max\,T(M_1,M_2) & \text{otherwise.}\end{cases}\tag{7.5}$$

In the latter case a real matrix $\Delta_0\in\mathbb{R}^{(l_1+l_2)\times(q_1+q_2)}$ satisfying $\|\Delta_0\|_2=1/\tau_1(M_1\oplus M_2)$ and $1\in\sigma(\Delta_0(M_1\oplus M_2))$ can be obtained as follows: Let $t_0=\max T(M_1,M_2)$ and let $y_k\in\mathbb{C}^{l_k}\setminus\{0\}$ be such that $\phi_{M_k}(t_0)=\frac{y_k^*(M_k^*M_k-t_0^2I_{l_k})y_k}{|y_k^T(M_k^TM_k-t_0^2I_{l_k})y_k|}$, $k\in\{1,2\}$. Let $w_k\in\mathbb{C}\setminus\{0\}$ be a square root of $y_k^T(M_k^TM_k-t_0^2I_{l_k})y_k$. Set $x_1:=w_1^{-1}y_1$, $x_2:=i\,w_2^{-1}y_2$ and $\Delta_0:=\begin{bmatrix}x_1&\overline{x_1}\\x_2&\overline{x_2}\end{bmatrix}\begin{bmatrix}M_1x_1&\overline{M_1x_1}\\M_2x_2&\overline{M_2x_2}\end{bmatrix}^\dagger$. Then Δ_0 has the required properties.

As a corollary to Theorem 7.1.1 we obtain the following result on real spectral value sets of block diagonal matrices with respect to the spectral norm.

Corollary 7.1.2 *For any family of matrix triples $(A_j, B_j, C_j) \in L_{n_j,l_j,q_j}$, $j \in \underline{m}$, $m \geq 2$, $\rho \geq 0$,*

$$\begin{aligned}\sigma_{\mathbb{R}}^{(2)}\left(\bigoplus_{j\in\underline{m}} A_j,\, \bigoplus_{j\in\underline{m}} B_j,\, \bigoplus_{j\in\underline{m}} C_j,\, \rho\right) &= \bigcup_{1\leq j<k\leq m} \sigma_{\mathbb{R}}^{(2)}(A_j \oplus A_k, B_j \oplus B_k, C_j \oplus C_k,\ \rho) \\ &\supseteq \bigcup_{j\in\underline{m}} \sigma_{\mathbb{R}}^{(2)}(A_j, B_j, C_j,\ \rho)\,.\end{aligned}$$

Proof of Corollary 7.1.2: Let $G(s) = \left(\bigoplus_{j\in\underline{m}} C_j\right)\left(sI - \bigoplus_{j\in\underline{m}} A_j\right)^{-1}\left(\bigoplus_{j\in\underline{m}} B_j\right)$ and $G_j(s) = C_j(sI_{l_j} - A)^{-1}B_j$, $j \in \underline{m}$. Then Theorem 7.1.1 yields

$$\tau_1(G(s)) = \tau_1(\bigoplus_{j\in\underline{m}} G_j(s)) = \max_{1\leq j<k\leq m} \tau_1\left(G_j(s) \oplus G_k(s)\right) \geq \max_{j\in\underline{m}} \tau_1(G_j(s)).$$

This implies the result. □

The proof of Theorem 7.1.1 is based on a Proposition 7.1.4 below which in turn is a consequence of the following lemma.

Lemma 7.1.3 *Let $\lambda_1, \ldots, \lambda_m \in \mathbb{R}$ and $z_1, \ldots, z_m \in \mathbb{C}$ be such that*

$$\textstyle\sum_{j\in\underline{m}} \lambda_j \geq \left|\sum_{j\in\underline{m}} z_j\right|. \tag{7.6}$$

Then at least one of the following conditions (a) and (b) holds.

(*a*) $\lambda_j \geq |z_j|$ *for some* $j \in \underline{m}$.

(*b*) $\frac{\lambda_j}{|z_j|} + \frac{\lambda_k}{|z_k|} \geq 0$ *for some* $j, k \in \underline{m}$ *with* $j \neq k$ *and* $|z_j| \neq 0 \neq |z_k|$.

Proof: The claim is obvious if $m = 1$, so let $m \geq 2$. Suppose that (7.6) holds and neither (*a*) nor (*b*) is satisfied. Then we have for some $j_0 \in \underline{m}$,

$$0 \leq \lambda_{j_0} < |z_{j_0}| \text{ and } \lambda_j < -\frac{\lambda_{j_0}}{|z_{j_0}|}|z_j| \text{ for all } j \neq j_0.$$

This yields the contradiction

$$\textstyle\sum_{j\in\underline{m}} \lambda_j < \frac{\lambda_{j_0}}{|z_{j_0}|}\left(|z_{j_0}| - \sum\limits_{j\neq j_0} |z_j|\right) \leq \frac{\lambda_{j_0}}{|z_{j_0}|}\left|\sum_{j\in\underline{m}} z_j\right| < \left|\sum_{j\in\underline{m}} z_j\right|.$$ □

Proposition 7.1.4 *Let $H_j \in \mathbb{C}^{n_j\times n_j}$ be hermitian and $S_j \in \mathbb{C}^{n_j\times n_j}$ be symmetric matrices, $j \in \underline{m}$, $m \geq 2$. Then the following assertions (i) and (ii) are equivalent.*

(*i*) *There is $x \in \mathbb{C}^n \setminus \{0\}$, $n = \sum_{k\in\underline{m}} n_k$, such that*

$$x^*\left(\bigoplus_{j\in\underline{m}} H_j\right)x \geq \left|x^T\left(\bigoplus_{j\in\underline{m}} S_j\right)x\right|. \tag{7.7}$$

(*ii*) *One of the following conditions holds.*

(*a*) *There is* $j \in \underline{m}$ *and* $x_j \in \mathbb{C}^{n_j} \setminus \{0\}$ *such that* $x_j^* H_j x_j \geq |x_j^T S x_j|$.

(*b*) *There are* $j, k \in \underline{m}$ $j \neq k$, *and* $x_j \in \mathbb{C}^{n_j}$, $x_k \in \mathbb{C}^{n_k}$ *such that* $x_j^T S_j x_j \neq 0 \neq x_k^T S_k x_k$ *and*

$$\frac{x_j^* H_j x_j}{|x_j^T S_j x_j|} + \frac{x_k^* H_k x_k}{|x_j^T S_k x_k|} \geq 0.$$

Proof: Let $x = [x_1^T \dots x_m^T]^T \neq 0$, $x_j \in \mathbb{C}^{n_j}$, and $J = \{\, j \in \underline{m} \mid \ x_j \neq 0 \,\}$. Then the inequality (7.7) reads

$$\textstyle\sum_{j \in J} x_j^* H_j x_j \geq \left| \sum_{j \in J} x_j^T S_j x_j \right|.$$

Thus, (*i*) implies (*ii*) by Lemma 7.1.3. Obviously, (*a*) implies (*i*). Suppose now, that (*b*) holds. Let $w_j, w_k \neq 0$ be square roots of $x_j^T S_j x_j$ and $x_k^T S_k x_k$ respectively. Set $\widetilde{x}_j := w_j^{-1} x_j$ and $\widetilde{x}_k := i\, w_k^{-1} x_k$. Then

$$\widetilde{x}_j^* H_j \widetilde{x}_j + \widetilde{x}_k^* H_k \widetilde{x}_k = \frac{x_j^* H_j x_j}{|x_j^T S_j x_j|} + \frac{x_k^* H_k x_k}{|x_j^T S_k x_k|} \geq 0, \qquad \text{and} \qquad \widetilde{x}_j^T S_j \widetilde{x}_j + \widetilde{x}_k^T S_k \widetilde{x}_k = 0$$

Set $\widetilde{x}_\ell := 0 \in \mathbb{C}^{n_\ell}$, $\ell \in \underline{m} \setminus \{j, k\}$, and $\widetilde{x} := [\widetilde{x}_1^T \dots \widetilde{x}_m^T]^T$. Then $\widetilde{x}^* \left(\bigoplus_{\ell \in \underline{m}} H_\ell \right) \widetilde{x} \geq \left| \widetilde{x}^T \left(\bigoplus_{\ell \in \underline{m}} S_\ell \right) \widetilde{x} \right|$. Thus (*b*) implies (*i*). □

Proof of Theorem 7.1.1: First, recall that by Theorem 6.1.7 and Corllary 6.1.8 we have for any $t \geq 0$ and any complex matrix M that $t \leq \tau_1(M)$ if and only if there is a vector $x \neq 0$ such that $x^*(M^*M - t^2 I)x \geq |x^*(M^T M - t^2 I)x|$.

Let $t \geq 0$. Then $t \leq \tau_1 \left(\bigoplus_{j \in J} M_j \right)$ if and only if there is an $x \neq 0$ such that

$$x^* \left(\left(\bigoplus_{j \in \underline{m}} M_j \right)^* \left(\bigoplus_{j \in \underline{m}} M_j \right) - t^2 I \right) x \geq \left| x^T \left(\left(\bigoplus_{j \in \underline{m}} M_j \right)^T \left(\bigoplus_{j \in \underline{m}} M_j \right) - t^2 I \right) x \right|;$$

equivalently,

$$x^* \left(\bigoplus_{j \in \underline{m}} (M_j^* M_j - t^2 I_{l_j}) \right) x \geq \left| x^T \left(\bigoplus_{j \in \underline{m}} (M_j^T M_j - t^2 I_{l_j}) \right) x \right|.$$

By Proposition 7.1.4 this is equivalent to the statement (A_t) below.

(A_t) The number t fulfills one of the following conditions.

(*a*) There is $j \in \underline{m}$ and $x_j \in \mathbb{C}^{n_j} \setminus \{0\}$ such that $x_j^*(M_j^* M_j - t^2 I_{l_j}) x_j \geq |x_j^T (M_j^T M_j - t^2 I_{l_j}) x_j|$.

(*b*) There are $j, k \in \underline{m}$ $j \neq k$, and $x_j \in \mathbb{C}^{n_j}$, $x_k \in \mathbb{C}^{n_k}$ such that $x_j^T (M_j^T M_j - t^2 I_{l_j}) x_j \neq 0 \neq x_k^T (M_k^T M_k - I_{l_k}) x_k$ and

$$\frac{x_j^* (M_j^* M_j - t^2 I_{l_j}) x_j}{|x_j^T (M_j^T M_j - t^2 I_{l_j}) x_j|} + \frac{x_k^* (M_k^* M_k - t^2 I_{l_k}) x_k}{|x_j^T (M_k^T M_k - t^2 I_{l_k}) x_k|} \geq 0.$$

But the statement (A_t) is equivalent to the assertion (B_t):

(B_t) There is a pair $j,k\in\underline{m}$, $j<k$, such that one of the following conditions holds.

- (a_{jk}) There is an $x_j\in\mathbb{C}^{n_j}\setminus\{0\}$ such that $x_j^*(M_j^*M_j-t^2I_{l_j})x_j \geq |x_j^T(M_j^TM_j-t^2I_{l_j})x_j|$ or, there is an $x_k\in\mathbb{C}^{n_k}\setminus\{0\}$ such that $x_k^*(M_k^*M_k-t^2I_{l_k})x_k \geq |x_k^T(M_k^TM_k-t^2I_{l_k})x_k|$.
- (b_{jk}) Condition (a_{jk}) fails and there are $x_j\in\mathbb{C}^{n_j}$, $x_k\in\mathbb{C}^{n_k}$ such that $x_j^T(M_j^TM_j-t^2I_{l_j})x_j \neq 0 \neq x_k^T(M_k^TM_k-t^2I_{l_k})x_k$ and

$$\frac{x_j^*(M_j^*M_j-t^2I_{l_j})x_j}{|x_j^T(M_j^TM_j-t^2I_{l_j})x_j|}+\frac{x_k^*(M_k^*M_k-t^2I_{l_k})x_k}{|x_j^T(M_k^TM_k-t^2I_{l_k})x_k|}\geq 0.$$

Condition (B_t) is in turn equivalent to the statement (C_t):

(C_t) There is a pair $j,k\in\underline{m}$, $j<k$, satisfying one of the following conditions.

- (a'_{jk}) $t\leq \max\{\tau_1(M_j),\tau_1(M_k)\}$
- (b'_{jk}) $t> \max\{\tau_1(M_j),\tau_1(M_k)\}$ and $t\in T(M_j,M_k)$.

Thus, we have shown that the equivalence

$$t\leq \tau_1\left(\bigoplus\nolimits_{j\in\underline{m}}M_j\right)\quad\Leftrightarrow\quad (C_t). \tag{7.8}$$

holds for all $t\geq 0$. In particular, we have for the case $m=2$,

$$t\leq\tau_1(M_1\oplus M_2)\quad\Leftrightarrow\quad\begin{cases} t\leq\max\{\tau_1(M_1),\tau_1(M_2)\}\\ \text{or}\\ t>\max\{\tau_1(M_1),\tau_1(M_1)\}\text{ and }t\in T(M_1,M_2).\end{cases}$$

This yields Formula (7.5), i.e.

$$\tau_1(M_1\oplus M_2)=\begin{cases}\max\{\tau_1(M_1),\tau_1(M_2)\} & \text{if } T(M_1,M_2)=\emptyset,\\ \max T(M_1,M_2) & \text{otherwise.}\end{cases}$$

Hence, for arbitrary m the assertion (C_t) is equivalent to the statement

(D_t) There is a pair $j,k\in\underline{m}$, $j<k$, such that $t\leq\tau_1(M_j\oplus M_k)$.

On replacing (C_t) by (D_t) in (7.8) we obtain Formula (7.4).

Suppose now, that $T(M_1,M_2)\neq\emptyset$ and let $t_0=\tau_1(M_1\oplus M_2)$. Then the continuity of the function $t\mapsto\phi_{M_1}(t)+\phi_{M_2}(t)$ yields $\phi_{M_1}(t_0)+\phi_{M_2}(t_0)=0$. Hence, the vectors y_1, y_2, mentioned in the theorem, satisfy

$$\frac{y_1^*(M_1^*M_1-t_0^2I_{l_1})y_1}{|y_1^T(M_1^TM_1-t_0^2I_{l_1})y_1|}+\frac{y_2^*(M_2^*M_2-t_0^2I_{l_2})y_2}{|y_2^T(M_2^TM_2-t_0^2I_{l_2})y_2|}=0.$$

Let $x=[x_1^T\ x_2^T]^T$ where $x_1=w_1^{-1}y_1$, $x_2=iw_2^{-1}y_2$ and $w_k=\sqrt{y_k^T(M_k^TM_k-t_0^2I_{l_k})y_k}$. Then

$$\begin{aligned} x^*((M_1\oplus M_2)^*(M_1\oplus M_2)-t_0^2)x &= x_1^*(M_1^*M_1-t_0^2I_{I_1})x_1+x_2^*(M_2^*M_2-t_0^2I_{I_1})x_2=0,\\ x^T((M_1\oplus M_2)^T(M_1\oplus M_2)-t_0^2)x &= x_1^T(M_1^TM_1-t_0^2I_{I_1})x_1+x_2^T(M_2^TM_2-t_0^2I_{I_1})x_2=0.\end{aligned}$$

Hence,

$$\begin{bmatrix} \|(M_1 \oplus M_2)x\|_2^2 & \overline{x^T(M_1 \oplus M_2)^T(M_1 \oplus M_2)x} \\ x^T(M_1 \oplus M_2)^T(M_1 \oplus M_2)x & \|(M_1 \oplus M_2)x\|_2^2 \end{bmatrix} = t_0^2 \begin{bmatrix} \|x\|_2^2 & \overline{x^T x} \\ x^T x & \|x\|_2^2 \end{bmatrix}.$$

Thus Propositions 6.6.2 and claim (iii) of Proposition 6.6.3 yield that $\Delta_0(M_1 \oplus M_2)x = x$ and $\|\Delta_0\|_2 = t_0^{-1}$, where

$$\begin{aligned} \Delta_0 &= [x \ \ \overline{x}\,][(M_1 \oplus M_2)x \ \ \overline{(M_1 \oplus M_2)x}\ ]^\dagger \\ &= \begin{bmatrix} x_1 & \overline{x_1} \\ x_2 & \overline{x_2} \end{bmatrix} \begin{bmatrix} M_1x_1 & \overline{M_1x_1} \\ M_2x_2 & \overline{M_2x_2} \end{bmatrix}^\dagger \in \mathbb{R}^{(l_1+l_2)\times(q_1+q_2)}. \end{aligned}$$

□

Next, we consider the minimal real perturbation value, $\widetilde{\tau}_{\min}(\cdot)$, of direct sums. By Corollary 6.1.8 we have for any square matrix $M \in \mathbb{C}^{n\times n}$,

$$\widetilde{\tau}_{\min}(M) = \min\{\ t \geq 0 \mid \ \exists x \in \mathbb{C}^n \backslash \{0\} : \ \ x^*(t^2 I_n - M^*M)x \geq |x^T(t^2I_n - M^TM)x| \ \ \}.$$

Moreover, $\widetilde{\tau}_{\min}(M) = 0$ if and only if M is singular. Thus, for nonsingular $M \in \mathbb{C}^{n\times n}$ the interval $[0, \widetilde{\tau}_{\min}(M)\,)$ is nonempty, and $t \in [0, \widetilde{\tau}_{\min}(M)\,)$ implies that $(t^2I_n - M^*M, t^2I_n - M^TM) \in \mathcal{HS}_n^0$. Hence, the function

$$[0, \widetilde{\tau}_{\min}(M)\,) \ni t \ \longmapsto \ \widetilde{\phi}_M(t) := \max_{x\in\mathbb{C}^n\backslash\{0\}} \frac{x^*(t^2I_n - M^*M)x}{|x^T(t^2I_n - M^*M)x|} \tag{7.9}$$

is well defined and continuous. For a pair of square matrices $M_1 \in \mathbb{C}^{n_1\times n_1}$, $M_2 \in \mathbb{C}^{n_2\times n_2}$ let

$$\widetilde{T}(M_1, M_2) := \{\ t \geq 0 \ \mid \ \ t < \min\{\tau_{\min}(M_1), \tau_{\min}(M_2)\}, \ \ \widetilde{\phi}_{M_1}(t) + \widetilde{\phi}_{M_2}(t) \geq 0 \ \ \}$$

Theorem 7.1.5 *Let $M_j \in \mathbb{C}^{n_j\times n_j}$, $j \in \underline{m}$, $m \geq 2$. Then*

$$\widetilde{\tau}_{\min}\left(\bigoplus\nolimits_{j\in\underline{m}} M_j\right) \ = \ \min_{1\leq j<k\leq m} \widetilde{\tau}_{\min}\left(M_j \oplus M_k\right). \tag{7.10}$$

Moreover,

$$\widetilde{\tau}_{\min}(M_1 \oplus M_2) = \begin{cases} \min\{\widetilde{\tau}_{\min}(M_1), \widetilde{\tau}_{\min}(M_2)\} & \text{if } \widetilde{T}(M_1, M_2) = \emptyset, \\ \min \widetilde{T}(M_1, M_2) & \text{otherwise.} \end{cases}$$

In the latter case a real matrix $\Delta_0 \in \mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$ satisfying $\|\Delta_0\|_2 = \widetilde{\tau}_{\min}(M_1 \oplus M_2)$ and $\det((M_1 \oplus M_2) - \Delta_0) = 0$ can be obtained as follows: Let $t_0 = \min \widetilde{T}(M_1, M_2)$ and let $y_k \in \mathbb{C}^{n_k}\backslash\{0\}$ be such that $\widetilde{\phi}_{M_k}(t_0) = \frac{y_k^*(t_0^2 I_{l_k} - M_k^* M_k)y_k}{|y_k^T(t_0^2 I_{n_k} - M_k^T M_k)y_k|}$, $k \in \{1, 2\}$. Let $w_k \in \mathbb{C} \setminus \{0\}$ be a square root of $y_k^T(t_0^2 I_{n_k} - M_k^T M_k)y_k$. Set $x_1 := w_1^{-1}y_1$, $x_2 := i\,w_2^{-1}y_2$ and $\Delta_0 := \begin{bmatrix} M_1x_1 & \overline{M_1x_1} \\ M_2x_2 & \overline{M_2x_2} \end{bmatrix} \begin{bmatrix} x_1 & \overline{x_1} \\ x_2 & \overline{x_2} \end{bmatrix}^\dagger$. Then Δ_0 has the required properties.

Theorem 7.1.5 can be proved in the same way as Theorem 7.1.1 via Proposition 7.1.4 . An alternative method to show Theorem 7.1.5 is by using Theorem 7.1.4 and the fact that $\widetilde{\tau}_{\min}(M) = \tau_1(M^{-1})^{-1}$. The details are left to the reader.

Either from Theorem 7.1.5 or by specializing Corollary 7.1.2 we obtain

Corollary 7.1.6 *Let $A_j \in \mathbb{C}^{n_j\times n_j}$, $j \in \underline{m}$, $m \geq 2$. Then for any $\rho \geq 0$,*

$$\sigma_{\mathbb{R}}^{(2)}\left(\bigoplus\nolimits_{j\in\underline{m}} A_j,\, \rho\right) \ = \ \bigcup_{1\leq j<k\leq m} \sigma_{\mathbb{R}}^{(2)}\left(A_j \oplus A_k,\, \rho\right) \supseteq \bigcup_{j\in\underline{m}} \sigma_{\mathbb{R}}^{(2)}\left(A_j,\ \rho\right).$$

7.2 τ_1 for diagonal matrices

In this section we give explicit formulas for the maximum real perturbation value of diagonal matrices. Theorem 7.1.1 yields that for any diagonal matrix

$$\tau_1(\,\mathrm{diag}(a_1,\ldots,a_n)\,) = \max_{1\le j<k\le n} \tau_1(\,\mathrm{diag}(a_j,a_k)\,).$$

Thus it is enough to consider the 2×2 case.

Proposition 7.2.1 *Let $a_1, a_2 \in \mathbb{C}$ and $|a_1| \ge |a_2|$. Then*

$$\tau_1(\mathrm{diag}(a_1,a_2)) = \begin{cases} |a_1| & \text{if } a_1\in\mathbb{R}, \\ |a_2| & \text{if } a_1\notin\mathbb{R},\ \ a_2\in\mathbb{R}, \\ t_0(a_1,a_2), & \text{if } a_1\notin\mathbb{R},\ \ a_2\notin\mathbb{R}, \end{cases}$$

where $t_0(a_1,a_2)$ is the zero of the strictly decreasing function

$$t \mapsto \frac{|a_1|^2-t^2}{|a_1^2-t^2|} + \frac{|a_2|^2-t^2}{|a_2^2-t^2|}, \qquad t\ge 0.$$

The zero $t_0(a_1,a_2)$ can be calculated as follows: Let $\alpha := (\Im a_1)^2 - (\Im a_2)^2$. If $\alpha = 0$ then

$$t_0(a_1,a_2) = \sqrt{\frac{|a_1|^2+|a_2|^2}{2}}. \tag{7.11}$$

If $\alpha\ne 0$ set $\beta := \alpha^{-1}((\Re a_1)^2(\Im a_2)^2 - (\Re a_2)^2(\Im a_1)^2)$, $\gamma := \alpha^{-1}(|a_2|^2(\Im a_1)^2 - |a_1|^2(\Im a_2)^2)$ and $\zeta := \frac{|a_1|^2-\beta}{|a_1^2-\beta|} + \frac{|a_2|^2-\beta}{|a_2^2-\beta|}$. Then

$$t_0(a_1,a_2) = \sqrt{\beta + sign(\zeta)\sqrt{\beta^2-\gamma}}. \tag{7.12}$$

In the proof of Proposition 7.2.1 and of Proposition 7.4.2 in the next section we will use the easily verified

Lemma 7.2.2 *Let $m_1, m_2, d_1, d_2 \in \mathbb{R}$ be such that $m_k + d_k > 0$ and $m_k - d_k \ge 0$. Then the following equivalence holds.*

$$\sqrt{\frac{m_1-d_1}{m_1+d_1}} - \sqrt{\frac{m_2-d_2}{m_2+d_2}} = 0 \quad\Leftrightarrow\quad m_1d_2 - m_2d_1 = 0.$$

Proof of Proposition 7.2.1: For a 1×1-matrix $[a] = a\in\mathbb{C}$ we have

$$\tau_1(a) = \begin{cases} \sigma_1(a) = |a|, & \text{if } a\in\mathbb{R} \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, the function ϕ, defined in (7.3) reads in the 1×1 case:

$$\phi_a(t) = \max_{x\in\mathbb{C}\setminus\{0\}} \frac{\overline{x}(|a|^2-t^2)x}{|x(a^2-t^2)x|} = \frac{|a|^2-t^2}{|a^2-t^2|}, \qquad t\ge 0,\ t^2\ne a^2.$$

If $a \in \mathbb{R}$ then $\phi_a(t)$ is not defined for $t = |a|$, and $\phi_a(t) = \begin{cases} 1 & \text{if } t < |a|, \\ -1 & \text{if } t > |a|. \end{cases}$
If a is nonreal then ϕ_a is defined for all $t \geq 0$. By differentiating ϕ_a, using the fact that $|a^2 - t^2|^2 = t^4 - 2\Re(a^2)t^2 + |a|^4$, one obtains

$$\phi_a'(t) = \frac{-2t(|a|^2 - \Re(a^2))(|a|^2 + t^2)}{|a^2 - t^2|^3}.$$

Thus, the function ϕ_a is strictly decreasing. Moreover, $\phi_a(0) = 1$, $\phi_a(|a|) = 0$ and $\lim_{t\to\infty} \phi_a(t) = -1$. We need the following representation of ϕ_a for $a \notin \mathbb{R}$. Set

$$m_a(t) := |a|^4 - 2(\Re a)^2 t^2 + t^4, \qquad d_a(t) := 2(\Im a)^2 t^2.$$

Then $(|a|^2 - t^2)^2 = m_a(t) - d_a(t)$ and $|a^2 - t^2|^2 = m_a(t) + d_a(t)$. Therefore

$$\phi_a(t) = \begin{cases} \sqrt{\frac{m_a(t) - d_a(t)}{m_a(t) + d_a(t)}} & \text{if } t \leq |a|, \\ -\sqrt{\frac{m_a(t) - d_a(t)}{m_a(t) + d_a(t)}} & \text{if } t > |a|. \end{cases}$$

We have to distinguish the following three cases which are illustrated in Figure 7.1. Note that, by assumption, $|a_1| \geq |a_2|$.

- a_1 is real. Then $\max\{\tau_1(a_1), \tau_1(a_2)\} = |a_1|$ and for $t > |a_1|$, $\phi_{a_1}(t) + \phi_{a_2}(t) < 0$. Thus Theorem 7.1.1 yields $\tau_1(\mathrm{diag}(a_1, a_2)) = |a_1|$.

- a_1 is nonreal and a_2 is real. Then $\max\{\tau_1(a_1), \tau_1(a_2)\} = |a_2|$ and for $t > |a_2|$, $\phi_{a_1}(t) + \phi_{a_2}(t) < 0$. Thus $\tau_1(\mathrm{diag}(a_1, a_2)) = |a_2|$.

- a_1 and a_2 are both nonreal. Then $\max\{\tau_1(a_1), \tau_2(a_2)\} = 0$. Moreover, the function $t \mapsto \phi_{a_1}(t) + \phi_{a_2}(t)$ is strictly decreasing and has a unique zero t_0. We have $|a_2| \leq t_0 \leq |a_1|$, and Theorem 7.1.1 yields $\tau_1(\mathrm{diag}(a_1, a_2)) = t_0$. It remains to verify the formulas for the computation of t_0. For $|a_2| \leq t \leq |a_1|$ we have

$$\phi_{a_1}(t) + \phi_{a_2}(t) = \sqrt{\frac{m_{a_1}(t) - d_{a_1}(t)}{m_{a_1}(t) + d_{a_1}(t)}} - \sqrt{\frac{m_{a_2}(t) - d_{a_2}(t)}{m_{a_2}(t) + d_{a_2}(t)}}.$$

  Hence, by Lemma 7.2.2, t_0 is a zero of the polynomial

$$\begin{aligned} p(t) &:= (\, m_{a_1}(t)\, d_{a_2}(t) - m_{a_2}(t)\, d_{a_1}(t)\,)\, (1/2) t^{-2} \\ &= (\, |a_1|^4 - 2(\Re a_1)^2 t^2 + t^4)\, (\Im a_2)^2 - (\, |a_2|^4 - 2(\Re a_2)^2 t^2 + t^4)\, (\Im a_1)^2 \\ &= ((\Im a_2)^2 - (\Im a_1)^2)\, t^4 \; - \; 2((\Re a_1)^2 (\Im a_2)^2 - (\Re a_2)^2 (\Im a_1)^2)\, t^2 \\ &\quad + |a_1|^4 (\Im a_2)^2 - |a_2|^4 (\Im a_1)^2 \end{aligned}$$

Let $\alpha = (\Im a_2)^2 - (\Im a_1)^2$. If $\alpha = 0$ then we obtain

$$\begin{aligned} t_0^2 &= \frac{1}{2}\frac{|a_1|^4(\Im a_2)^2 - |a_2|^4(\Im a_1)^2}{(\Re a_1)^2(\Im a_2)^2 - (\Re a_2)^2(\Im a_1)^2} \\ &= \frac{1}{2}\frac{|a_1|^4 - |a_2|^4}{(\Re a_1)^2 - (\Re a_2)^2} \\ &= \frac{1}{2}\frac{(\Re a_1)^4 - (\Re a_2)^4 + 2((\Re a_1)^2 - (\Re a_2)^2)(\Im a_1)^2}{(\Re a_1)^2 - (\Re a_2)^2} \\ &= (1/2)((\Re a_1)^2 + (\Re a_2)^2 + 2(\Im a_1)^2) \\ &= (1/2)(|a_1|^2 + |a_2|^2) \end{aligned}$$

If $\alpha \neq 0$ then t_0^2 equals one of the numbers

$$\theta_k = \beta + (-1)^k\sqrt{\beta^2 - \gamma}, \quad k = 1, 2,$$

where $\beta = \alpha^{-1}((\Re a_1)^2(\Im a_2)^2 - (\Re a_2)^2(\Im a_1)^2)$, $\gamma = \alpha^{-1}(|a_2|^2(\Im a_1)^2 - |a_1|^2(\Im a_2)^2)$. If $\phi_{a_1}(\beta) + \phi_{a_2}(\beta) < 0$ then $t_0^2 = \theta_1$. Otherwise $t_0^2 = \theta_2$. $\square$

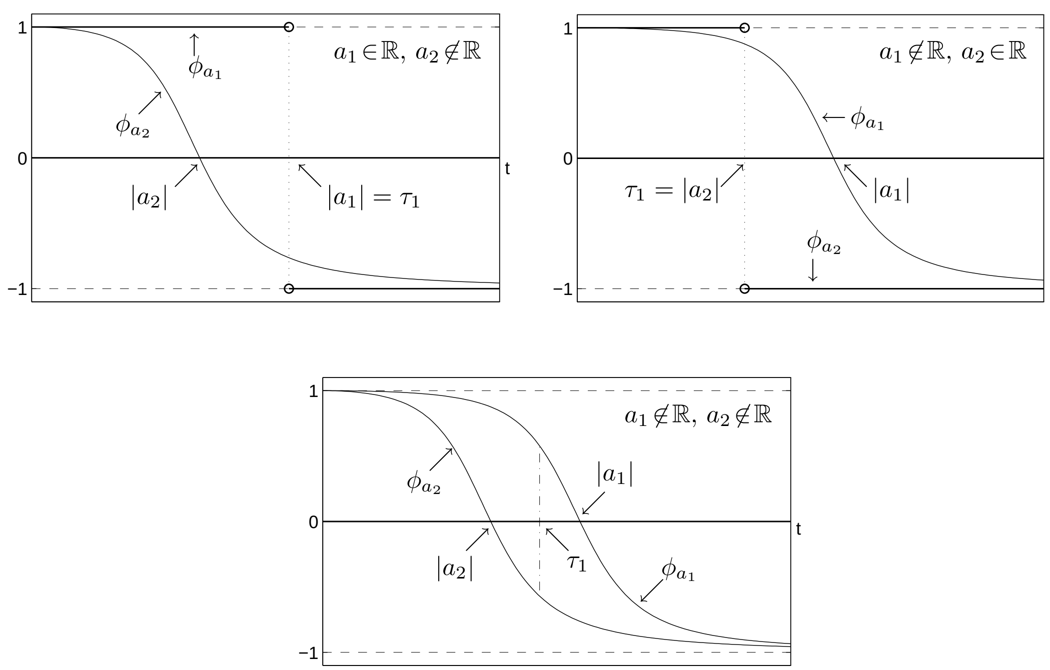


Figure 7.1: Computation of $\tau_1(\mathrm{diag}(a_1, a_2))$.

There is an interesting special case of Proposition 7.2.1:

Corollary 7.2.3 *Let $r_1 \geq r_2 \geq 0$ and $z \in \mathbb{C}$. Then*

$$\tau_1(\mathrm{diag}(zr_1, zr_2)) = \begin{cases} |z|r_1 & \text{if } z \in \mathbb{R}, \\ |z|\sqrt{r_1 r_2} & \text{otherwise.} \end{cases}$$

Proof: If $z \in \mathbb{R}$ or $r_2 = 0$ then the result is immediate from Proposition 7.2.1. Otherwise one has to verify that $t_0 := |z|\sqrt{r_1 r_2}$ is the zero of the function

$$t \mapsto \frac{|zr_1|^2 - t^2}{|z^2 r_1^2 - t^2|} + \frac{|zr_2|^2 - t^2}{|z^2 r_2^2 - t^2|}, \qquad t \geq 0.$$

However, we have

$$\frac{|zr_1|^2 - t_0^2}{|z^2 r_1^2 - t_0^2|} + \frac{|zr_2|^2 - t_0^2}{|z^2 r_2^2 - t_0^2|} = \frac{|r_1| - |r_2|}{|wr_1 - r_2|} + \frac{|r_2| - |r_1|}{|wr_2 - r_1|}, \qquad w := z^2/|z^2|,$$

and, since $|w| = 1,\ w^{-1} = \overline{w}$,

$$|wr_1 - r_2| = |\,\overline{wr_1 - r_2}\,| = |w|\,|\overline{w}r_1 - r_2| = |r_1 - wr_2|.$$

□

The next proposition is an extension of Corollary 7.2.3. We have already used it in the proof of Proposition 6.9.4.

Proposition 7.2.4 *Let $R \in \mathbb{R}^{n\times n}$ be a real matrix, and let $z \in \mathbb{C}$. Then*

$$\tau_1(zR) = \begin{cases} |z|\sigma_1(R) & \text{if } z \in \mathbb{R}, \\ |z|\sqrt{\sigma_1(R)\,\sigma_2(R)} & \text{otherwise.} \end{cases}$$

Proof: Let $R = U\Sigma V^T$ be a singular value decomposition of R with $\Sigma = \operatorname{diag}(\sigma_1(R), \ldots, \sigma_n(R))$ and real orthogonal matrices $U, V \in \mathbb{R}^{n\times n}$. Then by Theorem 7.1.1 and the invariance of the real perturbation values under multiplication with real orthogonal matrices,

$$\tau_1(zR) = \tau_1(z\Sigma) = \max_{1\leq j<k\leq n} \tau_1(\operatorname{diag}(z\sigma_j(R), z\sigma_k(R))).$$

Now, Corollary 7.2.3 yields the result. □

7.3 Examples

In this section we give some examples of real spectral value sets with respect to spectral norm,

$$\sigma_{\mathbb{R}}^{(2)}(A, B, C, \rho) = \sigma(A) \cup \{\, s \in \mathbb{C}\backslash\sigma(A) \mid \ \tau_1(G(s)) \geq \rho^{-1} \,\}, \qquad G(s) = C(sI - A)^{-1}B.$$

Moreover, we compare them with their complex counterparts

$$\sigma_{\mathbb{C}}^{(2)}(A, B, C, \rho) = \sigma(A) \cup \{\, s \in \mathbb{C}\backslash\sigma(A) \mid \ \sigma_1(G(s)) \geq \rho^{-1} \,\}.$$

The examples are based on the results of the last section.

In the following let $R \in \mathbb{R}^{q\times l}\backslash\{0\}$ be a real matrix and let $g(s) \in \mathbb{C}$ be a strictly proper rational function. Moreover, let $(A, B, C) \in L_{n',l',q'}$ be a minimal realization of the rational matrix function $G(s) := g(s)R$, i.e. $G(s) = C(sI - A)^{-1}B$. Then the realness locus of $G(s)$ is $\Re_G = \{\ s \in \mathbb{C} \mid \ \Im(g(s)) = 0\ \}$. Thus, $\Re_G$ is a real algebraic curve with finitely many points removed. By Proposition 7.2.4 we have

$$\tau_1(G(s)) \;=\; \begin{cases} |g(s)|\sigma_1(R) = \sigma_1(G(s)) & \text{if } s\in\Re_G, \\ |g(s)|\sqrt{\sigma_1(R)\sigma_2(R)} & \text{otherwise.} \end{cases} \tag{7.13}$$

Thus

$$\begin{aligned} \sigma_{\mathbb{C}}^{(2)}(A,B,C,\rho) &= \sigma(A)\cup\{\, s\in\mathbb{C}\setminus\sigma(A) \mid \ |g(s)|\sigma_1(R)\geq\rho^{-1}\,\}, \\ \sigma_{\mathbb{R}}^{(2)}(A,B,C,\rho) &= \sigma(A)\cup\{\, s\in\Re_G \mid \ |g(s)|\sigma_1(R)\geq\rho^{-1}\,\} \\ &\qquad \cup\{\, s\in\mathbb{C}\setminus(\sigma(A)\cup\Re_G) \mid \ |g(s)|\sqrt{\sigma_1(R)\sigma_2(R)}\geq\rho^{-1}\,\}. \end{aligned}$$

Thus, we have $\sigma_{\mathbb{R}}^{(2)}(A,B,C,\rho)=\mathcal{I}(R,\rho)\cup\mathcal{B}(R,\rho)$, where $\mathcal{I}(R,\rho):=\Re_G\cap\sigma_{\mathbb{C}}^{(2)}(A,B,C,\rho)$ is the intersection of the complex spectral value set with the realness locus of G and

$$\mathcal{B}(R,\rho):=\{\, s\in\mathbb{C}\setminus(\sigma(A)\cup\Re_G) \mid \ |g(s)|\sqrt{\sigma_1(R)\sigma_2(R)}\geq\rho^{-1}\}.$$

It can be shown that $\partial\mathcal{B}(R,\rho)=\partial(\operatorname{int}\mathcal{B}(R,\rho))$ and $\operatorname{int}(\sigma_{\mathbb{R}}^{(2)}(A,B,C,\rho))\subseteq\operatorname{cl}(\mathcal{B}(R,\rho))$. In our concrete examples we consider the following matrices and rational functions.

$$\begin{aligned} R_\kappa &:= \operatorname{diag}(1,\kappa), \qquad \kappa\in[0,1], \\ g_r(s) &:= \frac{((s+\frac{1}{2})^2+1)((s+2)^2+1)+r}{s((s-1)^2+1)((s+1)^2+4)}, \qquad r\geq 0. \end{aligned}$$

Note that $\sigma_1(R_\kappa)=1$ and $\sigma_2(R_\kappa)=\kappa$. Figure 7.2 shows the spectral value sets corresponding to the perturbation level $\rho=2$ and the rational matrix function

$$G(s)=g_0(s)R_{1/4}=\operatorname{diag}(g_0(s),(1/4)g_0(s)). \tag{7.14}$$

The left picture of Figure 7.2 is a plot of the realness locus $\Re_G$. The picture in the middle shows the sets $\Re_G$, $\mathcal{B}(R_{1/4},2)$ (shaded region) and the boundary of the complex spectral value set $\sigma_{\mathbb{C}}^{(2)}(A,B,C,2)$ (dashed curve). In the picture on the right one sees the real spectal value set $\sigma_{\mathbb{R}}^{(2)}(A,B,C,2)$. The crosses mark the poles of g_0.

In Figure 7.3 we vary the second singular value κ of R_κ. The corresponding rational functions are

$$G_\kappa(s):=g_0(s)R_\kappa=\operatorname{diag}(\,g_0(s),\kappa\,g_0(s)\,). \tag{7.15}$$

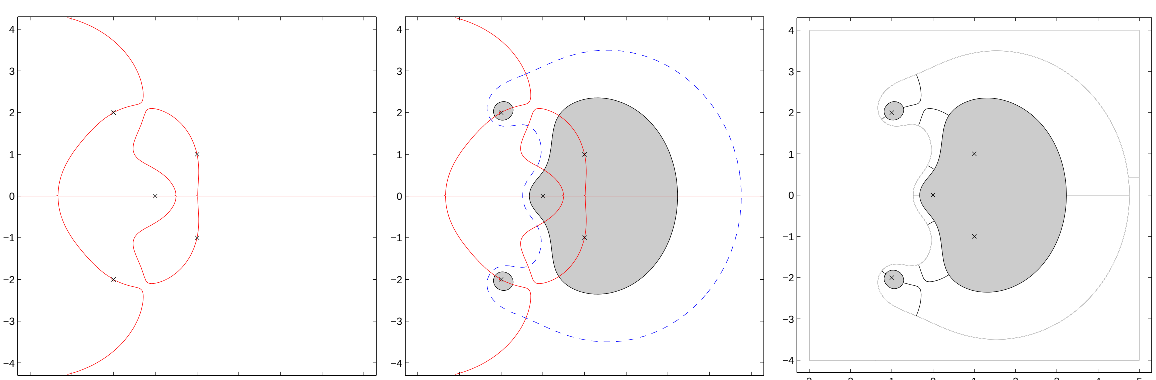


Figure 7.2: Real and spectral values sets to the rational matrix function $G(s)$ defined in (7.14).

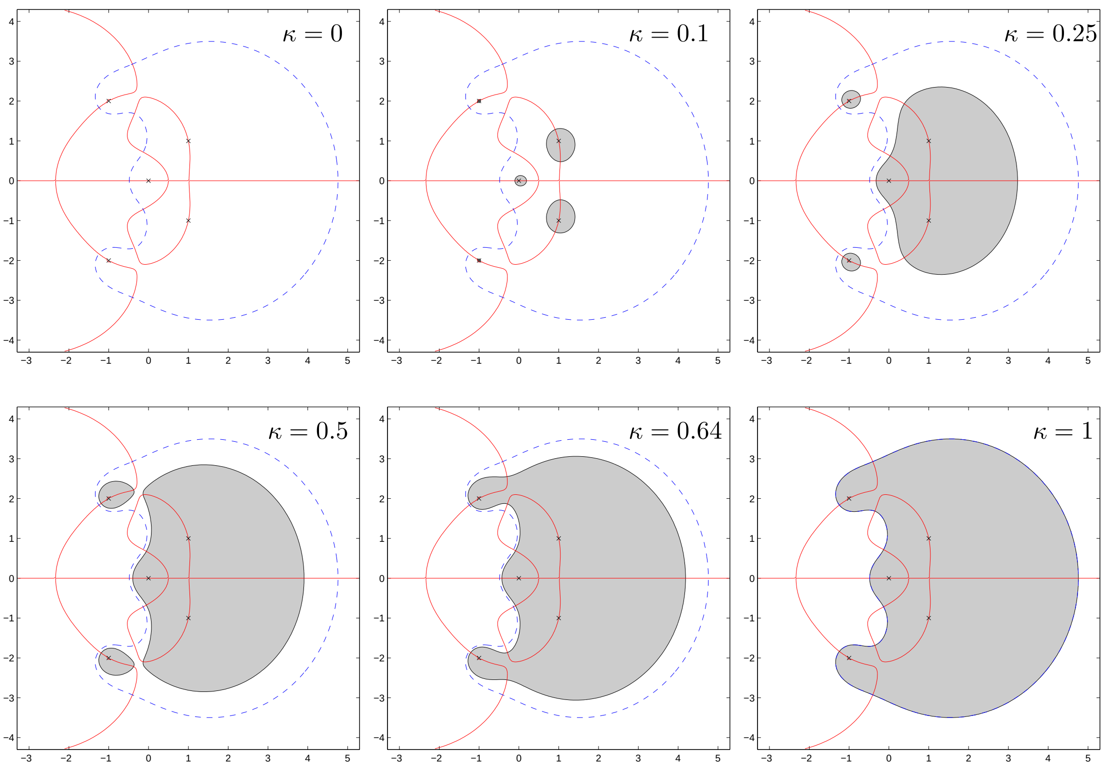


Figure 7.3: Real and complex spectral value sets to the functions $G_\kappa(s)$ defined in (7.15).

The pictures in Figure 7.3 show the sets $\Re_{G_\kappa}$, $\mathcal{B}(R_\kappa, 2)$ and the boundary

$$\partial\sigma_{\mathbb{C}}^{(2)}(A_\kappa, B_\kappa, C_\kappa) = \{ s \in \mathbb{C} \mid \; G_\kappa(s) = 1/2 \} \quad \text{(dashed curve)},$$

where $\kappa \in \{0, 0.1, 0.25, 0.5, 0.64, 1\}$ and $(A_\kappa, B_\kappa, C_\kappa, 2)$ is a minimal realization of G_κ. Note that the complex spectral value set corresponding to G_κ does not depend on κ. Note further that

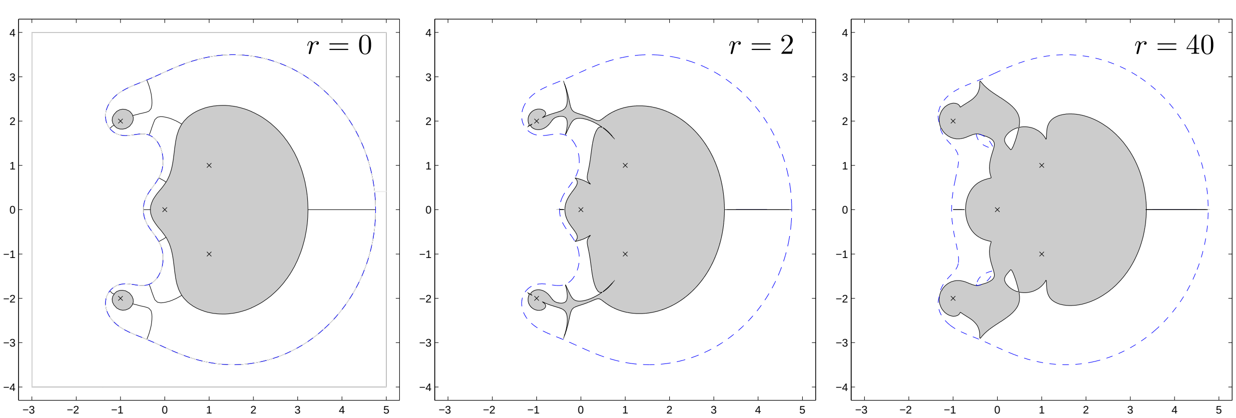


Figure 7.4: Real and complex spectral value sets to the functions $G^r(s)$ defined in (7.16).

the corresponding real spectral value set coincides with the complex one if $\kappa = 1$. It coincides with the $\mathcal{I}(R_\kappa, 2)$ if $\kappa = 0$.

In Figure 7.4 we vary the rational function in the second component of G. More precisely, the pictures in Figure 7.4 show the real spectral value sets and the boundaries of its complex counterparts (dashed curves) for the perturbation level $\rho = 2$ and the rational functions

$$G^r(s) = \operatorname{diag}(g_0(s), (1/4)g_r(s)), \tag{7.16}$$

where $r = 0, 2, 40$. Note that the formula (7.13) is not applicable to these examples (for $\kappa \neq 0$). The plots have been generated using the Proposition 7.2.1.

7.4 Normal matrices

In this section we determine the real pseudospectra of real normal matrices. The underlying norm is the spectral norm. Recall that by Corollary 5.3.8 a complex pseudospectrum of a normal matrix $A \in \mathbb{C}^{n\times n}$ is just a union of disks: $\sigma_{\mathbb{C}}^{(2)}(A, \rho) = \bigcup_{\lambda\in\sigma(A)} \mathcal{D}(\lambda, \rho)$. The structure of *real* pseudospectra of normal matrices is more complicated. However, it turns out that each real pseudospectrum of a real normal matrix is a union of disks, of pseudospectra of 2×2 diagonal matrices, and of pseudospectra of the form $\sigma_{\mathbb{R}}^{(2)}(N(a, b, \lambda), \rho)$, where

$$N(a, b, \lambda) = \begin{bmatrix} a & -b & 0 \\ b & a & 0 \\ 0 & 0 & \lambda \end{bmatrix}, \qquad a, b, \lambda \in \mathbb{R}.$$

More precisely, the following holds.

Theorem 7.4.1 *Let $A \in \mathbb{R}^{n\times n}$, $n \geq 2$, be normal with real eigenvalues $\lambda_1, \ldots, \lambda_p$ and nonreal eigenvalues $a_1 \pm ib_1, \ldots, a_q \pm ib_q$ $(a_j, b_j \in \mathbb{R}, b_j > 0)$ counting multiplicities. Then*

$$\begin{aligned} \sigma_{\mathbb{R}}^{(2)}(A, \rho) \quad &= \bigcup_{1\leq j<k\leq p} \sigma_{\mathbb{R}}^{(2)}(\operatorname{diag}(\lambda_j, \lambda_k), \rho) \cup \bigcup_{1\leq j<k\leq q} (\mathcal{D}(a + ib, \rho) \cup \mathcal{D}(a - ib, \rho)) \\ &\cup \bigcup_{\substack{1\leq j\leq q \\ 1\leq k\leq p}} \sigma_{\mathbb{R}}^{(2)}(N(a_j, b_j, \lambda_k), \rho). \end{aligned}$$

If A has no real (resp. nonreal) eigenvalues then the corresponding unions are not present.

Proof: Since A is real and normal there is a real orthogonal matrix Q such that

$$Q^{-1}AQ = \operatorname{diag}(\lambda_1, \ldots, \lambda_p) \oplus \bigoplus_{j\in\underline{q}} \begin{bmatrix} a_j & -b_j \\ b_j & a_j \end{bmatrix}.$$

However, real pseudospectra with respect to spectral norm are invariant under similarity transformations with real orthogonal matrices. Hence, $\sigma_{\mathbb{R}}^{(2)}(A, \rho) = \sigma_{\mathbb{R}}^{(2)}\big(Q^{-1}AQ, \rho\big)$. By applying Corollary 7.1.6 to the block diagonal matrix $Q^{-1}AQ$ we obtain

$$\begin{aligned} \sigma_{\mathbb{R}}^{(2)}(A, \rho) \quad &= \bigcup_{1\leq j<k\leq p} \sigma_{\mathbb{R}}^{(2)}(\operatorname{diag}(\lambda_j, \lambda_k), \rho) \cup \bigcup_{1\leq j<k\leq q} \sigma_{\mathbb{R}}^{(2)}\left(\begin{bmatrix} a_j & -b_j \\ b_j & a_j \end{bmatrix} \oplus \begin{bmatrix} a_k & -b_k \\ b_k & a_k \end{bmatrix}, \rho\right) \\ &\cup \bigcup_{\substack{1\leq j\leq q \\ 1\leq k\leq p}} \sigma_{\mathbb{R}}^{(2)}(N(a_j, b_j, \lambda_k), \rho). \end{aligned}$$

The matrix $\begin{bmatrix} a_j & -b_j \\ b_j & a_j \end{bmatrix} \oplus \begin{bmatrix} a_k & -b_k \\ b_k & a_k \end{bmatrix}$ is permutation similar to $A_{jk} := \begin{bmatrix} \mathrm{diag}(a_j, a_k) & -\mathrm{diag}(b_j, b_k) \\ \mathrm{diag}(b_j, b_k) & \mathrm{diag}(a_j, a_k) \end{bmatrix}$, and by Corollary 6.8.5 the real and the complex pseudospectra of A_{jk} coincide. Thus

$$\begin{aligned}
\sigma_{\mathbb{R}}^{(2)}\left(\begin{bmatrix} a_j & -b_j \\ b_j & a_j \end{bmatrix} \oplus \begin{bmatrix} a_k & -b_k \\ b_k & a_k \end{bmatrix}, \rho \right) &= \sigma_{\mathbb{R}}^{(2)}(A_{jk}, \rho) \\
&= \sigma_{\mathbb{C}}^{(2)}(A_{jk}, \rho) \qquad \text{(by Corollary 6.8.5)} \\
&= \textstyle\bigcup_{\lambda \in \sigma(A_{jk})} \mathcal{D}(\lambda, \rho) \qquad \text{(by Corollary 5.3.8)} \\
&= \mathcal{D}(a_j + ib_j, \rho) \cup \mathcal{D}(a_j - ib_j, \rho) \\
&\quad \cup \mathcal{D}(a_k + ib_k, \rho) \cup \mathcal{D}(a_k - ib_k, \rho). \qquad \square
\end{aligned}$$

In the remainder of this section we analyze the structure of the sets $\sigma_{\mathbb{R}}^{(2)}(\mathrm{diag}(\lambda_1, \lambda_2), \rho)$ and $\sigma_{\mathbb{R}}^{(2)}(N(a_j, b_j, \lambda_k), \rho)$. To this end we compute the corresponding minimum real perturbation values.

Proposition 7.4.2 *Let $\lambda_1, \lambda_2 \in \mathbb{R}$ and $s \in \mathbb{C}$. Then*

$$\widetilde{\tau}_{\min}(\mathrm{diag}(\lambda_1 - s, \lambda_2 - s)) = \begin{cases} \min\{|s - \lambda_1|, |s - \lambda_2|\} & \text{if } \Im s = 0, \\ \sqrt{\left(\Re s - \frac{\lambda_1 + \lambda_2}{2}\right)^2 + \left(\frac{\lambda_1 - \lambda_2}{2}\right)^2 + (\Im s)^2} & \text{otherwise.} \end{cases}$$

Proof: The case $\Im s = 0$ is trivial since then

$$\widetilde{\tau}_{\min}(\mathrm{diag}(\lambda_1 - s, \lambda_2 - s)) = \sigma_{\min}(\mathrm{diag}(\lambda_1 - s, \lambda_2 - s)).$$

The formula for $\Im s \neq 0$ can be derived from Theorem 7.2.1 using the fact that

$$\widetilde{\tau}_{\min}(\mathrm{diag}(\lambda_1 - s, \lambda_2 - s)) = 1/\tau_1\big(\mathrm{diag}((\lambda_1 - s)^{-1}, (\lambda_2 - s)^{-1}\big).$$

However, we give a proof via Theorem 7.1.5 since we will need a part of this proof later on. We first treat the case when s is pure imaginary.

Let $\lambda, y \in \mathbb{R}$, $y \neq 0$. Then we have for the 1×1-matrix $[\lambda - iy] = \lambda - iy$ that $\widetilde{\tau}_{\min}(\lambda - iy) = \infty$. Moreover, the function $\widetilde{\phi}$ defined in (7.9) satisfies

$$\begin{aligned}
\widetilde{\phi}_{\lambda - iy}(t) &= \max_{x \in \mathbb{C} \setminus \{0\}} \frac{\overline{x}(t^2 - |\lambda - iy|^2)x}{|x(t^2 - (\lambda - iy)^2)x|} = \frac{t^2 - |\lambda - iy|^2}{|t^2 - (\lambda - iy)^2|} \\
&= \begin{cases} \sqrt{\frac{(t^2 - |\lambda - iy|^2)^2}{|t^2 - (\lambda - iy)^2|^2}} = \sqrt{\frac{m_{\lambda,y}(t) - y^2 t^2}{m_{\lambda,y}(t) + y^2 t^2}} & \text{if } t \geq |\lambda - iy| \\ -\sqrt{\frac{(t^2 - |\lambda - iy|^2)^2}{|t^2 - (\lambda - iy)^2|^2}} = -\sqrt{\frac{m_{\lambda,y}(t) - y^2 t^2}{m_{\lambda,y}(t) + y^2 t^2}} & \text{otherwise,} \end{cases}
\end{aligned}$$

where

$$m_{\lambda,y}(t) := t^4 - 2\lambda^2 t^2 + (\lambda^2 + y^2)^2.$$

By differentiating $\phi_{\lambda - iy}$ one shows that $0 \leq t \mapsto \widetilde{\phi}_{\lambda - iy}(t)$ is strictly increasing. Moreover, we have $\widetilde{\phi}_{\lambda - iy}(0) = -1$, $\widetilde{\phi}_{\lambda - iy}(|\lambda - iy|) = 0$, and $\lim_{t \to \infty} \widetilde{\phi}_{\lambda - iy}(t) = 1$.

Let $\lambda_1, \lambda_2 \in \mathbb{R}$ be such that $|\lambda_1 - iy| \geq |\lambda_2 - iy|$. Then Theorem 7.1.5 yields that $t_0 := \widetilde{\tau}_{\min}(\operatorname{diag}(\lambda_1 - iy, \lambda_2 - iy))$ equals the unique zero of the increasing function $0 < t \mapsto \widetilde{\phi}_{\lambda - iy_1}(t) + \widetilde{\phi}_{\lambda_2 - iy}(t)$. If $\lambda_1^2 = \lambda_2^2$ then $t_0 = |\lambda_1 - iy|^2 = \sqrt{\lambda_1^2 + y^2}$. Otherwise t_0 is contained in the open interval $(\,|\lambda_2 - iy|, |\lambda_1 - iy|\,)$. Therefore

$$\begin{aligned} 0 &= \widetilde{\phi}_{\lambda_1 - iy}(t_0) + \widetilde{\phi}_{\lambda_2 - iy}(t_0) \\ &= \sqrt{\frac{m_{\lambda_2,y}(t) - y^2 t_0^2}{m_{\lambda_2,y}(t) + y^2 t_0^2}} - \sqrt{\frac{m_{\lambda_1,y}(t) - y^2 t_0^2}{m_{\lambda_1,y}(t) + y^2 t_0^2}}. \end{aligned}$$

Equivalently (according to Lemma 7.2.2),

$$\begin{aligned} 0 &= m_{\lambda_1,y}(t_0) - m_{\lambda_1,y}(t_0) \\ &= 2(\lambda_2^2 - \lambda_1^2)t_0^2 + (\lambda_2^2 + y^2)^2 - (\lambda_1^2 + y^2)^2 \\ &= 2(\lambda_2^2 - \lambda_1^2)(\,t_0^2 - y^2 - (1/2)(\lambda_1^2 + \lambda_2^2)). \end{aligned}$$

Thus

$$t_0^2 = (\lambda_1^2 + \lambda_2^2) + y^2 = \left(\frac{\lambda_1 + \lambda_2}{2}\right)^2 + \left(\frac{\lambda_1 - \lambda_2}{2}\right)^2 + y^2.$$

On replacing in the computations above y by $\Im s \neq 0$ and λ_k by $\lambda_k - \Re s$ one obtains the formula in the proposition. $\square$

Proposition 7.4.2 yields that the pseudospectrum

$$\sigma_{\mathbb{R}}^{(2)}(\operatorname{diag}(\lambda_1, \lambda_2), \rho) = \{\, s \in \mathbb{C} \mid \widetilde{\tau}_{\min}(\operatorname{diag}(s - \lambda_1, s - \lambda_2)) \leq \rho \,\}$$

is either the union of two intervals or the union of a disk and an interval. The precise statement is as follows. We leave its easy proof to the reader.

Corollary 7.4.3 *Let $\lambda_1, \lambda_2 \in \mathbb{R}$ and $\rho \geq 0$.*

(i) If $\rho \leq \frac{1}{2}(\lambda_1 + \lambda_2)$ then

$$\sigma_{\mathbb{R}}^{(2)}(\operatorname{diag}(\lambda_1, \lambda_2), \rho) = [\lambda_1 - \rho, \lambda_1 + \rho] \cup [\lambda_2 - \rho, \lambda_2 + \rho].$$

(i) If $\rho > \frac{1}{2}(\lambda_1 + \lambda_2)$ then

$$\sigma_{\mathbb{R}}^{(2)}(\operatorname{diag}(\lambda_1, \lambda_2), \rho) = \mathcal{D}\Big(\tfrac{\lambda_1 + \lambda_2}{2},\, r_\rho\Big) \cup [\min\{\lambda_1, \lambda_2\} - \rho, \max\{\lambda_1, \lambda_2\} + \rho],$$

where

$$r_\rho = \sqrt{\rho^2 - \Big(\tfrac{\lambda_1 - \lambda_2}{2}\Big)^2}.$$

Observe that $\mathcal{D}\Big(\frac{\lambda_1 + \lambda_2}{2}, r_\rho\Big)$ is the largest disk with center $\frac{\lambda_1 + \lambda_2}{2}$ lying in the complex pseudospectrum $\sigma_{\mathbb{C}}^{(2)}(\operatorname{diag}(\lambda_1, \lambda_2), \rho) = \mathcal{D}(\lambda_1, \rho) \cup \mathcal{D}(\lambda_2, \rho)$. This is illustrated in Figure 7.5.

It remains to compute the pseudospectra

$$\sigma_{\mathbb{R}}^{(2)}(\,N(a, b, \lambda)\,,\, \rho\,) = \{\, s \in \mathbb{C} \mid \widetilde{\tau}_{\min}(N(a, b, \lambda) - sI_3) \leq \rho \,\}.$$

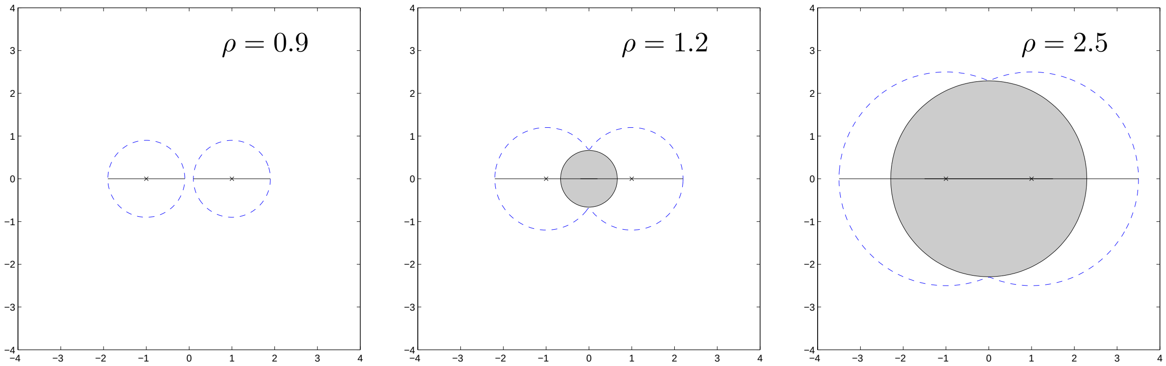


Figure 7.5: The real pseudospectra $\sigma_{\mathbb{C}}^{(2)}(\mathrm{diag}(-1,1),\rho)$, $\rho = 0.9, 1.2, 2.5$. The dashed curves are the boundaries of the corresponding complex pseudospectra.

In the next proposition we use the notation

$$\mathcal{S}(a,b,\lambda) := \{\ s \in \mathbb{C}\ |\ \ |\lambda - s| < \min\{|a+ib-s|, |a-ib-s|\}\ \ \}.$$

In words, $\mathcal{S}(a,b,\lambda)$ is the set of $s \in \mathbb{C}$ whose distance to the real eigenvalue of $N(a,b,\lambda)$ is less then the distance to the nonreal eigenvalues. If $\lambda \neq a$ then $\mathcal{S}(a,b,\lambda)$ is an angle in the complex plane, see Figure 7.6. If $\lambda = a$ then $\mathcal{S}(a,b,\lambda)$ is a stripe symmetric to the real axis.

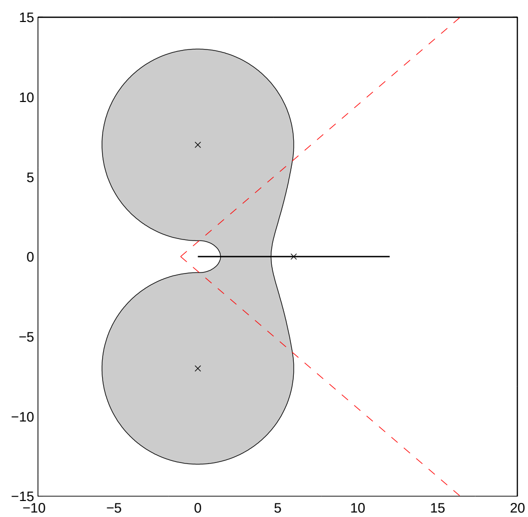


Figure 7.6: The real pseudospectrum $\sigma_{\mathbb{R}}^{(2)}(N(a,b,\lambda),\rho)$ for $a = 0,\ b = 7,\ \lambda = 6,\ \rho = 6.$ The dashed lines are the boundary of $\mathcal{S}(a,b,\lambda)$.

Proposition 7.4.4 *Let $a, b, \lambda \in \mathbb{R}$ and $s \in \mathbb{C}$.*

(*i*) *If $\Im s = 0$ then* $\quad \widetilde{\tau}_{\min}(N(a,b,\lambda) - sI_3) = \min\{|a+ib-s|, |\lambda - s|\}.$

(*ii*) *If $\Im s \neq 0$ and $s \notin \mathcal{S}(a,b,\lambda)$ then*

$$\widetilde{\tau}_{\min}(N(a,b,\lambda) - sI_3) = \min\{|a+ib-s|, |a-ib-s|\}.$$

(iii) If $\Im s \neq 0$ *and* $s \in \mathcal{S}(a,b,\lambda)$,

$$\widetilde{\tau}_{\min}(N(a,b,\lambda) - sI_3) = \sqrt{\frac{(a-\Re s)^2 + b^2 + (\lambda - \Re s)^2}{2} + \frac{(a-\Re s)^2 - b^2 - (\lambda-\Re s)^2}{(a-\Re s)^2 + b^2 - (\lambda - \Re s)^2}(\Im s)^2}.$$

In case (iii) we have $\widetilde{\tau}_{\min}(N(a,b,\lambda) - sI_3) < \min\{\,|a+ib-s|, |a-ib-s|\,\}$.

Corollary 7.4.5 *Let* $a, b, \lambda \in \mathbb{R}$. *Then we have for all* $\rho \geq 0$,

$$\sigma_{\mathbb{R}}^{(2)}(\, N(a,b,\lambda)\, ,\, \rho\,) \supseteq [\lambda - \rho, \lambda + \rho] \cup \mathcal{D}(a+ib,\rho) \cup \mathcal{D}(a-ib,\rho).$$

Equality holds if and only if $\mathcal{D}(a+ib,\rho) \cap \mathcal{S}(a,b,\lambda) = \emptyset$.

A gallery of pseudospectra $\sigma_{\mathbb{R}}^{(2)}(N(a,b,\lambda),\rho)$ computed by means of the formulas in Proposition 7.4.4 is shown in Figure 7.7.

For the proof of Proposition 7.4.4 we need lemma on the quotient of a hermitian and a symmetric form in the 2×2 case.

Lemma 7.4.6 *Let* $H \in \mathbb{C}^{2\times 2}$ *be a hermitian matrix and* $S \in \mathbb{C}^{2\times 2}$ *be a nonsingular symmetric matrix. Then*

(i) *There are linearly independent vectors* $v_1, v_2 \in \mathbb{C}^2$ *such that*

$$\{\, x \in \mathbb{C}^2 \mid\, x^T S x = 0 \,\} = \mathbb{C}v_1 \cup \mathbb{C}v_2. \tag{7.17}$$

(ii) *Let* $v_1, v_2 \in \mathbb{C}^2$ *be linearly independent vectors satisfying* (7.17). *Suppose that* $v_1^* H v_1 < 0$ *and* $v_2^* H v_2 < 0$. *Set* $\frac{x^*Hx}{|x^TSx|} := -\infty$ *for* $x \in \mathbb{C}v_1 \cup \mathbb{C}v_2$, *and*

$$x_0 := \left(w\sqrt{|v_2^* H v_2|}\right) v_1 + \sqrt{|v_1^* H v_1|}\, v_2, \qquad w := \begin{cases} 1 & \text{if } v_1^* H v_2 = 0, \\ v_1^* H v_2 / |v_1^* H v_2| & \text{otherwise.} \end{cases}$$

Then

$$\max_{x \in \mathbb{C}^2 \setminus \{0\}} \frac{x^* H x}{|x^T S x|} = \frac{x_0^* H x_0}{|x_0^T S x_0|} = \frac{|v_1^* H v_2| - \sqrt{(v_1^* H v_1)(v_2^* H v_2)}}{|v_1^T S v_2|}.$$

Proof: (*i*) Let $S = \begin{bmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{bmatrix}$, $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$. Then $x^T S x = s_{11}x_1^2 + 2s_{12}x_1x_2 + s_{22}x_2^2$, and, since S is assumed to be nonsingular, $0 \neq \det(S) = s_{11}s_{22} - s_{12}^2$. Suppose first that $s_{11} = 0$. Then $s_{12} \neq 0$ and $x^T S x = (2s_{12}x_1 + s_{22}x_2)x_2$. Thus, $x^T S x = 0$ if and only if x is a scalar multiple of one of the linearly independent vectors $v_1 = \begin{bmatrix} 1 & 0 \end{bmatrix}^T$, $v_2 = \begin{bmatrix} -s_{22} & 2s_{12} \end{bmatrix}^T$. Assume now, that $s_{11} \neq 0$. Then $x^T S x = 0$ if and only if $x_2 = x_1\zeta_k$, $k = 1$ or $k = 2$, where $\zeta_1, \zeta_2 \in \mathbb{C}$ are the solutions of the quadratic equation $s_{11}\zeta^2 + 2s_{12}\zeta + s_{22} = 0$. The discriminant of this equation coincides with $-\det(S) \neq 0$. Thus $\zeta_1 \neq \zeta_2$. Thus (7.17) holds for the linearly independent vectors $v_k = \begin{bmatrix} 1 & \zeta_k \end{bmatrix}^T$, $k = 1, 2$.

(*ii*). Let v_1,v_2 be vectors as in (*i*). Then each $x \in \mathbb{C}^2$ can be written in the form $x = z_1v_1 + z_2v_2$, $z_k \in \mathbb{C}$, and we have $x^T S x = 2(v_1^T S v_2)z_1z_2$ and $x^* H x = h_1|z_1|^2 + h_2|z_2|^2 + 2\Re(h_3\overline{z_1}z_2)$, where $h_k := v_k^* H v_k$, $k = 1, 2$ and $h_3 := v_1^* H v_2$. Thus, if $z_1 \neq 0 \neq z_2$,

$$\begin{aligned} \frac{x^* H x}{|x^T S x|} &= \frac{h_1|z_1|^2 + h_2|z_2|^2 + 2\Re(h_3\overline{z_1}z_2)}{2|(v_1^T S v_2)z_1z_2|} \\ &= (2|v_1^T S v_2|)^{-1}\left(h_1|z| + h_2|z|^{-1} + 2\,\Re(h_3\,\overline{z}\,|z|^{-1})\right), \qquad z := z_1/z_2. \end{aligned}$$

Since, by assumption, $h_1, h_2 < 0$, the function $0 \neq z \mapsto h_1|z| + h_2|z|^{-1}$ attains its maximum, $-2\sqrt{h_1 h_2}$, at z if $|z| = \sqrt{|h_2|/|h_1|}$. The function $0 \neq z \mapsto \Re(h_3\, \overline{z}\, |z|^{-1})$ attains its maximum, $|h_3|$, at z if $z = \lambda h_3$, $\lambda \geq 0$. This completes the proof of (ii). $\square$

Proof of Proposition 7.4.4: Claim (i) follows from the fact that for real s,

$$\tau_{\min}(N(a,b,\lambda) - sI_3) = \sigma_{\min}(N(a,b,\lambda) - sI_3).$$

In the proof of (ii) and (iii) we first assume that s is purely imaginary, i.e. $s = iy$, $y \in \mathbb{R}\backslash\{0\}$. Set $A_y := \begin{bmatrix} a - iy & -b \\ b & a - iy \end{bmatrix}$. Then $N(a,b,\lambda) - iyI_3 = \operatorname{diag}(A_y, \lambda - iy) = A_y \oplus [\lambda - iy]$. Therefore, by Theorem 7.1.5,

$$\widetilde{\tau}_{\min}(N(a,b,\lambda) - iyI_3) = \begin{cases} \min\{\widetilde{\tau}_{\min}(A_y), \widetilde{\tau}_{\min}(\lambda - iy)\} & \text{if } \widetilde{T} = \emptyset, \\ \min \widetilde{T} & \text{otherwise,} \end{cases}$$

where

$$\widetilde{T} := \{\ 0 \leq t < \min\{\widetilde{\tau}_{\min}(A_y), \widetilde{\tau}_{\min}(\lambda - iy)\} \ \mid\ \widetilde{\phi}_{\lambda - iy}(t) + \widetilde{\phi}_{A_y}(t) \geq 0\ \},$$

and

$$\begin{aligned} \widetilde{\phi}_{\lambda-iy}(t) &= \max_{x \in \mathbb{C}\backslash\{0\}} \frac{\overline{x}(t^2 - |\lambda - iy|^2)x}{|x(t^2 - (\lambda - iy)^2)x|} = \frac{t^2 - |\lambda - iy|^2}{|t^2 - (\lambda - iy)^2|}, \\ \widetilde{\phi}_{A_y}(t) &= \max_{x \in \mathbb{C}^2\backslash\{0\}} \frac{x^* H_t x}{|x^T S_t x|}, \qquad H_t := t^2 I_2 - A_y^* A_y, \quad S_t := t^2 I_2 - A_y^T A_y. \end{aligned}$$

We have $\widetilde{\tau}_{\min}(\lambda - iy) = \infty$. Hence, either $\widetilde{T} = \emptyset$ and $\widetilde{\tau}_{\min}(N(a,b,\lambda) - iyI_3) = \widetilde{\tau}_{\min}(A_y)$, or $\widetilde{T} \neq \emptyset$ and $\widetilde{\tau}_{\min}(N(a,b,\lambda) - iyI_3) = \min \widetilde{T}$. Let us determine $\widetilde{\tau}_{\min}(A_y)$. A straightforward computation yields that for all $t \in \mathbb{R}$,

$$\begin{aligned} S_t &= (t^2 - (a^2 + b^2 - 2ayi - y^2))I_2, \\ H_t &= \begin{bmatrix} t^2 - (a^2 + b^2 + y^2) & 2iby \\ -2iby & t^2 - (a^2 + b^2 + y^2) \end{bmatrix}. \end{aligned}$$

Hence for every $x = [x_1\ x_2]^T \in \mathbb{C}^2$,

$$\begin{aligned} x^T S_t x &= (t^2 - (a^2 + b^2 - 2ayi - y^2))(x_1^2 + x_2^2), \\ x^* H_t x &= (t^2 - (a^2 + b^2 + y^2))(|x_1|^2 + |x_2|^2) + 4\,\Re(iyb\,\overline{x_1}\,x_2) \end{aligned}$$

Suppose that $t^2 < a^2 + b^2 + y^2 - 2|yb| = a^2 + (|b| - |y|)^2$. Then for every $x \neq 0$,

$$\begin{aligned} x^* H_t x &< -2|yb|(|x_1|^2 + |x_2|^2) + 4\,\Re(iyb\,\overline{x_1}\,x_2) \\ &\leq -2|yb|(|x_1|^2 + |x_2|^2) + 4\,|yb|\,|x_1|\,|x_2| \\ &= -2|yb|(|x_1|^2 - |x_2|^2)^2 \\ &\leq 0. \end{aligned}$$

Thus for $t < \sqrt{a^2 + (|b| - |y|)^2}$ the hermitian-symmetric inequality $x^* H_t x \geq |x^T S_t x|$ has no solution $x \neq 0$. Therefore $\widetilde{\tau}_{\min}(A_y) \geq \sqrt{a^2 + (|b| - |y|)^2}$. On the other hand if $t = \sqrt{a^2 + (|b| - |y|)^2}$ and $x = [-sign(yb)\ i]^T$ then $x^* H_t x = 0 = |x^T S_t x|$. Thus

$$\widetilde{\tau}_{\min}(A_y) = \sqrt{a^2 + (|b| - |y|)^2} = \min\{|a + ib - iy|, |a - ib - iy|\}. \tag{7.18}$$

Next, we compute $\widetilde{\phi}_{A_y}(t)$. Set $v_1 := [1\ i]^T$, $v_2 := [1\ -i]^T$. Then the following identities hold for all $t \in \mathbb{R}$.

$$\begin{aligned}
v_1^* S_t v_1 &= 0,\\
v_2^* S_t v_2 &= 0,\\
v_1^* S_t v_2 &= 2|t^2 - (a^2 + b^2 - 2ayi - y^2)|,\\
v_1^* H_t v_1 &= 2(a^2 + b^2 + y^2 - t^2) - 4by = 2(t^2 - (a^2 + (b+y)^2)),\\
v_2^* H_t v_2 &= 2(a^2 + b^2 + y^2 - t^2) + 4by = 2(t^2 - (a^2 + (b-y)^2))\\
v_1^* H_t v_2 &= 0.
\end{aligned}$$

In the sequel let $0 \le t < \widetilde{\tau}_{\min}(A_y)$. Then $\widetilde{\phi}_{A_y}(t)$ is well defined. Moreover,

$$v_k^* H_t v_k < |v_k^* S_t v_k| = 0, \qquad k = 1, 2.$$

If $t^2 = a^2 + b^2 - 2ayi - y^2$ then $S_t = 0$ and therefore $x^* H_t x < |x^T S_t x| = 0$ for all $x \neq 0$. Thus $\widetilde{\phi}_{A_y}(t) = -\infty$. If $t^2 \neq a^2 + b^2 - 2ayi - y^2$ then S_t is nonsingular, and Lemma 7.4.6 yields

$$\widetilde{\phi}_{A_y}(t) = \frac{|v_1^* H_t v_2| - \sqrt{(v_1^* H_t v_1)(v_2^* H_t v_2)}}{|v_1^T S_t v_2|} = -\frac{\sqrt{(v_1^* H_t v_1)(v_2^* H_t v_2)}}{|v_1^T S_t v_2|} < 0.$$

Let us compute an explicite formula for $\widetilde{\phi}_{A_y}(t)$. Set

$$m_{A_y}(t) := t^4 - 2(a^2 + b^2)\, t^2 + (a^4 + b^4 + y^4 + 2a^2 b^2 + 2a^2 y^2 - 2b^2 y^2).$$

Then

$$\begin{aligned}
(1/4)\,(v_1^* H_t v_1)(v_2^* H_t v_2) &= (t^2 - (a^2 + (b+y)^2))(t^2 - (a^2 + (b-y)^2))\\
&= t^4 - 2(a^2 + b^2 + y^2)\, t^2 + (a^2 + (b+y)^2)(a^2 + (b-y)^2)\\
&= t^4 - 2(a^2 + b^2 + y^2)\, t^2 + (a^4 + b^4 + y^4 + 2a^2 b^2 + 2a^2 y^2 - 2b^2 y^2).\\
&= m_{A_y}(t) - 2y^2 t^2.
\end{aligned}$$

and

$$\begin{aligned}
(1/4)\,|v_1^T S_t v_2|^2 &= (t^2 - a^2 - b^2 + y^2)^2 + 4a^2 y^2\\
&= t^4 - 2(a^2 + b^2 - y^2)\, t^2 + (y^2 - a^2 - b^2)^2 + 4a^2 y^2\\
&= t^4 - 2(a^2 + b^2 - y^2)\, t^2 + (a^4 + b^4 + y^4 + 2a^2 b^2 + 2a^2 y^2 - 2b^2 y^2)\\
&= m_{A_y}(t) + 2y^2 t^2.
\end{aligned}$$

Thus,

$$\widetilde{\phi}_{A_y}(t) = -\frac{\sqrt{(v_1^* H_t v_1)(v_2^* H_t v_2)}}{|v_1^T S_t v_2|} = -\sqrt{\frac{m_{A_y}(t) - 2y^2 t^2}{m_{A_y}(t) + 2y^2 t^2}}.$$

Recall from the proof of Proposition 7.4.2 that

$$\widetilde{\phi}_{\lambda - iy}(t) = \begin{cases} \sqrt{\frac{m_{\lambda,y}(t) - y^2 t^2}{m_{\lambda,y}(t) + y^2 t^2}} & \text{if } t \ge |\lambda - iy| \\ -\sqrt{\frac{m_{\lambda,y}(t) - y^2 t^2}{m_{\lambda,y}(t) + y^2 t^2}} & \text{otherwise,} \end{cases}$$

where

$$m_{\lambda,y}(t) := t^4 - 2\lambda^2 t^2 + (\lambda^2 + y^2)^2.$$

We have $\widetilde{\phi}_{\lambda - iy}(t) + \widetilde{\phi}_{A_y}(t) < 0$ for $t < \min\{\,|\lambda - iy|, \widetilde{\tau}_{\min}(A_y)\}$. Now there are two cases:

- Suppose $|\lambda - iy| \geq \widetilde{\tau}_{\min}(A_y)$. Then $\widetilde{\phi}_{A_y}(t) + \widetilde{\phi}_{\lambda - iy}(t) < 0$ for all $t < \widetilde{\tau}_{\min}(A_y)$. Thus, $\widetilde{T} = \emptyset$ and

$$\widetilde{\tau}_{\min}(N(a,b,\lambda) - iyI_3) = \widetilde{\tau}_{\min}(A_y) = \min\{|a+ib-iy|, |a-ib-iy|\}. \tag{7.19}$$

- Suppose, $|\lambda - iy| < \widetilde{\tau}_{\min}(A_y)$. Then $\widetilde{T} \neq \emptyset$ and $\min \widetilde{T} = \widetilde{\tau}_{\min}(N(a,b,\lambda) - iyI_3) =: t_0$. The number t_0 is the smallest zero of the function

$$[\,|\lambda - iy|, \widetilde{\tau}_{\min}(A_y)\,) \ni t \longmapsto \widetilde{\phi}_{\lambda - iy}(t) + \widetilde{\phi}_{A_y}(t) = \sqrt{\frac{m_{\lambda,y}(t) - y^2t^2}{m_{\lambda,y}(t) + y^2t^2}} - \sqrt{\frac{m_{A_y}(t) - 2y^2t^2}{m_{A_y}(t) + 2y^2t^2}}.$$

By Lemma 7.2.2 the zeros of this function are zeros of the polynomial $t \mapsto (m_{A_y}(t) - m_{\lambda,y}(t))y^2t^2$. Since $t_0 \neq 0$, it follows that

$$\begin{aligned} 0 &= m_{A_y}(t_0) - m_{\lambda,y}(t_0) \\ &= -2(a^2+b^2-\lambda^2)t_0^2 + a^4 + b^4 - \lambda^4 + 2a^2b^2 + 2(a^2-b^2-\lambda^2)y^2 \\ &= -2(a^2+b^2-\lambda^2)t_0^2 + (a^2+b^2-\lambda^2)(a^2+b^2+\lambda^2) + 2(a^2-b^2-\lambda^2)y^2. \end{aligned}$$

Thus,

$$\widetilde{\tau}_{\min}(N(a,b,\lambda) - iyI_3) = t_0^2 = \frac{a^2+b^2+\lambda^2}{2} + \frac{a^2-b^2-\lambda^2}{a^2+b^2-\lambda^2}\, y^2. \tag{7.20}$$

The latter equation completes the proof of the proposition for the case that $s = iy$ is pure imaginary. However, we have for arbitrary $s \in \mathbb{C}$,

$$N(a,b,\lambda) - sI_3 = N(a - \Re s, b, \lambda - \Re s) - (\Im s)I_3.$$

Thus we obtain the full statement of the proposition on replacing in (7.4) and (7.20) y by $\Im s$, a by $a - \Re s$ and λ by $\lambda - \Re s$. □

7.5 The 2×2 case

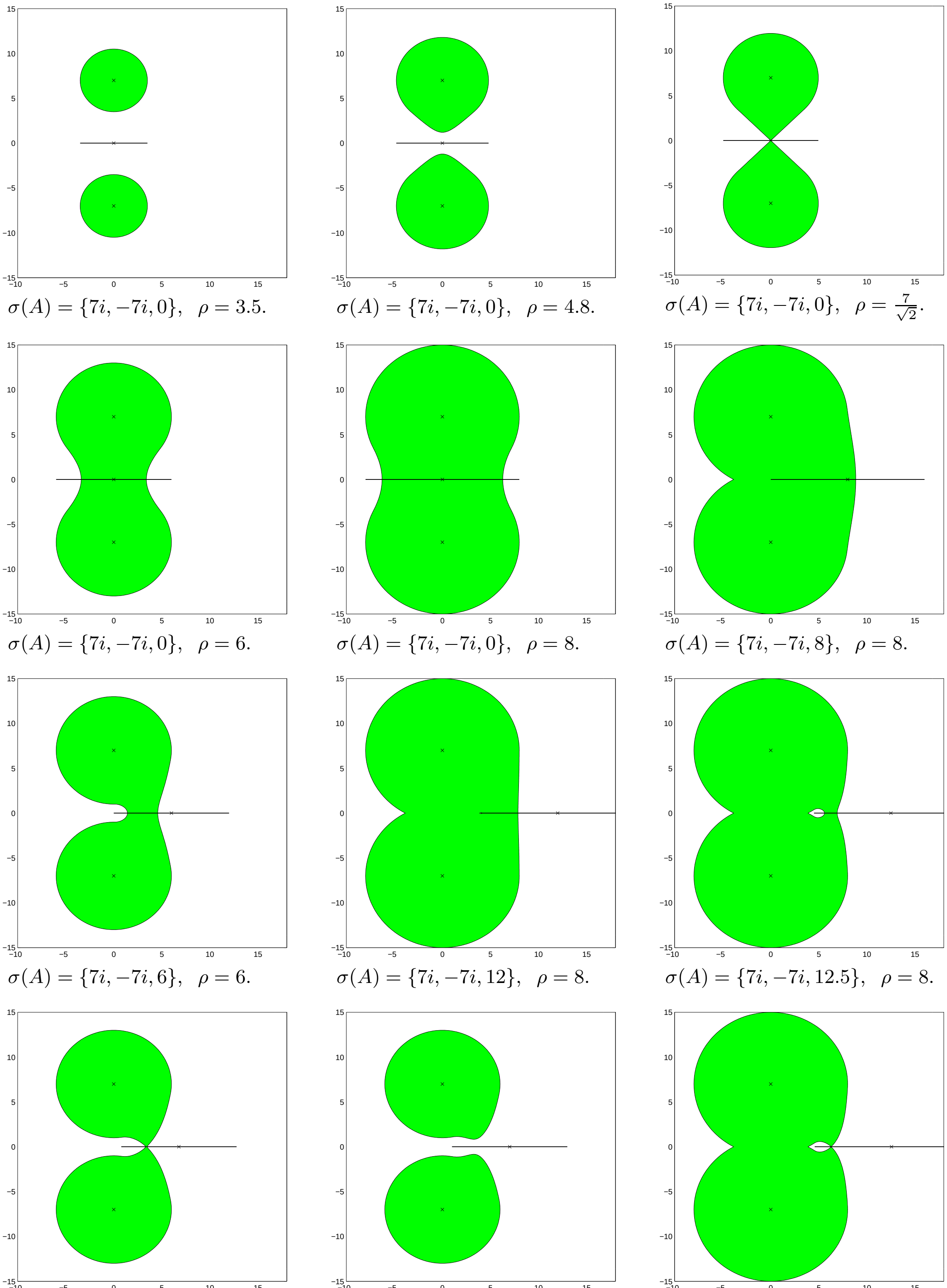
In this section we discuss real pseudospectra of real 2×2 matrices with respect to unitarily invariant norms. These sets yield lower bounds for the pseudospectra of matrices of larger size. More generally, we have the following result whose proof is analogous to the proof of Corollary 5.2.8.

Proposition 7.5.1 *Let $V \in \mathbb{R}^{n\times k}$ be a* real *matrix whose columns form an orthonormal basis of an invariant subspace of $A \in \mathbb{R}^{n\times n}$. Then we have for every unitarily invariant norm and every $\rho \geq 0$,*

$$\sigma_{\mathbb{R}}^{\|\cdot\|}(V^TAV, \rho) \subseteq \sigma_{\mathbb{R}}^{\|\cdot\|}(A, \rho).$$

The Proposition below treats the case of 2-dimensional invariant subspaces.

Proposition 7.5.2 *Let $A \in \mathbb{R}^{n\times n}$ and $\rho \geq 0$.*



$\sigma(A) = \{7i, -7i, 0\}, \quad \rho = 3.5.$ $\sigma(A) = \{7i, -7i, 0\}, \quad \rho = 4.8.$ $\sigma(A) = \{7i, -7i, 0\}, \quad \rho = \frac{7}{\sqrt{2}}.$

$\sigma(A) = \{7i, -7i, 0\}, \quad \rho = 6.$ $\sigma(A) = \{7i, -7i, 0\}, \quad \rho = 8.$ $\sigma(A) = \{7i, -7i, 8\}, \quad \rho = 8.$

$\sigma(A) = \{7i, -7i, 6\}, \quad \rho = 6.$ $\sigma(A) = \{7i, -7i, 12\}, \quad \rho = 8.$ $\sigma(A) = \{7i, -7i, 12.5\}, \quad \rho = 8.$

$\sigma(A) = \{7i, -7i, 6.78\}, \quad \rho = 6.$ $\sigma(A) = \{7i, -7i, 7\}, \quad \rho = 6.$ $\sigma(A) = \{7i, -7i, 12.57\}, \quad \rho = 8.$

Figure 7.7: Real spectral value sets $\sigma_{\mathbb{R}}(A, \rho)$, where $A \in \mathbb{R}^{3\times 3}$ is normal.

(*i*) *Let* $x \in \mathbb{R}^n \setminus \{0\}$ *be a generalized eigenvector of* A *such that* $(A - \lambda I_n)^{\nu-1} x \neq 0$ *and*

$(A-\lambda I_n)^\nu x = 0$, $\lambda \in \mathbb{R}$, $\nu \geq 2$. *Then*

$$\sigma_{\mathbb{R}}^{\|\cdot\|}\left(\begin{bmatrix}\lambda & \delta_x \\ 0 & \lambda\end{bmatrix}, \rho\right) \subseteq \sigma_{\mathbb{R}}^{\|\cdot\|}(A,\rho), \tag{7.21}$$

where

$$\delta_x := \left(\frac{\|(A-\lambda I_n)^{\nu-2}x\|_2^2}{\|(A-\lambda I_n)^{\nu-1}x\|_2^2} - \frac{|\,((A-\lambda I_n)^{\nu-2}x)^T(A-\lambda I_n)^{\nu-1}x\,|^2}{\|(A-\lambda I_n)^{\nu-1}x\|_2^4}\right)^{-\frac{1}{2}}.$$

(*ii*) *Let $x_1, x_2 \in \mathbb{R}^n \setminus \{0\}$ be eigenvectors of A to the eigenvalues $\lambda_1, \lambda_2 \in \mathbb{R}$. If $\lambda_1 \neq \lambda_2$ then*

$$\sigma_{\mathbb{R}}^{\|\cdot\|}\left(\begin{bmatrix}\lambda_1 & \delta_{x_1,x_2} \\ 0 & \lambda_2\end{bmatrix}, \rho\right) \subseteq \sigma_{\mathbb{R}}^{\|\cdot\|}(A,\rho),$$

where

$$\delta_{x_1,x_2} := |\lambda_1 - \lambda_2| \cot\phi, \qquad \phi := \arccos\frac{|x_1^T x_2|}{\|x_1\|_2\|x_2\|_2}.$$

(*iii*) *Let $x \in \mathbb{C}^n \setminus \{0\}$ be an eigenvector of A to the nonreal eigenvalue $a+ib$ ($a, b \in \mathbb{R}$, $b > 0$). Then*

$$\sigma_{\mathbb{R}}^{\|\cdot\|}\left(\begin{bmatrix}a & -\theta_x b \\ \theta_x^{-1} b & a\end{bmatrix}, \rho\right) \subseteq \sigma_{\mathbb{R}}^{\|\cdot\|}(A,\rho),$$

where

$$\theta_x := \sqrt{\frac{\|x\|_2^2 + |x^T x|}{\|x\|_2^2 - |x^T x|}}.$$

Proof: The proofs of (i) and (ii) are analogous to the proofs of Propositions 5.5.6 and 5.5.7. We therefore only prove claim (iii). Let $x \neq 0$ be an eigenvector of A to the eigenvalue $a+ib$, $b \neq 0$. Then $\overline{x}$ is an eigenvector of A to the eigenvalue $a-ib$. Thus x and $\overline{x}$ are linearly independent. Hence, by the Cauchy-Schwarz-inequality, $|x^T x| = |\overline{x}^* x| < \|\overline{x}\|_2\|x\|_2 = \|x\|_2^2$. Thus θ_x is well defined. The continuous functions $z \mapsto g_1(z) := \Im(z^2\, x^T x)$, $z \mapsto g_2(z) := \Re(z^2\, x^T x)$, $z \in \mathbb{C}$, satisfy $g_k(iz) = -g_k(z)$, $g_k(-z) = g_k(z)$. Thus there is $z \in \mathbb{C} \setminus \{0\}$ such that $\Im(z^2\, x^T x) = 0$ and $\Re(z^2\, x^T x) \geq 0$. Set $x_0 := zx$ and $x_1 := \Re x_0$ and $x_2 := \Im x_0$. Then $x_1^T x_2 = (1/2)\Im(x_0^T x_0) = 0$ and $\|x_1\|_2^2 - \|x_2\|_2^2 = \Re(x_0^T x_0) \geq 0$. Thus $|x_0^T x_0| = \|x_1\|_2^2 - \|x_2\|_2^2$ and

$$\theta_x^2 = \frac{\|x\|_2^2 + |x^T x|}{\|x\|_2^2 - |x^T x|} = \frac{\|x_0\|_2^2 + |x_0^T x_0|}{\|x_0\|_2^2 - |x_0^T x_0|} = \frac{\|x_1\|_2^2 + \|x_2\|_2^2 + (\|x_1\|_2^2 - \|x_2\|_2^2)}{\|x_1\|_2^2 + \|x_2\|_2^2 - (\|x_1\|_2^2 - \|x_2\|_2^2)} = \left(\frac{\|x_1\|_2}{\|x_2\|_2}\right)^2.$$

From the relation $Ax_0 = (a+ib)x_0$ we obtain

$$Ax_1 = ax_1 - bx_2, \qquad Ax_2 = bx_1 + ax_2.$$

Let $v_1 := x_1/\|x_1\|_2$, $v_2 := -x_2/\|x_2\|_2$. Then v_1, v_2 form an orthonormal basis of the A-invariant subspace span$\{x_1, x_2\}$, and we have

$$Av_1 = av_1 + \theta_x^{-1} bv_2, \qquad Av_2 = -\theta_x\, bv_1 + av_2.$$

Thus, $V^*AV = \begin{bmatrix}a & -\theta_x b \\ \theta_x^{-1} b & a\end{bmatrix}$, where $V = [v_1 \;\; v_2]$. □

If $A \in \mathbb{R}^{2\times 2}$ then the inclusions in Proposition 7.5.2 become equalities. The matrices $\begin{bmatrix}\lambda_1 & \delta \\ 0 & \lambda_2\end{bmatrix}$, $\begin{bmatrix}a & -\theta b \\ \theta^{-1} & b\end{bmatrix}$ are the real Schur forms of real 2×2 matrices. More precisely, the following holds.

(*i*) If $A\in\mathbb{R}^{2\times 2}$ has different real eigenvalues λ_1,λ_2 or a real eigenvalue $\lambda_1=\lambda_2$ of algebraic multiplicity 2 then there is an orthogonal matrix $V\in\mathbb{R}^{2\times 2}$ such that $V^TAV=\begin{bmatrix}\lambda_1 & \delta\\ 0 & \lambda_2\end{bmatrix}$, $\delta\geq 0$.

(*ii*) If $A\in\mathbb{R}^{2\times 2}$ has conjugate complex eigenvalues $a\pm ib$, $b\neq 0$, then there is an orthogonal matrix $V\in\mathbb{R}^{2\times 2}$ and such that $V^TAV=\begin{bmatrix}a & -\theta_A b\\ \theta_A^{-1}b & a\end{bmatrix}$, $\theta_A\geq 1$.

The number δ in case (*i*) coincides with the departure of normality of A defined in (5.11), i.e.

$$\delta=\delta(A)=\sqrt{\mathrm{tr}(A^TA)-(1/2)(\mathrm{tr}(A)^2+|\mathrm{tr}(A)^2-4\mathrm{det}(A)|)},\tag{7.22}$$

If A has nonreal eigenvalues then by calculating the right hand side of (7.22) one obtains the following relation between $\delta(A)$ and θ_A.

$$\delta(A)=|b|\sqrt{\theta_A^2+\theta_A^{-2}-2}\,.\tag{7.23}$$

Thus $\delta(A)$ is an increasing function of θ_A, and $\theta_A=g((\delta(A)/|b|)^2+2)$ where $g(\cdot)$ is the inverse of the strictly increasing function $\theta\mapsto\theta^2+\theta^{-2}$, $\theta\geq 1$.

The next Theorem is the main result of this section. It shows how the function

$$s\longmapsto\widetilde{\mu}_{\mathbb{R}}^{\|\cdot\|}(A-sI_2),\qquad A\in\mathbb{R}^{2\times 2},\ s\in\mathbb{C}$$

can be calculated by solving a 1-parameter minimization problem provided that the underlying norm is unitarily invariant.

Theorem 7.5.3 *Let $\|\cdot\|$ be an a unitarily invariant norm on $\mathbb{C}^{2\times 2}$ and let $|\!|\!|\cdot|\!|\!|$ be the absolute and symmetric norm on $\mathbb{R}^{2\times 2}$ such that $\|X\|=|\!|\!|\,[\sigma_{\max}(X),\sigma_{\min}(X)]^T\,|\!|\!|$ for all $X\in\mathbb{R}^{2\times 2}$. Let $A=\begin{bmatrix}a & c\\ b & d\end{bmatrix}\in\mathbb{R}^{2\times 2}$ and $s\in\mathbb{C}$.*

(*i*) *If $s\in\mathbb{R}$ then*

$$\widetilde{\mu}_{\mathbb{R}}^{\|\cdot\|}(A-sI_2)\;=\;\sigma_{\min}(A-sI_2)\;=\;\left|\sqrt{\left(s-\tfrac{a+d}{2}\right)^2+\left(\tfrac{b-c}{2}\right)^2}-\sqrt{\left(\tfrac{a-d}{2}\right)^2+\left(\tfrac{b+c}{2}\right)^2}\,\right|$$

(*ii*) *Suppose $s\notin\mathbb{R}$. For $\eta\in\mathbb{R}\setminus(-\Im s,\Im s)$ set*

$$\begin{aligned}
f(\eta)&:=\sqrt{\left(\Re s-\tfrac{a+d}{2}\right)^2+\left(\eta-\tfrac{b-c}{2}\right)^2},\\
g(\eta)&:=\left|\sqrt{\eta^2-(\Im s)^2}-\sqrt{\left(\tfrac{a-d}{2}\right)^2+\left(\tfrac{b+c}{2}\right)^2}\right|,\\
\psi(\eta)&:=\left|\!\left|\!\left|\begin{bmatrix}f(\eta)+g(\eta)\\ f(\eta)-g(\eta)\end{bmatrix}\right|\!\right|\!\right|.
\end{aligned}$$

Moreover, let

$$R\;:=\;\max\left\{\tfrac{|b-c|}{2},\,\left(\tfrac{a-d}{2}\right)^2+\left(\tfrac{b+c}{2}\right)^2+(\Im s)^2\right\}.$$

Then we have
$$\widetilde{\mu}_{\mathbb{R}}^{\|\cdot\|}(A - sI_2) = \min\{\, \psi(\eta) \mid |\Im s| \le |\eta| \le R \,\}.$$
Let $\eta_0 \in \mathbb{R}$ *be such that* $|\Im s| \le |\eta_0| \le R$ *and* $\psi(\eta_0) = \widetilde{\mu}_{\mathbb{R}}^{\|\cdot\|}(A - sI_2)$. *Then a real matrix* Δ_0 *with* $\|\Delta_0\| = \widetilde{\mu}_{\mathbb{R}}^{\|\cdot\|}(A - sI_2)$ *and* $s \in \sigma(A + \Delta_0))$ *is given by*
$$\Delta_0 = \begin{bmatrix} q(\eta_0)\,\frac{a-d}{2} & q(\eta_0)\,\frac{b+c}{2} - \eta_0 \\ q(\eta_0)\,\frac{b+c}{2} + \eta_0 & -q(\eta_0)\,\frac{a-d}{2} \end{bmatrix} - A + (\Re s) I_2,$$
where
$$q(\eta_0) := \sqrt{\eta_0^2 - (\Im s)^2} \Big/ \sqrt{\left(\tfrac{a-d}{2}\right)^2 + \left(\tfrac{b+c}{2}\right)^2}.$$

By means of Theorem 7.5.3 real pseudspectra of 2×2 matrices can be computed numerically. Figure 7.8 shows some examples for different norms. For the spectral norm the computation can be simplified by using the following explicite formula which generalizes the result of Proposition 7.4.2.

Theorem 7.5.4 *Let* $A = \begin{bmatrix} a & c \\ b & d \end{bmatrix} \in \mathbb{R}^{2\times 2}$ *and* $s \in \mathbb{C} \setminus \mathbb{R}$. *Then*
$$\widetilde{\tau}_{\min}(A - sI_2) = \sqrt{\left(\Re s - \tfrac{a+d}{2}\right)^2 + \left(\tfrac{|b-c|}{2} - \sqrt{\left(\tfrac{b+c}{2}\right)^2 + \left(\tfrac{a-d}{2}\right)^2 + (\Im s)^2}\,\right)^2}\,.$$
A real matrix Δ_0 *with* $\|\Delta\| = \widetilde{\tau}_{\min}(s\,I_2 - A)$ *and* $s \in \sigma(A + \Delta_0)$ *is given by*
$$\Delta_0 = \begin{bmatrix} \Re s - \frac{a+d}{2} & -\left(\frac{c-b}{2} + \kappa\sqrt{\left(\frac{b+c}{2}\right)^2 + \left(\frac{a-d}{2}\right)^2 + (\Im s)^2}\right) \\ \frac{c-b}{2} + \kappa\sqrt{\left(\frac{b+c}{2}\right)^2 + \left(\frac{a-d}{2}\right)^2 + (\Im s)^2} & \Re s - \frac{a+d}{2} \end{bmatrix}$$
where $\quad \kappa := \begin{cases} 1 & \text{if } c - b \le 0, \\ -1 & \text{if } c - b > 0. \end{cases}$

Observe that the matrix Δ_0 in Theorem 7.5.4 is normal. It has been used in [51] to construct lower bounds for real pseudospectra. However, the optimality of Δ_0 has not been shown in that paper.

For the proof of Theorem 7.5.3 and Theorem 7.5.4 we need the following three lemmas.

Lemma 7.5.5 *Let* $y \in \mathbb{R}\setminus\{0\}$. *Then*
$$\{\; X \in \mathbb{R}^{2\times 2} \mid iy \in \sigma(X) \;\} = \left\{ \begin{bmatrix} \xi & \zeta - \eta \\ \zeta + \eta & -\xi \end{bmatrix} \;\middle|\; \xi, \eta, \zeta \in \mathbb{R}, \quad \eta^2 - \xi^2 - \zeta^2 = y^2 \right\}.$$

Proof: This follows from the equivalence
$$iy \in \sigma(X) \quad \Leftrightarrow \quad (\; \mathrm{tr}(X) = 0, \text{ and } \det(X) = y^2 \;),$$
which holds for all $X \in \mathbb{R}^{2\times 2}$ □

For the next lemma we need some notation.

Notation 7.5.6 *In the sequel,*

$$\mathcal{R} := \left\{ \begin{bmatrix} x & -y \\ y & x \end{bmatrix} \;\middle|\; x, y \in \mathbb{R} \right\}, \qquad \mathcal{S} := \left\{ \begin{bmatrix} x & y \\ y & -x \end{bmatrix} \;\middle|\; x, y \in \mathbb{R} \right\}.$$

For $A = \begin{bmatrix} a & c \\ b & d \end{bmatrix} \in \mathbb{R}^{2\times 2}$ *we set* $A_{\mathcal{R}} := \begin{bmatrix} \frac{a+d}{2} & -\frac{b-c}{2} \\ \frac{b-c}{2} & \frac{a+d}{2} \end{bmatrix}$, $A_{\mathcal{S}} := \begin{bmatrix} \frac{a-d}{2} & \frac{b+c}{2} \\ \frac{b+c}{2} & -\frac{a-d}{2} \end{bmatrix}$.

Lemma 7.5.7 *The following statements hold.*

(*i*) *We have the direct decomposition* $\mathbb{R}^{2\times 2} = \mathcal{R} \oplus \mathcal{S}$ *with accompanying projections* $A \mapsto A_{\mathcal{R}} \in \mathcal{R}$, $A \mapsto A_{\mathcal{S}} \in \mathcal{S}$.

(*ii*) *For all* $x, y \in \mathbb{R}$ *and every* $v \in \mathbb{R}^2$,

$$\left\| \begin{bmatrix} x & -y \\ y & x \end{bmatrix} v \right\|_2 = \left\| \begin{bmatrix} x & y \\ y & -x \end{bmatrix} v \right\|_2 = \sqrt{x^2 + y^2}\, \|v\|_2.$$

Thus,

$$\left\| \begin{bmatrix} x & -y \\ y & x \end{bmatrix} \right\|_2 = \left\| \begin{bmatrix} x & y \\ y & -x \end{bmatrix} \right\|_2 = \sqrt{x^2 + y^2}.$$

(*iii*) *The singular values of* $A \in \mathbb{R}^{2\times 2}$ *are*

$$\sigma_{\max}(A) = \|A\|_2 = \|A_{\mathcal{R}}\|_2 + \|A_{\mathcal{S}}\|_2, \quad \sigma_{\min}(A) = |\, \|A_{\mathcal{R}}\|_2 - \|A_{\mathcal{S}}\|_2 \,|.$$

Proof: Claims (*i*) and (*ii*) are easily checked. Claim (*iii*) could be verified by a straightforward computation. However, we prefer the following more geometric proof: (*iii*) follows from (*ii*) if $A_{\mathcal{R}} = 0$ or $A_{\mathcal{S}} = 0$. So let $A_{\mathcal{R}} \neq 0$, $A_{\mathcal{S}} \neq 0$. Then $\det(A_{\mathcal{R}}) > 0, \det(A_{\mathcal{S}}) < 0$ and hence $\det(A_{\mathcal{S}} A_{\mathcal{R}}^{-1}) < 0$, so the pencil $A_{\mathcal{S}} - \lambda A_{\mathcal{R}}$ has a positive and a negative eigenvalue, i.e. there are $\lambda_+ > 0, \lambda_- < 0, v_+, v_- \in \mathbb{R}^2 \setminus \{0\}$ such that

$$A_{\mathcal{S}} v_+ = \lambda_+ A_{\mathcal{R}} v_+, \qquad A_{\mathcal{S}} v_- = \lambda_- A_{\mathcal{R}} v_-.$$

We have

$$\begin{aligned}
|\, \|A_{\mathcal{R}}\|_2 - \|A_{\mathcal{S}}\|_2 \,|\; \|v_-\|_2 &= |\, \|A_{\mathcal{R}}\|_2\, \|v_-\|_2 - \|A_{\mathcal{S}}\|_2\, \|v_-\|_2 \,| \\
&= |\, \|A_{\mathcal{R}}\, v_-\|_2 - \|A_{\mathcal{S}}\, v_-\|_2 \,| \\
&= |\, \|A_{\mathcal{R}}\, v_-\|_2 - |\lambda_-|\, \|A_{\mathcal{R}}\, v_-\|_2 \,| \\
&= |1 - |\lambda_-||\, \|A_{\mathcal{R}}\, v_-\|_2 \\
&= |1 + \lambda_-\,|\, \|A_{\mathcal{R}}\, v_-\|_2 \\
&= \|\,(1 + \lambda_-) A_{\mathcal{R}}\, v_-\|_2 \\
&= \|\, A_{\mathcal{R}}\, v_- + A_{\mathcal{S}}\, v_-\|_2 \\
&= \|A v_-\|_2,
\end{aligned}$$

and for all $v \in \mathbb{R}^2$,

$$\begin{aligned}
|\, \|A_{\mathcal{R}}\|_2 - \|A_{\mathcal{S}}\|_2 \,|\; \|v\|_2 &= |\, \|A_{\mathcal{R}}\|_2\, \|v\|_2 - \|A_{\mathcal{S}}\|_2\, \|v\|_2 \,| \\
&= |\, \|A_{\mathcal{R}}\, v\|_2 - \|A_{\mathcal{S}}\, v\|_2 \,| \\
&\leq \|A_{\mathcal{R}}\, v + A_{\mathcal{S}}\, v\|_2 \\
&= \|Av\|_2.
\end{aligned}$$

Thus $|\,\|A_{\mathcal{R}}\|_2 - \|A_{\mathcal{S}}\|_2\,| = \min_{v\neq 0}\frac{\|Av\|_2}{\|v\|_2} = \sigma_{\min}(A)$. Analogously one obtains the relations $(\|A_{\mathcal{R}}\| + \|A_{\mathcal{S}}\|_2)\,\|v_+\|_2 = \|Av_+\|_2$ and $\|A_{\mathcal{R}}\|_2 + \|A_{\mathcal{S}}\|_2 = \max_{v\neq 0}\frac{\|Av\|_2}{\|v\|_2} = \sigma_{\max}(A)$. □

Lemma 7.5.8 *Let $\|\!|\cdot|\!\|$ be an absolute and symmetric norm on $\mathbb{R}^2$ and let $\phi(x,t) = \left\|\!\left|\begin{bmatrix} x+t \\ x-t \end{bmatrix}\right|\!\right\|$, where $x, t\in\mathbb{R}$. Then the functions $x\mapsto\phi(x,t)$, $t\mapsto\phi(x,t)$ are increasing on $[0,\infty)$. They are decreasing on $(-\infty,0]$.*

Proof: We have for every $x,t\in\mathbb{R}$,

$$\phi(-x,t) = \left\|\!\left|\begin{bmatrix} -x+t \\ -x-t \end{bmatrix}\right|\!\right\| = \left\|\!\left|\begin{bmatrix} -x-t \\ -x+t \end{bmatrix}\right|\!\right\| = \left\|\!\left|\begin{bmatrix} |-x-t| \\ |-x+t| \end{bmatrix}\right|\!\right\| = \left\|\!\left|\begin{bmatrix} x+t \\ -x+t \end{bmatrix}\right|\!\right\| = \phi(x,t),$$

and, analogously, $\phi(x,-t) = \phi(x,t)$. Thus, the claim follows from convexity of the functions $x\mapsto\phi(x,t)$, $t\mapsto\phi(x,t)$. □

Proof of Theorem 7.5.3 and Theorem 7.5.4: Let $A = \begin{bmatrix} a & c \\ b & d \end{bmatrix}\in\mathbb{R}^{2\times 2}$ and $s\in\mathbb{R}$. Then we have according to Corollary 2.5.5 and Lemma 7.5.7,

$$\begin{aligned}
\widetilde{\mu}_{\mathbb{R}}^{\|\cdot\|}(A - sI_2) &= \sigma_{\min}(A - sI_2) = |\,\|(A-sI_2)_{\mathcal{R}}\| - \|(A-sI_2)_{\mathcal{S}}\|\,| \\
&= \left|\sqrt{\left(s - \frac{a+d}{2}\right)^2 + \left(\frac{b-c}{2}\right)^2} - \sqrt{\left(\frac{a-d}{2}\right)^2 + \left(\frac{b+c}{2}\right)^2}\right|.
\end{aligned}$$

Next, we are going to determine $\widetilde{\mu}_{\mathbb{R}}^{\|\cdot\|}(A - sI_2)$ for $s\notin\mathbb{R}$. Let $y := \Im s \neq 0$ and $D_y := \{\ \eta\in\mathbb{R} \mid |\eta| \geq |y|\ \}$. For $\eta\in D_y$ set

$$P_\eta := \{\ [\xi\ \eta]^T \mid \xi^2 + \zeta^2 = \eta^2 - y^2\ \}, \qquad X_{\eta,\xi,\zeta} := \begin{bmatrix} \xi & \zeta-\eta \\ \zeta+\eta & -\xi \end{bmatrix}$$

Then by Lemma 7.5.5, $\left\{\ X\in\mathbb{R}^{2\times 2}\ \middle|\ iy\in\sigma(X)\ \right\} = \left\{\ X_{\eta,\xi,\zeta} \mid \eta\in D_y,\ [\xi\ \eta]^T\in P_\eta\ \right\}$. Thus

$$\begin{aligned}
\widetilde{\mu}_{\mathbb{R}}^{\|\cdot\|}(A - s\,I_2) &= \min\left\{\ \|\Delta\| \mid \Delta\in\mathbb{R}^{2\times 2},\ s\in\sigma(A+\Delta)\ \right\} \\
&= \min\left\{\ \|\Delta\| \mid \Delta\in\mathbb{R}^{2\times 2},\ iy\in\sigma(A - (\Re s)I_2 + \Delta)\ \right\} \\
&= \min\left\{\ \|X - A + (\Re s)I_2\| \mid X\in\mathbb{R}^{2\times 2},\ iy\in\sigma(X)\ \right\} \\
&= \min\left\{\ \|X_{\eta,\xi,\zeta} - A + (\Re s)I_2\| \mid \eta\in D_y,\ [\xi\ \eta]^T\in P_\eta\ \right\} \\
&= \min\left\{\ \psi(\eta) \mid \eta\in D_\eta\ \right\},
\end{aligned}$$

where

$$\psi(\eta) = \min\left\{\ \|X_{\eta,\xi,\zeta} - A + (\Re s)I_2\| \mid [\xi\ \eta]^T\in P_\eta\ \right\}.$$

We determine $\psi(\eta)$. Set $\alpha := (\Re s)I_2 - \frac{a+d}{2}$, $\beta := \frac{b-c}{2}$, $\gamma := \frac{a-d}{2}$, $\delta := \frac{b+c}{2}$. Then

$$\begin{aligned}
(X_{\eta,\xi,\zeta} - A + (\Re s)I_2)_{\mathcal{R}} &= \begin{bmatrix} \alpha & -(\eta-\beta) \\ \eta-\beta & \alpha \end{bmatrix}, \\
(X_{\eta,\xi,\zeta} - A + (\Re s)I_2)_{\mathcal{S}} &= \begin{bmatrix} \xi-\gamma & \zeta-\delta \\ \zeta-\delta & -(\xi-\gamma) \end{bmatrix},
\end{aligned}$$

and, by Lemma 7.5.7,

$$\begin{aligned}
\|(X_{\eta,\xi,\zeta} - A + (\Re s)I_2)_{\mathcal{R}}\|_2 &= \sqrt{\alpha^2 + (\eta-\beta)^2} =: f(\eta) \\
\|(X_{\eta,\xi,\zeta} - A + (\Re s)I_2)_{\mathcal{S}}\|_2 &= \sqrt{(\xi-\gamma)^2 + (\zeta-\delta)^2} =: h(\xi,\zeta), \\
\sigma_{\max}(X_{\eta,\xi,\zeta} - A + (\Re s)I_2) &= f(\eta) + h(\xi,\zeta), \\
\sigma_{\min}(X_{\eta,\xi,\zeta} - A + (\Re s)I_2) &= |f(\eta) - h(\xi,\zeta)|.
\end{aligned}$$

Thus

$$\begin{aligned}
\|X_{\eta,\xi,\zeta} - A + (\Re s)I_2\| &= \left|\!\left|\!\left| \begin{bmatrix} \sigma_{\max}(X_{\eta,\xi,\zeta} - A + (\Re s)I_2) \\ \sigma_{\min}(X_{\eta,\xi,\zeta} - A + (\Re s)I_2) \end{bmatrix} \right|\!\right|\!\right| \\
&= \left|\!\left|\!\left| \begin{bmatrix} f(\eta) + h(\xi,\zeta) \\ |f(\eta) - h(\xi,\zeta)| \end{bmatrix} \right|\!\right|\!\right| \\
&= \left|\!\left|\!\left| \begin{bmatrix} f(\eta) + h(\xi,\zeta) \\ f(\eta) - h(\xi,\zeta) \end{bmatrix} \right|\!\right|\!\right| \qquad \text{(since the norm } |\!|\!|\cdot|\!|\!| \text{ is absolute)} \\
&= \phi(f(\eta), h(\xi,\zeta)),
\end{aligned}$$

where for all $x, t \in \mathbb{R}$,

$$\phi(x,t) = \left|\!\left|\!\left| \begin{bmatrix} x+t \\ x-t \end{bmatrix} \right|\!\right|\!\right|.$$

By Lemma 7.5.8 the function $0 \le t \mapsto \phi(f(\eta), t)$ is increasing. Therefore,

$$\begin{aligned}
\psi(\eta) &= \min\left\{\ \|X_{\eta,\xi,\zeta} - A + (\Re s)I_2\| \ \mid\ [\xi\ \eta]^T \in P_\eta \right\} \\
&= \min\left\{\ \phi(f(\eta), h(\xi,\zeta)) \mid\ [\xi\ \eta]^T \in P_\eta \right\} \\
&= \phi(f(\eta), g(\eta)),
\end{aligned}$$

where

$$\begin{aligned}
g(\eta) &:= \min\left\{ h(\xi,\zeta) \mid\ [\xi\ \ \zeta]^T \in P_\eta \right\} \\
&= \min\left\{ \sqrt{(\xi-\gamma)^2 + (\zeta-\delta)^2} \mid\ [\xi\ \ \zeta]^T \in P_\eta \right\} \\
&= \min\left\{ \|\,[\xi\ \ \zeta]^T - [\gamma\ \delta]^T\,\|_2 \ \mid\ \xi,\zeta \in \mathbb{R},\ \|\,[\xi\ \ \zeta]^T\|_2 = \sqrt{\eta^2 - y^2} \right\}.
\end{aligned}$$

The minimum is attained for the vector

$$\begin{bmatrix} \xi(\eta) \\ \zeta(\eta) \end{bmatrix} := q(\eta) \begin{bmatrix} \gamma \\ \delta \end{bmatrix}, \qquad q(\eta) := \frac{\sqrt{\eta^2 - y^2}}{\sqrt{\gamma^2 + \delta^2}},$$

and we have

$$g(\eta) = h(\xi(\eta), \zeta(\eta)) = \left| \sqrt{\eta^2 - y^2} - \sqrt{\gamma^2 + \delta^2} \right|.$$

A matrix $\Delta_\eta \in \mathbb{R}^{2\times 2}$ satisfying $s \in \sigma(A + \Delta_\eta)$ and $\|\Delta_\eta\| = \psi(\eta)$ is given by

$$\begin{aligned}
\Delta_\eta &:= X_{\eta,\xi(\eta),\zeta(\eta)} - A + (\Re s)I_2 \\
&= \begin{bmatrix} q(\eta)\gamma & q(\eta)\delta - \eta \\ q(\eta)\delta + \eta & -q(\eta)\gamma \end{bmatrix} - A + (\Re s)I_2.
\end{aligned}$$

Next, we show that $\min_{\eta \in D_y} \psi(\eta)$ is attained for some η_0 in the interval $[-R,\, R]$, where $R = \max\left\{|\beta|\,,\ \sqrt{\gamma^2+\delta^2+y^2}\right\}$. The functions $f, g : D_y \to \mathbb{R}$ is differentiable on the set

$$D_y' := \left\{\ \eta \in \mathbb{R}\ \middle|\ |\eta| > |y|,\ \alpha^2 + (\eta-\beta)^2 \neq 0,\ \sqrt{\eta^2-y^2} \neq \sqrt{\gamma^2+\delta^2}\ \right\}$$

and we have

$$f'(\eta) \quad = \quad \frac{\eta-\beta}{\sqrt{\alpha^2+(\eta-\beta)^2}}, \qquad g'(\eta) = \operatorname{sign}\left(\sqrt{\eta^2-y^2} - \sqrt{\gamma^2+\delta^2}\right)\frac{\eta}{\sqrt{\eta^2-y^2}}.$$

The derivatives satisfy $f'(\eta), g'(\eta) \geq 0$ if $\eta \geq R$. and $f'(\eta), g'(\eta) \leq 0$ if $\eta \leq -R$. Hence,

$$\begin{aligned} f(R) \leq f(\eta),\ \ g(R) \leq g(\eta)\ \text{ if } \eta \geq R \\ f(-R) \leq f(\eta),\ \ g(-R) \leq g(\eta)\ \text{ if } \eta \leq -R. \end{aligned}$$

Thus, Lemma 7.5.8 yields

$$\begin{aligned} \psi(R) = \phi(f(R), g(R)) \leq \phi(f(\eta), g(\eta)) = \psi(\eta)\ \text{ if } \eta \geq R, \\ \psi(-R) = \phi(f(-R), g(-R)) \leq \phi(f(\eta), g(\eta)) = \psi(\eta)\ \text{ if } \eta \leq -R. \end{aligned}$$

Thus $\min_{\eta \in D_y} \psi(\eta)$ is attained for some η_0 satisfying $|\eta_0| \leq R$. Next, we determine $\widetilde{\tau}_{\min}(A - sI_2)$, i.e. we consider the special case in which $\|\cdot\|$ is the maximum norm on $\mathbb{R}^2$. In this case

$$\psi(\eta) = \phi(f(\eta), g(\eta)) = f(\eta) + g(\eta).$$

The derivatives of f and g satisfy

$$|f'(\eta)| \leq 1, \qquad |g'(\eta)| > 1, \qquad \operatorname{sign}(g'(\eta)) = \operatorname{sign}\left(\sqrt{\eta^2-y^2} - \sqrt{\gamma^2+\delta^2}\right)\operatorname{sign}(\eta).$$

Hence we have for $\eta_* := \sqrt{\gamma^2+\delta^2+y^2}$ and all $\eta \in D_y'$,

$$\psi'(\eta) = (f+g)'(\eta) \begin{cases} < 0 & \text{if } \eta < -\eta_*, \\ > 0 & \text{if } -\eta_* < \eta < -|y|, \\ < 0 & \text{if } |y| < \eta < \eta_*, \\ > 0 & \text{if } \eta_* < \eta. \end{cases}$$

These inequalities together with the continuity of ψ and the fact that the set $D_y \setminus D_y'$ is finite yield that the local minima of $\psi : D_y \to \mathbb{R}$ are at the points $\pm\eta_*$. Thus

$$\begin{aligned} \widetilde{\tau}_{\min}(A - sI_2) \quad &= \quad \inf\{\ \psi(\eta)\ |\ \eta \in D_y\ \} \\ &= \quad \min\{\psi(-\eta_*), \psi(\eta_*)\} \\ &= \quad \min\left\{\sqrt{\alpha^2 + \left(\sqrt{\gamma^2+\delta^2+y^2} - \beta\right)^2}, \sqrt{\alpha^2 + \left(-\sqrt{\gamma^2+\delta^2+y^2} - \beta\right)^2}\right\} \\ &= \quad \sqrt{\alpha^2 + \left(|\beta| - \sqrt{\gamma^2+\delta^2+y^2}\right)^2} \\ &= \quad \sqrt{\left(\frac{a+d}{2}\right)^2 + \left(\frac{|b-c|}{2} - \sqrt{\left(\frac{b+c}{2}\right)^2 + \left(\frac{a-d}{2}\right)^2 + y^2}\right)^2}. \end{aligned}$$

Finally, note that $q(\pm\eta_*) = 1$. Thus, a matrix $\Delta_{\pm\eta_*} \in \mathbb{R}^{2\times 2}$ satisfying $s \in \sigma(A + \Delta_{\pm\eta_*})$ and $\|\Delta_{\pm\eta_*}\|_2 = \psi(\pm\eta_*)$ is given by

$$\Delta_{\pm\eta_*} \;=\; \begin{bmatrix} \gamma & \delta - (\pm\eta_*) \\ \delta + (\pm\eta_*) & -\gamma \end{bmatrix} - A + (\Re s)I_2 \;=\; (\Re s)I_2 - A_{\mathcal{R}} + \begin{bmatrix} 0 & -(\pm\eta_*) \\ (\pm\eta_*) & 0 \end{bmatrix}.$$

□

We now give an application of the just proved result. Let $A_\delta := \begin{bmatrix} \lambda_1 & \delta \\ 0 & \lambda_2 \end{bmatrix} \in \mathbb{R}^{2\times 2}$. In Section 5.5 it has been shown that $0 \le \delta_1 < \delta_2$ implies $\sigma_{\mathbb{C}}^{(2)}(A_{\delta_1}, \rho) \subset \sigma_{\mathbb{C}}^{(2)}(A_{\delta_2}, \rho)$. Moreover, it has been proved that $\bigcup_{\delta\ge 0} \sigma_{\mathbb{C}}^{(2)}(A_\delta, \rho) = \mathbb{C}$ for any $\rho > 0$. The former statement also holds for real pseudospectra while the latter does not, as the next proposition shows. We continue with denoting by

$$\delta(A) = \sqrt{\operatorname{tr}(A^T A) - (1/2)(\operatorname{tr}(A)^2 + |\operatorname{tr}(A)^2 - 4\det(A)|\,)}$$

the departure from normality of the 2×2 matrix A.

Proposition 7.5.9 *For any $\rho > 0$ the following assertions hold.*

(i) Let $A, B \in \mathbb{R}^{2\times 2}$ with $\sigma(A) = \sigma(B)$ and $\delta(A) < \delta(B)$. Then $\sigma_{\mathbb{R}}^{(2)}(A, \rho) \subset \sigma_{\mathbb{R}}^{(2)}(B, \rho)$.

(ii) Let $(A_k)_{k\in\mathbb{N}_0}$ be a sequence in $\mathbb{R}^{2\times 2}$ satisfying $\sigma(A_k) = \sigma(A_0)$ for every $k \in \mathbb{N}$ and $\lim_{k\to\infty} \delta(A_k) = \infty$. Then

$$\bigcup_{k\in\mathbb{N}} \sigma_{\mathbb{R}}^{(2)}(A_k, \rho) = \mathbb{R} \cup \{\, s \in \mathbb{C} \mid \; |\Re s - (1/2)\operatorname{tr}(A_0)| \le \rho \,\}.$$

Proof: Let $\lambda_1, \lambda_2, a, b \in \mathbb{R}$. For $s \in \mathbb{C}$ and $\delta \ge 0$ we set

$$f_1(s,\delta) \;:=\; \left| \sqrt{\left(\Re s - \tfrac{\lambda_1+\lambda_2}{2}\right)^2 + \left(\tfrac{\delta}{2}\right)^2} - \sqrt{\left(\tfrac{\lambda_1-\lambda_2}{2}\right)^2 + \left(\tfrac{\delta}{2}\right)^2} \right|,$$

$$f_2(s,\delta) \;:=\; \sqrt{\left(\Re s - \tfrac{\lambda_1+\lambda_2}{2}\right)^2 + \left(\tfrac{\delta}{2} - \sqrt{\left(\tfrac{\lambda_1-\lambda_2}{2}\right)^2 + \left(\tfrac{\delta}{2}\right)^2 + (\Im s)^2}\right)^2},$$

$$f_3(s,\delta) \;:=\; \left| \sqrt{(\Re s - a)^2 + \left(\tfrac{\delta}{2}\right)^2 + b^2} \;-\; \frac{\delta}{2} \right|,$$

$$f_4(s,\delta) \;:=\; \sqrt{(\,x - a)^2 + \left(\sqrt{\left(\tfrac{\delta}{2}\right)^2 + b^2} - \sqrt{\left(\tfrac{\delta}{2}\right)^2 + (\Im s)^2}\right)^2}.$$

Then, by applying Theorem 7.5.4 to the real Schur form of $A \in \mathbb{R}^{2\times 2}$ we obtain after a direct computation:

- If $\sigma(A) = \{\lambda_1, \lambda_2\}$ then

$$\widetilde{\tau}_{\min}(A - sI_2) = \widetilde{\tau}_{\min}\left(\begin{bmatrix} \lambda_1 & \delta(A) \\ 0 & \lambda_2 \end{bmatrix} - sI_2\right) = \begin{cases} f_1(s, \delta(A)) & \text{if } s \in \mathbb{R}, \\ f_2(s, \delta(A)) & \text{otherwise.} \end{cases}$$

- If $\sigma(A)=\{a+ib, a-ib\}$ then (using relation (7.23)),

$$\widetilde{\tau}_{\min}(A-sI_2)=\widetilde{\tau}_{\min}\left(\begin{bmatrix} a & -\theta_A\, b \\ \theta_A^{-1} b & a \end{bmatrix} - sI_2\right) = \begin{cases} f_3(s,\delta(A)) & \text{if } s\in\mathbb{R}, \\ f_4(s,\delta(A)) & \text{otherwise.} \end{cases}$$

Thus, the claim is a consequence of the following easily verified facts.

- For each $s\in\mathbb{C}$ the functions $0\geq\delta\mapsto f_k(s,\delta)$, $k=1,\ldots,4$ are strictly decreasing and

$$\begin{aligned} \lim_{\delta\to\infty} f_1(s,\delta) &= \lim_{\delta\to\infty} f_3(s,\delta)=0, \\ \lim_{\delta\to\infty} f_2(s,\delta) &= \left|\Re s - \frac{\lambda_1+\lambda_2}{2}\right|, \\ \lim_{\delta\to\infty} f_4(s,\delta) &= \left|\Re s - a\right|. \end{aligned}$$

□

From claim (i) of Proposition 7.5.9 we obtain the following corollary.

Corollary 7.5.10 *Let $A\in\mathbb{R}^{2\times 2}$ be normal. Then for any nonsingular $S\in\mathbb{R}^{2\times 2}$ and any $\rho>0$,*

$$\sigma_{\mathbb{R}}^{(2)}(A,\rho)\subseteq\sigma_{\mathbb{R}}^{(2)}(SAS^{-1},\rho). \tag{7.24}$$

In the next section we discuss the question whether (7.24) holds for all normal $A\in\mathbb{R}^{n\times n}$, n arbitrary.

7.6 Two conjectures

In Section 5.3 we have shown, that if $A\in\mathbb{C}^{n\times n}$ is normal then for any nonsingular $S\in\mathbb{C}^{n\times n}$,

$$\sigma_{\mathbb{C}}^{(2)}(A,\rho)\subseteq\sigma_{\mathbb{C}}^{(2)}(SAS^{-1},\rho).$$

We conjecture that this statement holds with $\mathbb{C}$ replaced by $\mathbb{R}$:

Conjecture 7.6.1 *Let $A\in\mathbb{R}^{n\times n}$ be normal. Then for any nonsingular $S\in\mathbb{R}^{n\times n}$ and any $\rho>0$,*

$$\sigma_{\mathbb{R}}^{(2)}(A,\rho)\subseteq\sigma_{\mathbb{R}}^{(2)}(SAS^{-1},\rho).$$

In the last section we have seen that Conjecture 7.6.1 holds true for $n=2$. Below we will show that Conjecture 7.6.1 is closely related to the following.

Conjecture 7.6.2 *Let $M_k\in\mathbb{C}^{q_k\times l_k}$, $k=1,2$ and $M_3\in\mathbb{C}^{q_1\times l_2}$. Then the function*

$$[0,\infty)\ni r\longmapsto\tau_1\left(\begin{bmatrix} M_1 & rM_3 \\ 0 & M_2 \end{bmatrix}\right)$$

is increasing.

Conjecture 7.6.2 has been confirmed by numerous numerical experiments. However, all our attempts to find a proof failed. We now discuss the consequences of Conjecture 7.6.2. In particular

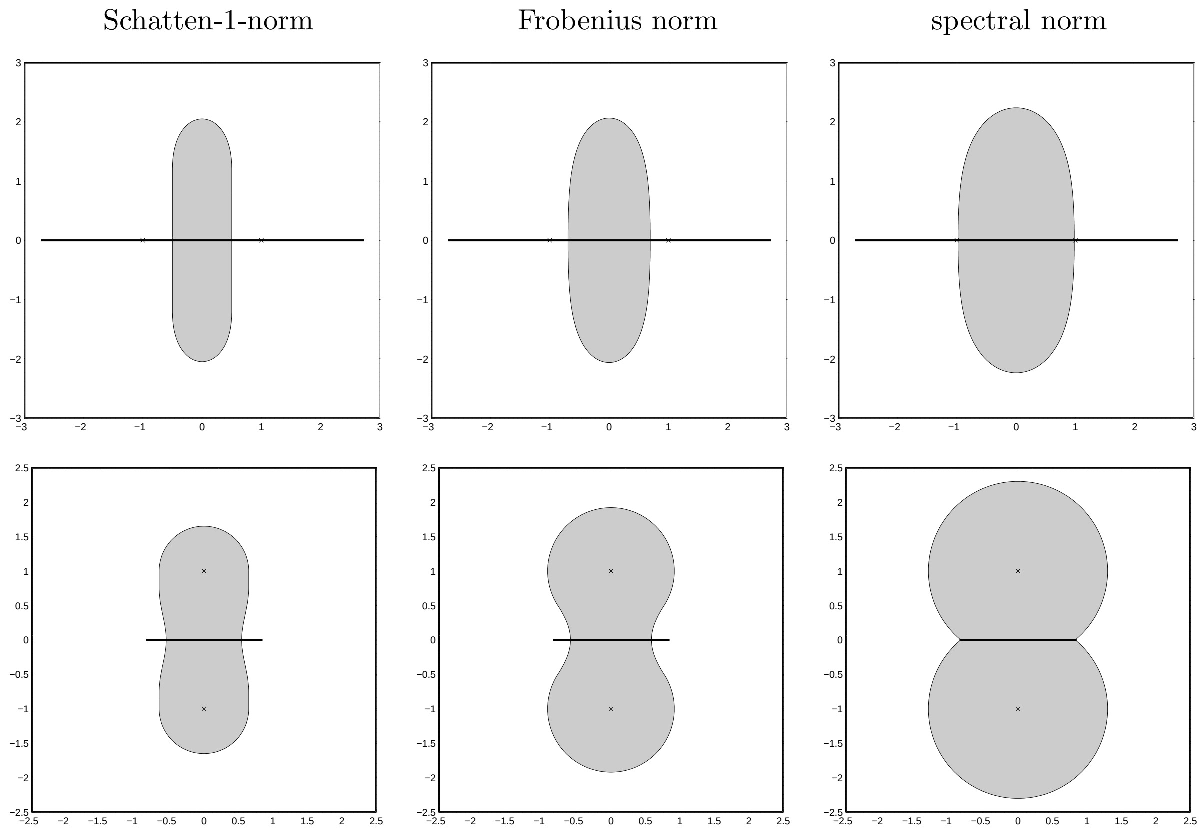


Figure 7.8: Pseudospectra $\sigma_{\mathbb{R}}^{\|\cdot\|}(A,\rho)$ of 2×2 matrices. Upper row: $\rho = 1$, $\sigma(A) = \{-1, 1\}$, $\delta(A) = 5$. Lower row: $\rho = 1.3$, $\sigma(A) = \{-i, i\}$, $\delta(A) = 0$.

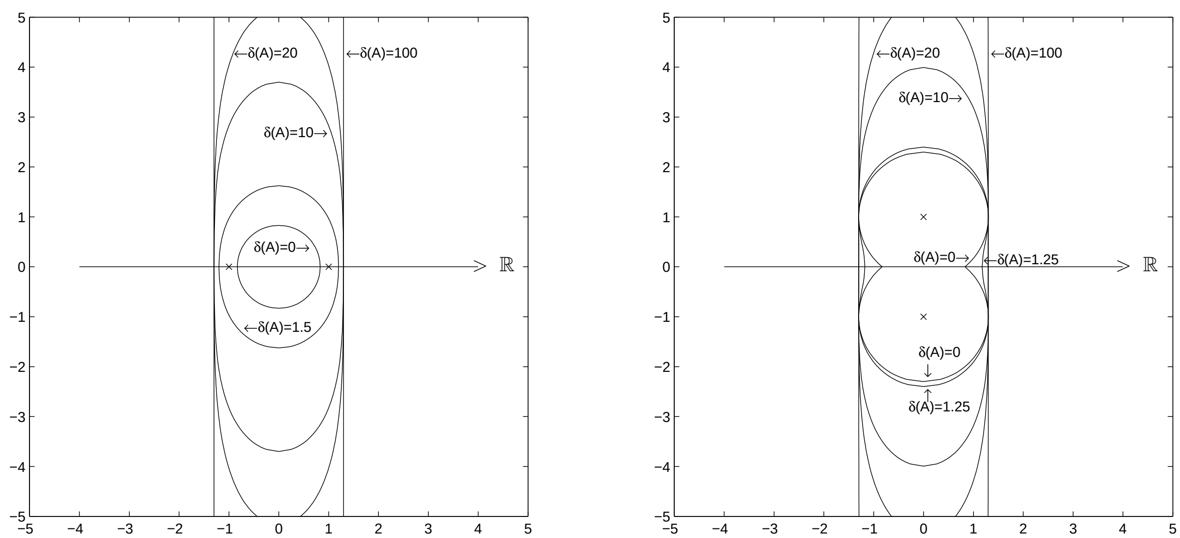


Figure 7.9: The sets $\{\ s\in\mathbb{C}\setminus\mathbb{R} \mid\ \widetilde{\tau}_{\min}(sI-A)=\rho\ \}$, $A\in\mathbb{R}^{2\times 2}$. Left: $\sigma(A) = \{-1, 1\}$, $\rho = 1.3$ and $\delta(A)\in\{0, 1.5, 10, 20, 100\}$. Right: $\sigma(A) = \{-i, i\}$, $\rho = 1.3$ and $\delta(A)\in\{0, 1.25, 10, 20, 100\}$

we show that Conjecture 7.6.2 together with an additional assumption implies Conjecture 7.6.1. Let $A_k \in \mathbb{C}^{n_k \times n_k}$, $k = 1, 2$, and $A_3 \in \mathbb{C}^{n_1 \times n_2}$. Set

$$A_r := \begin{bmatrix} A_1 & rA_3 \\ 0 & A_2 \end{bmatrix}, \qquad r \geq 0.$$

Then, for any $s \notin \sigma(A_r) = \sigma(A_1) \cup \sigma(A_2)$ we have

$$\begin{aligned} (sI_{n_1+n_2} - A_r)^{-1} &= \begin{bmatrix} sI_{n_1} - A_{n_1} & -A_2 \\ 0 & sI_{n_2} - A_{n_2} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} (sI_{n_1} - A_{n_1})^{-1} & r\,(sI_{n_1} - A_{n_1})^{-1} A_2 (sI_{n_2} - A_{n_2})^{-1} \\ 0 & (sI_{n_2} - A_{n_2})^{-1} \end{bmatrix}. \end{aligned}$$

Thus, the Conjecture 7.6.2 implies that the function

$$[0, \infty) \ni r \longmapsto \tau_1(\,(sI_{n_1+n_2} - A_r)^{-1}\,) = (\widetilde{\tau}_{\min}(sI_{n_1+n_2} - A_r))^{-1}$$

is increasing. Since the real pseudospectra of A_r satisfy

$$\sigma_{\mathbb{R}}^{(2)}(\,A_r\,,\,\rho\,) = \sigma(A_r) \cup \{\, s \in \mathbb{C} \backslash \sigma(A_r) \;\mid\; \widetilde{\tau}_{\min}(sI_{n_1+n_2} - A_r) \leq \rho\,\}.$$

we obtain the following result.

Corollary 7.6.3 *Suppose the Conjecture 7.6.2 holds true. Then $0 \leq r_1 \leq r_2$ implies that*

$$\sigma_{\mathbb{R}}^{(2)}(\,A_{r_1}\,,\,\rho\,) \subseteq \sigma_{\mathbb{R}}^{(2)}(\,A_{r_2}\,,\,\rho\,), \qquad \rho \geq 0. \tag{7.25}$$

Corollary 7.6.3 is the real counterpart to claim (a) of Proposition 5.6.1. The latter has been extended to the multiparamenter case in Theorem 5.6.6. Let us state the real analog to that Theorem. For $r = [r_1, \ldots, r_{m-1}]^T \in [0, \infty)^{m-1}$ consider the block upper triangular matrices

$$A_r := \begin{bmatrix} A_{11} & r_1 A_{12} & r_1 r_2 A_{13} & r_1 r_2 r_3 A_{13} & \cdots & r_1 r_2 r_3 \ldots r_{m-1} A_{1m} \\ & A_{22} & r_2 A_{23} & r_2 r_3 A_{24} & & r_2 r_3 \ldots r_{m-1} A_{2m} \\ & & A_{33} & r_3 A_{34} & & r_3 \ldots r_{m-1} A_{3m} \\ & & & \ddots & & \vdots \\ & 0 & & & & \\ & & & & A_{m-1,m-1} & r_{m-1} A_{m-1\,m} \\ & & & & & A_{mm} \end{bmatrix},$$

where $A_{jk} \in \mathbb{C}^{n_j \times n_k}$.

Proposition 7.6.4 *Let $r, r' \in [0, \infty)^{m-1}$. Suppose Conjecture 7.6.2 holds true. Then $r' \leq r$ (componentwise) implies that $\sigma_{\mathbb{R}}^{(2)}(\,A_{r'}\,,\,\rho\,) \subseteq \sigma_{\mathbb{R}}^{(2)}(\,A_r\,,\,\rho\,)$ for all $\rho \geq 0$. In particular we have*

$$\sigma_{\mathbb{R}}^{(2)}(\,A_r\,,\,\rho\,) \supseteq \sigma_{\mathbb{R}}^{(2)}(\,A_0\,,\,\rho\,) = \sigma_{\mathbb{R}}^{(2)}\left(\bigoplus\nolimits_{j \in \underline{m}} A_{jj}\,,\,\rho\right) = \bigcup_{1 \leq j < k \leq m} \sigma_{\mathbb{R}}^{(2)}(A_{jj} \oplus A_{kk}\,,\,\rho\,)\,.$$

Proof: This follows from Proposition 5.6.1 and the fact that for each index $k \leq m-1$ the matrix A_r can be written in the form $A_r = \begin{bmatrix} \hat{A}_{1k}(r) & r_k \hat{A}_{3k}(r) \\ 0 & \hat{A}_{2k}(r) \end{bmatrix}$, where the matrices $\hat{A}_{jk}(r)$ do not depend on the component r_k of r. (Observe that we used the same argument as in the proof of Theorem 5.6.6). $\square$

We remind the reader of the real Schur form of a real square matrix (see e.g. [61]). Let $A\in\mathbb{R}^{n\times n}$, $n\geq 2$, have real eigenvalues $\lambda_1,\dots,\lambda_p$ and nonreal eigenvalues $a_1\pm ib_1,\dots,a_q\pm ib_q$ $(a_j,b_j\in\mathbb{R}, b_j>0)$ counting multiplicities. Then there is an orthogonal matrix $Q\in\mathbb{R}^{n\times n}$ such that QAQ^{-1} is block upper triangular and the block diagonal D of QAQ^{-1} is of the form

$$D=\operatorname{diag}(\lambda_1,\dots,\lambda_p)\oplus\bigoplus_{j\in\underline{q}}\begin{bmatrix}a_j & -\theta_j b_j\\ \theta_j^{-1}b_j & a_j\end{bmatrix},\qquad \theta_j\geq 1.$$

If Conjecture 7.6.2 holds then Proposition 7.6.4 yields

$$\begin{aligned}\sigma_{\mathbb{R}}^{(2)}(A,\rho) &= \sigma_{\mathbb{R}}^{(2)}(QAQ^{-1},\rho)\\ &\supseteq \sigma_{\mathbb{R}}^{(2)}(D,\rho)\\ &= \bigcup_{1\leq j<k\leq p}\sigma_{\mathbb{R}}^{(2)}\left(\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix},\rho\right)\cup\bigcup_{1\leq j<k\leq q}\sigma_{\mathbb{R}}^{(2)}\left(\begin{bmatrix}a_j & -\theta_j\, b_j\\ \theta_j^{-1}b_j & a_j\end{bmatrix}\oplus\begin{bmatrix}a_k & -\theta_k\, b_k\\ \theta_k^{-1}b_k & a_k\end{bmatrix},\rho\right)\\ &\quad\cup\bigcup_{j\in\underline{q},k\in\underline{p}}\sigma_{\mathbb{R}}^{(2)}(N(a_j,b_j,\theta_j,\lambda_k),\rho),\end{aligned}\tag{7.26}$$

where

$$N(a,b,\theta,\lambda):=\begin{bmatrix}a & -\theta\, b & \\ \theta^{-1}b & a & 0\\ 0 & 0 & \lambda\end{bmatrix},\qquad\text{for } a,b,\lambda\in\mathbb{R},\ \theta\geq 1.$$

The inclusion 7.26 can be simplified using the fact that

$$\begin{aligned}\sigma_{\mathbb{R}}^{(2)}\left(\begin{bmatrix}a_j & -\theta_j\, b_j\\ \theta_j^{-1}b_j & a_j\end{bmatrix}\oplus\begin{bmatrix}a_k & -\theta_k\, b_k\\ \theta_k^{-1}b_k & a_k\end{bmatrix},\rho\right) &\supseteq \sigma_{\mathbb{R}}^{(2)}\left(\begin{bmatrix}a_j & -\theta_j\, b_j\\ \theta_j^{-1}b_j & a_j\end{bmatrix},\rho\right)\cup\sigma_{\mathbb{R}}^{(2)}\left(\begin{bmatrix}a_k & -\theta_k\, b_k\\ \theta_k^{-1}b_k & a_k\end{bmatrix},\rho\right)\\ &\quad\text{(by Corollary 7.1.6)}\\ &\supseteq \sigma_{\mathbb{R}}^{(2)}\left(\begin{bmatrix}a_j & -b_j\\ b_j & a_j\end{bmatrix},\rho\right)\cup\sigma_{\mathbb{R}}^{(2)}\left(\begin{bmatrix}a_k & -b_k\\ b_k & a_k\end{bmatrix},\rho\right)\\ &\quad\text{(by Proposition 7.5.9)}\\ &= \mathcal{D}(a_j+ib_j,\rho)\cup\mathcal{D}(a_k+ib_k,\rho).\end{aligned}$$

Hence,

$$\begin{aligned}\sigma_{\mathbb{R}}^{(2)}(A,\rho) \supseteq &\bigcup_{1\leq j<k\leq p}\sigma_{\mathbb{R}}^{(2)}\left(\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix},\rho\right)\cup\bigcup_{j\in\underline{q}}\mathcal{D}(a_j+ib_j,\rho)\\ &\cup\bigcup_{j\in\underline{q},k\in\underline{p}}\sigma_{\mathbb{R}}^{(2)}(N(a_j,b_j,\theta_j,\lambda_k),\rho).\end{aligned}\tag{7.27}$$

We now claim the following:

Claim 7.6.5 *For all $a, b, \lambda \in \mathbb{R}$, $s \in \mathbb{C}$ the function*

$$1 \leq \theta \longmapsto \widetilde{\tau}_{\min}(N(a,b,\theta,\lambda) - sI_3)$$

is decreasing.

The claim can be shown by means of the methods used in the proof of Proposition 7.4.4. However, the computations are tedious and will not be given here. The claim implies that for any $\rho \geq 0$,

$$\sigma_{\mathbb{R}}^{(2)}(\, N(a,b,\theta,\lambda),\, \rho\,) \supseteq \sigma_{\mathbb{R}}^{(2)}(\, N(a,b,1,\lambda),\, \rho\,)\,.$$

By combining this with (7.27) we obtain

Proposition 7.6.6 *Let $A \in \mathbb{R}^{n \times n}$, $n \geq 2$, have real eigenvalues $\lambda_1, \ldots, \lambda_p$ and nonreal eigenvalues $a_1 \pm ib_1, \ldots, a_q \pm ib_q$ $(a_j, b_j \in \mathbb{R}, b_j > 0)$ counting multiplicities. Suppose that Conjecture 7.6.2 and Claim 7.6.5 hold. Then for any $\rho \geq 0$,*

$$\begin{aligned} \sigma_{\mathbb{R}}^{(2)}(\, A,\, \rho\,) \quad &\supseteq \quad \bigcup_{1 \leq j < k \leq p} \sigma_{\mathbb{R}}^{(2)}\left(\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}, \rho \right) \cup \bigcup_{j \in \underline{q}} \mathcal{D}(a_j + ib_j, \rho) \cup \bigcup_{j \in \underline{q}, k \in \underline{p}} \sigma_{\mathbb{R}}^{(2)}(\, N(a_j, b_j, \theta_j, \lambda_k),\, \rho\,) \\ &= \quad \sigma_{\mathbb{R}}^{(2)}(\, A_0,\, \rho\,)\,, \end{aligned} \tag{7.28}$$

where

$$A_0 := \operatorname{diag}(\lambda_1, \ldots, \lambda_p) \oplus \bigoplus_{j \in \underline{q}} \begin{bmatrix} a_j & -b_j \\ b_j & a_j \end{bmatrix}.$$

Equation (7.28) holds by Theorem 7.4.1. Assume now that A is diagonalizable. Then there is a nonsingular matrix $S \in \mathbb{R}^{n \times n}$ such that $A = SA_0S^{-1}$ and the inclusion in Proposition 7.6.6 reads

$$\sigma_{\mathbb{R}}^{(2)}\left(\, SA_0S^{-1},\, \rho\,\right) \supseteq \sigma_{\mathbb{R}}^{(2)}(\, A_0,\, \rho\,)\,.$$

Hence, Conjecture 7.6.2 and Claim 7.6.5 together imply Conjecture 7.6.1.

7.7 Root sets of uncertain polynomials

We mentioned in Section 6.1 that a computable formula for $\mu_{\mathbb{R}}^{\|\cdot\|}(M)$ and arbitrary $M \in \mathbb{C}^{q\times l}$ is known only for the case that the underlying norm is the spectral norm. However, if M is a column or a row vector then one has the following result which holds for arbitrary norms. It reduces the computation of $\mu_{\mathbb{R}}^{\|\cdot\|}(M)$ to a 1-parameter optimization problem.

Theorem 7.7.1 *[42, 48, 53] Let $M \in \mathbb{C}^{m\times 1}$ $(M \in \mathbb{C}^{1\times m})$ and let $\|\cdot\|$ be a norm on $\mathbb{R}^{1\times m}$ $(\mathbb{R}^{m\times 1})$. Furthermore let $x = \Re M$, $y = \Im M$. Then*

$$\mu_{\mathbb{R}}^{\|\cdot\|}(M) \quad = \quad \min_{t\in\mathbb{R}} \|x - t\,y\|^D \tag{7.29}$$

where $\|\cdot\|^D$ denotes the dual to the norm $\|\cdot\|$.

In other words, Theorem 7.7.1 states that $\mu_{\mathbb{R}}^{\|\cdot\|}(x+iy)$ equals the distance of the vector $x\in\mathbb{R}^m$ to the one-dimensional vector space $\mathbb{R}y$ with respect to the norm $\|\cdot\|^D$. By means of this result one can compute root sets of interval polynomials or, more generally, root sets of polynomials whose coefficients are subject to linear real perturbations. We are now going to discuss this in detail. In doing so we also deal with complex perturbations.

Let p and $q_1, \dots q_m$ be polynomials of arbitrary degree with complex coefficients. We consider perturbations of the nominal polynomial p of the form

$$p(s) \rightsquigarrow p_\delta(s) := p(s) - \sum_{j=1}^{m} \delta_j q_j(s), \qquad \delta = [\delta_1, \dots, \delta_m] \in \mathbb{F}^{1\times m}. \tag{7.30}$$

The associated root sets with respect to the perturbation level $\rho \geq 0$ and the underlying norm $\|\cdot\|$ on $\mathbb{F}^{1\times m}$ are defined by

$$\mathcal{R}_{\mathbb{F}}^{\|\cdot\|}(p; q_1, \dots, q_m, \rho) := \{\ s\in\mathbb{C} \mid\ \exists\, \delta\in\mathbb{F}^{1\times m} : (\|\delta\| \leq \rho \text{ and } p_\delta(s) = 0\,)\ \}.$$

If the degree of the polynomials q_j is less than the degree of p then the root sets $\mathcal{R}_{\mathbb{F}}^{\|\cdot\|}(p; q_1, \dots, q_m, \rho)$ are spectral value sets. This can be seen as follows. Without loss of generality we may assume that p is monic. Then p and the polynomials q_j of degree $\deg(q_j) < \deg(p) =: n$ can be written in the form

$$p(s) = s^n + \sum_{k=1}^{n} a_k s^{k-1}, \qquad q_j(s) = \sum_{k=1}^{n} c_{jk} s^{k-1}$$

with coeffients $a_k, c_{jk} \in \mathbb{C}$. The polynomials p_δ then are

$$p_\delta(s) \quad = \quad p(s) - \sum_{j=1}^{m} \delta_j\, q_j(s) = s^n + \sum_{k=1}^{n} a_k(\delta)\, s^{k-1},$$

where

$$a_k(\delta) := a_k - \sum_{j=1}^{m} \delta_j\, c_{jk}.$$

The roots of p_δ coincide with the eigenvalues of its companion matrix $A_\delta \in \mathbb{C}^{n\times n}$. This matrix can be written in the form

$$\underbrace{\begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -a_1(\delta) & & -a_{n-1}(\delta) & -a_n(\delta) \end{bmatrix}}_{=A_\delta} = \underbrace{\begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -a_1 & & -a_{n-1} & -a_n \end{bmatrix}}_{=A_0} + \underbrace{\begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ 1 \end{bmatrix}}_{=:e_n} [\delta_1 \ \dots \ \delta_m] \underbrace{\begin{bmatrix} c_{11} & \dots & c_{1m} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ c_{m1} & \dots & c_{mm} \end{bmatrix}}_{=:C}$$

Thus, we have for all $\rho \geq 0$,

$$\begin{aligned} \mathcal{R}_{\mathbb{F}}^{\|\cdot\|}(p; q_1, \dots, q_m, \rho) &= \sigma_{\mathbb{F}}^{\|\cdot\|}(A_0, e_n, C, \rho) \\ &= \sigma(A_0) \cup \{\, s \in \mathbb{C} \backslash \sigma(A_0) \mid \mu_{\mathbb{F}}^{\|\cdot\|}(C(sI_n - A_0)^{-1} e_n) \geq \rho^{-1} \,\}. \end{aligned} \tag{7.31}$$

The vector $C(sI_n - A_0)^{-1}e_n$ can be written in a more explicite form: For $s \in \mathbb{C}$ let $\hat{s} = [\,1 \ s \ s^2 \ \dots \ s^{n-1}\,]^T$. Then $(sI_n - A_0)\hat{s} = p(s)\, e_n$. Thus,

$$\begin{aligned} C(sI_n - A_0)^{-1}e_n &= C(\, p(s)^{-1}\hat{s}\,) \\ &= p(s)^{-1}\, C\hat{s} \\ &= p(s)^{-1}[\, q_1(s) \ q_2(s) \ \dots q_m(s)\,]^T. \end{aligned}$$

By combining this with (7.31) we obtain

$$\mathcal{R}_{\mathbb{F}}^{\|\cdot\|}(p; q_1, \dots, q_m, \rho) = \mathcal{R}(p) \cup \{\, s \in \mathbb{C} \backslash \mathcal{R}(p) \mid \ f(s) \geq \rho^{-1} \,\}, \tag{7.32}$$

where $\mathcal{R}(p)$ denotes the set of roots of p and

$$f(s) := \mu_{\mathbb{F}}^{\|\cdot\|}(\, p(s)^{-1}[\, q_1(s) \ q_2(s) \ \dots \ q_m(s)\,]^T \,). \tag{7.33}$$

However, the derivation of (7.32) via spectral value sets is somewhat indirect. There is an alternative method which also works for the case that $\deg(q_j) \geq \deg(p)$ for some $j \in \underline{m}$:

Note first that the definition of μ for a column vector $M \in \mathbb{C}^{m\times 1} = \mathbb{C}^m$ reads

$$\begin{aligned} \mu_{\mathbb{F}}^{\|\cdot\|}(M) &= \big(\min\{\|\delta\| \mid \ \delta \in \mathbb{F}^{1\times m}, \ 1 \in \sigma(\delta M)\,\}\big)^{-1} && (7.34) \\ &= \big(\min\{\|\delta\| \mid \ \delta \in \mathbb{F}^{1\times m}, \ \delta M = 1\,\}\big)^{-1}. && (7.35) \end{aligned}$$

Observe further that if $s \notin \mathcal{R}(p)$ then $0 = p_\delta(s) = p(s) - \sum_{j=1}^m \delta_j \, q_j(s)$ if and only if

$$[\, \delta_1 \ \delta_2 \ \dots \ \delta_m\,]\, p(s)^{-1}[\, q_1(s) \ q_2(s) \ \dots q_m(s)\,]^T = 1.$$

Thus, the function f defined in (7.33) satisfies

$$\begin{aligned} f(s) &= \mu_{\mathbb{F}}^{\|\cdot\|}(\, p(s)^{-1}[\, q_1(s) \ q_2(s) \ \dots \ q_m(s)\,]^T \,) \\ &= \big(\min\{\|\delta\| \mid \ \delta \in \mathbb{F}^{1\times m}, \ \delta\, p(s)^{-1}\,[\, q_1(s) \ q_2(s) \ \dots q_m(s)\,]^T = 1\,\}\big)^{-1} \\ &= \big(\min\{\|\delta\| \mid \ \delta \in \mathbb{F}^{1\times m}, \ p_\delta(s) = 0\,\}\big)^{-1}. \end{aligned}$$

This yields (7.32). We remark that with an analogous reasoning one can treat spectral value sets of matrix polynomials. This has been done in [23, 40, 41, 44]. We note further, that if

$\deg(p_\delta)$ is not constant for all $\delta \in \mathbb{F}^m$ satisfying $\|\delta\| \le \rho$ then the root set $\mathcal{R}_{\mathbb{F}}^{\|\cdot\|}(p; q_1, \ldots, q_m, \rho)$ is unbounded. The proof is left to the reader.

We now discuss μ-functions of column vectors in detail. If $M \in \mathbb{F}^m = \mathbb{F}^{m\times 1}$ and $\|\cdot\|$ is a norm on $\mathbb{F}^{1\times m}$ then

$$\mu_{\mathbb{F}}^{\|\cdot\|}(M) = \big(\min\{\|\delta\| \mid \delta\in\mathbb{F}^{1\times m},\ \delta M = 1\,\}\big)^{-1} = \max_{\substack{\delta\in\mathbb{F}^{1\times m}\\ \|\delta'\|=1}} |\delta' M| = \|M\|^D. \tag{7.36}$$

For the case $M\in\mathbb{C}^m$, $\mathbb{F} = \mathbb{R}$, we have by Theorem 7.7.1,

$$\mu_{\mathbb{R}}^{\|\cdot\|}(M) = \min_{t\in\mathbb{R}} \|x - t\,y\|^D,$$

where $x = \Re M$, $y = \Im M$. Let us give the proof. The relation $\delta(x+iy) = 1$ with $\delta \in \mathbb{R}^{1\times m}$ implies $\delta x = 1$ and $\delta y = 0$. Thus $\delta \neq 0$ and $1 = \delta(x - ty)$ for all $t\in\mathbb{R}$. Hence,

$$1 = |\delta(x-ty)| \le \|\delta\|\,\|x-ty\|^D, \qquad \text{for all } t\in\mathbb{R}.$$

Consequently, $\|\delta\|^{-1} \le \min_{t\in\mathbb{R}} \|x - t\,y\|^D$. Since this inequality holds for all δ satisfying $\delta M = 1$ we conclude

$$\mu_{\mathbb{R}}^{\|\cdot\|}(M) \le \min_{t\in\mathbb{R}} \|x - t\,y\|^D. \tag{7.37}$$

We show that the latter inequality is actually an equality. Assume first that $x = t_0\,y$ for some $t_0\in\mathbb{R}$. Then there is no δ such that $\delta x = 1$ and $\delta y = 0$. Thus $\mu_{\mathbb{R}}^{\|\cdot\|}(M) = 0 = \min_{t\in\mathbb{R}} \|x - t\,y\|^D$. If $y = 0$ then $\min_{t\in\mathbb{R}} \|x - t\,y\|^D = \|x\|^D = \|M\|^D = \mu_{\mathbb{R}}^{\|\cdot\|}(M)$. The latter equation is (7.36). The remaining case is that x and y are linearly independent. In this case there is a linear functional $g : \mathrm{span}_{\mathbb{R}}\{x,y\} \to \mathbb{R}$ such that $g(x) = 1$, $g(y) = 0$. Its operator norm with respect to the norm $\|\cdot\|^D$ is

$$\begin{aligned}
|\!|\!|g|\!|\!| &= \max_{(\lambda_1,\lambda_2)\in\mathbb{R}^2\setminus\{0\}} \frac{|g(\lambda_1 x - \lambda_2 y)|}{\|\lambda_1 x - \lambda_2 y\|^D}\\
&= \max_{(\lambda_1,\lambda_2)\in\mathbb{R}^2\setminus\{0\}} \frac{|\lambda_1|}{\|\lambda_1 x - \lambda_2 y\|^D}\\
&= \max_{\substack{(\lambda_1,\lambda_2)\in\mathbb{R}^2\setminus\{0\}\\ \lambda_1\neq 0}} \frac{1}{\|x - (\lambda_2/\lambda_1) y\|^D}\\
&= \left(\min_{t\in\mathbb{R}} \|x - ty\|^D\right)^{-1}
\end{aligned}$$

By the Hahn-Banach theorem g can be extended to a linear functional $\widetilde{g} : \mathbb{R}^m \to \mathbb{R}$ with the same operator norm, $|\!|\!|\widetilde{g}|\!|\!| = |\!|\!|g|\!|\!|$. Let $\delta \in \mathbb{R}^{1\times m}$ be such that $\widetilde{g}(\xi) = \delta\xi$ for all $\xi \in \mathbb{R}^m$. Then $\|\delta\| = (\|\delta\|^D)^D = \max_{\xi\neq 0} \frac{|\delta\xi|}{\|\xi\|^D} = |\!|\!|\widetilde{g}|\!|\!| = |\!|\!|g|\!|\!|$. Thus, equality holds in (7.37). □

Since $\|\cdot\|_2^D = \|\cdot\|_2$ we have for $x, y\in\mathbb{R}^m$,

$$\mu_{\mathbb{R}}^{(2)}(x+iy) = \min_{t\in\mathbb{R}} \|x - ty\|_2 = \sqrt{\|x\|_2^2 - (x^T y)^2}. \tag{7.38}$$

The proposition below treats the μ-problem for the maximum norm.

Proposition 7.7.2 *Let* $x = [\,x_1 \dots x_m]^T \in \mathbb{R}^m$ *and* $y = [\,y_1 \dots y_m]^T \in \mathbb{R}^m \setminus \{0\}$. *Then*

$$\mu_{\mathbb{R}}^{(\infty)}(x+iy) = \min_{\substack{k\in \underline{m} \\ y_k \neq 0}} \|x - (x_k/y_k)y\|_1. \tag{7.39}$$

Proof: Since $\|\cdot\|_\infty^D = \|\cdot\|_1$ we have by Theorem 7.7.1 that $\mu_{\mathbb{R}}^{(\infty)}(x+iy) = \min_{t\in\mathbb{R}} h(t)$, where

$$h(t) := \|x - ty\|_1 = \sum_{k=1}^{m} |x_k - ty_k|.$$

Let $P = \{\, x_k/y_k \mid k \in \underline{m},\ y_k \neq 0 \,\}$. Then $\mathbb{R} \setminus P$ is a disjoint union of intervals. The closures of these intervals cover $\mathbb{R}$. Let J be such an interval with the property that $\min_{t\in\mathbb{R}} h(t)$ is attained at $\mathrm{cl}(J)$. Since $J \cap P = \emptyset$ the functions $t \mapsto x_k - ty_k$, $k \in \underline{m}$, do not change sign on J. Hence the restriction $h|_{\mathrm{cl}(J)}$ is an affine function and attains its minimum at $\partial J \subseteq P$. $\square$

By using (7.38) and (7.39) the sets root sets $\mathcal{R}_{\mathbb{R}}^{\|\cdot\|_\alpha}(p; q_1, \dots, q_m, \rho)$, $\alpha \in \{0, \infty\}$, can be determined numerically. An example is given in Figure 7.10. It extends an example from [42]. The figure shows the root sets $\mathcal{R}_{\mathbb{R}}^{\|\cdot\|_\alpha}(p; q_1, q_2, \rho)$ and the boundaries of $\mathcal{R}_{\mathbb{C}}^{\|\cdot\|_\alpha}(p; q_1, q_2, \rho)$ (dashed curves) for $\alpha = 2, \infty$, and the polynomials

$$\begin{aligned} p(s) &= (s+2)^2(s+0.5+0.5i)(s+0.5-0.5i) \\ &= s^4 + 5\,s^3 + 8.5\,s^2 + 6\,s + 2, \qquad (7.40) \\ q_1(s) &= s^2 + 4\,s + 3.25, \\ q_2(s) &= 1.5\,s + 1. \end{aligned}$$

We now consider root sets of interval polynomials. The latter are defined as follows. Let $\alpha = [\,\alpha_1 \dots \alpha_{n+1}]^T \in \mathbb{R}^{n+1}$, $\beta = [\,\beta_1 \dots \beta_{n+1}]^T \in \mathbb{R}^{n+1}$ be real vectors such that $\alpha \leq \beta$ (conponentwise). Then the interval polynomial $\mathcal{P}[\alpha, \beta]$ is the set of polynomials p of the form

$$p(s) = \sum_{k=1}^{n+1} a_k s^{k-1}, \qquad \alpha_k \leq a_k \leq \beta_k.$$

The corresponding root sets are

$$\mathcal{R}[\alpha, \beta] := \{\, s \in \mathbb{C} \mid p(s) = 0 \text{ for some } p \in \mathcal{P}[\alpha, \beta] \,\}.$$

Let

$$p^c(s) := \sum_{k=1}^{n+1} \frac{\alpha_k + \beta_k}{2}\, s^{k-1}, \qquad q_j(s) := \frac{\beta_j - \alpha_j}{2}\, s^{j-1}, \;\; j \in \underline{n+1}.$$

The polynomial p^c is called the center polynomial of $\mathcal{P}[\alpha, \beta]$. We have

$$\mathcal{P}[\alpha, \beta] = \{\, p^c + \textstyle\sum_{j=1}^{n+1} \delta_j\, q_j \;\mid\; \forall j \in \underline{n+1} : \delta_j \in \mathbb{R},\ |\delta_j| \leq 1 \,\},$$

and therefore

$$\mathcal{R}[\alpha, \beta] \;=\; \mathcal{R}_{\mathbb{R}}^{\|\cdot\|_\infty}(p^c; q_1, \dots, q_{n+1}, 1) = \mathcal{R}(p^c) \cup \{\, s \in \mathbb{C} \setminus \mathcal{R}(p^c) \;\mid\; f_{\alpha,\beta}(s) \geq 1 \,\},$$

where $\mathcal{R}(p^c)$ is the set of roots of p^c and

$$f_{\alpha,\beta}(s) := \mu_{\mathbb{R}}^{(\infty)}(\, p^c(s)^{-1}[\, q_1(s)\; q_2(s)\; \ldots q_{n+1}(s)\,]^T\,). \tag{7.41}$$

The function $f_{\alpha,\beta}(s)$ can be evaluated via Formula (7.39). The complex counterpart to $\mathcal{R}[\alpha,\beta]$ is the set

$$\mathcal{R}^{\mathbb{C}}[\alpha,\beta] := \{\, s \in \mathbb{C} \mid p(s) = 0 \text{ for some } p \in \mathcal{P}^{\mathbb{C}}[\alpha,\beta]\,\},$$

where

$$\mathcal{P}^{\mathbb{C}}[\alpha,\beta] := \{\, p^c + \textstyle\sum_{j=1}^{n+1} \delta_j\, q_j \;\mid\; \forall j \in \underline{n+1} : \delta_j \in \mathbb{C},\ |\delta_j| \le 1\,\}.$$

Figure 7.11 shows the sets $\mathcal{R}[\alpha^{(\kappa)},\beta^{(\kappa)}]$ and the boundaries of the sets $\mathcal{R}^{\mathbb{C}}[\alpha^{(\kappa)},\beta^{(\kappa)}]$, $\kappa = 1,2$, where

$$\begin{aligned} \alpha^{(1)} &= [\,1.5\;\; -9\;\; 7.5\;\; 4.5\;\; 1\,]^T, \\ \beta^{(1)} &= [\,2.5\;\; 21\;\; 9.5\;\; 5.5\;\; 1\,]^T, \\ \alpha^{(2)} &= [\,1.5\;\; -6\;\; 8.5\;\; 4.5\;\; 0.5\,]^T, \\ \beta^{(2)} &= [\,2.5\;\; 18\;\; 8.5\;\; 5.5\;\; 1.5\,]^T. \end{aligned} \tag{7.42}$$

In both cases the center polynomial is the polynomial p defined in (7.40).

Figure 7.12 shows the root set of the interval polynomial

$$p_\delta(s) = \prod_{k=1}^{12}(s+k) + \sum_{j=1}^{12} \delta_j\, s^{j-1}, \qquad \delta_j \in \mathbb{R},\ |\delta_j| \le 7 \text{ for all } j \in \underline{12}. \tag{7.43}$$

The figure indicates that the roots -2 and -1 of the center polynomial are insensitive to perturbations.

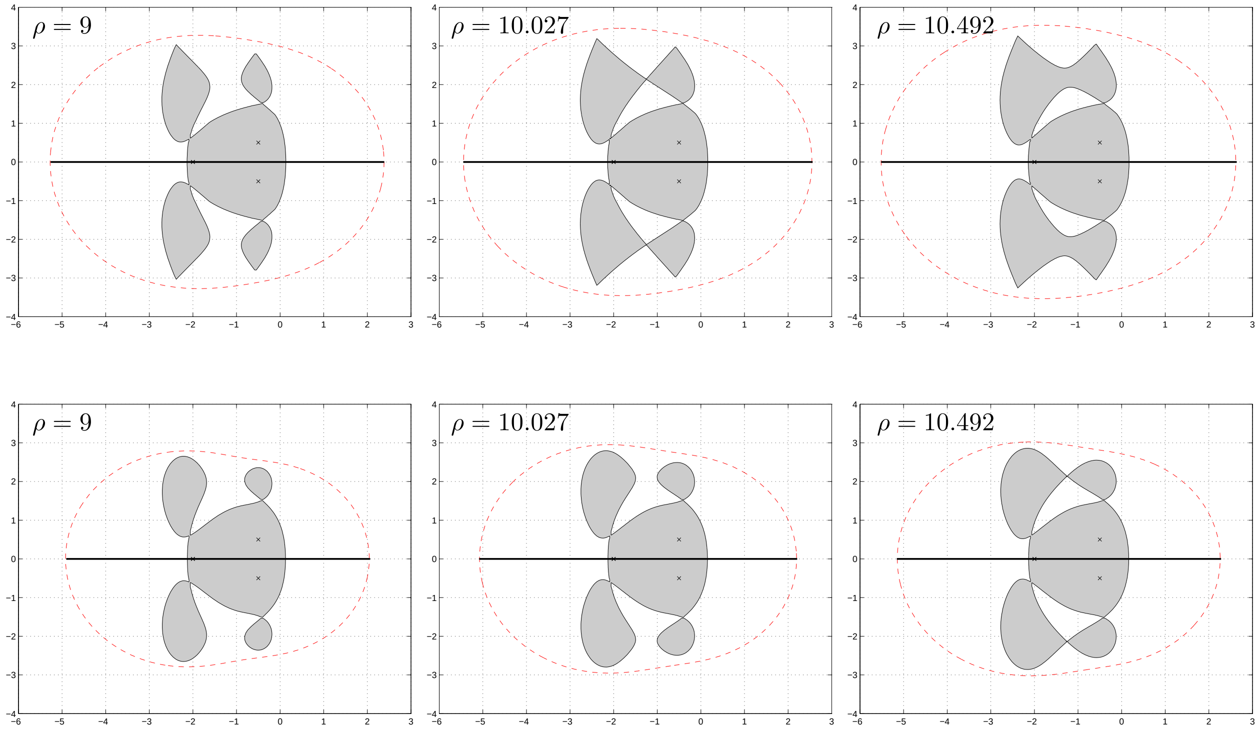


Figure 7.10: The root sets $\mathcal{R}_{\mathbb{R}}^{\|\cdot\|_\alpha}(p; q_1, q_2, \rho)$ for the polynomials p, q_1, q_2 defined in (7.40). Upper row: $\alpha = \infty$. Lower row: $\alpha = 2$. The crosses mark the zeros of p.
.

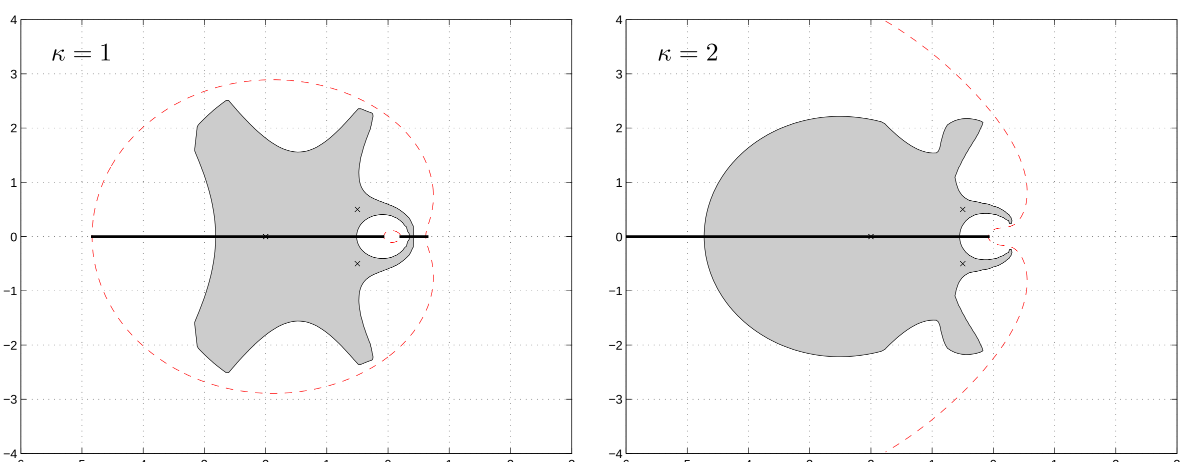


Figure 7.11: The root sets of the interval polynomials $\mathcal{R}[\alpha^{(\kappa)}, \beta^{(\kappa)}]$ and the boundaries of the sets $\mathcal{R}^{\mathbb{C}}[\alpha^{(\kappa)}, \beta^{(\kappa)}]$ (dashed curves) for $\alpha^{(\kappa)}, \beta^{(\kappa)}$ defined in (7.42).
.

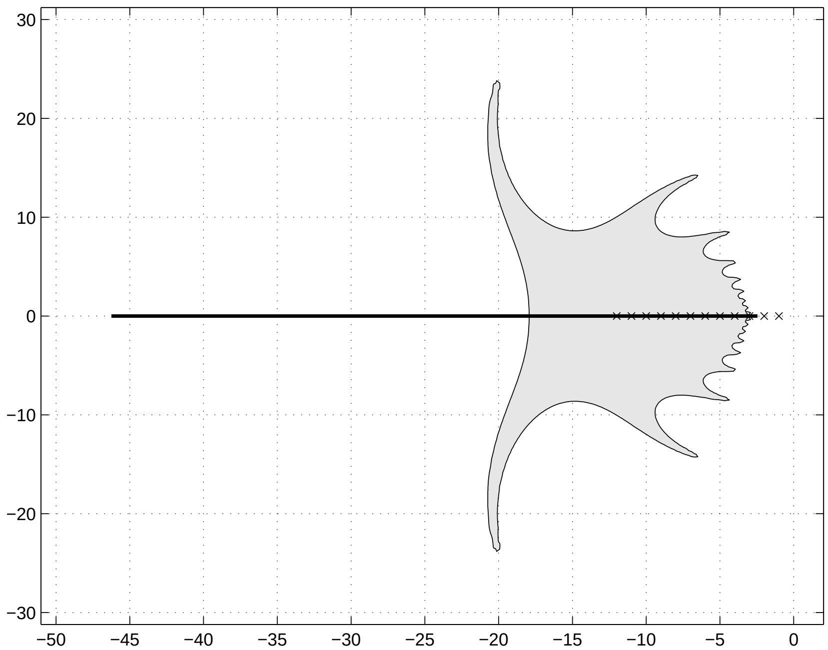


Figure 7.12: The root set of the interval polynomial defined in (7.43). The crosses mark the zeros of the center polynomial.
.

Appendix A

Hermitian-symmetric inequalities

The appendix deals with hermitan-symmetric inequalities. The material is organized as follows. In Section A.1 we derive a lower bound for the positive semi-definite inertia of a hermitian symmetric matrix pair. The latter yields a lower bound for real perturbation values which extends the result of Corollary 6.7.2. In Section A.2 we discuss the canonical form for hermitian-symmetric matrix pairs. The canonical form is needed in the proof of the BQR-theorem (Theorem 6.1.11) which will be given in Section A.3.

Recall from Chapter 6 the notation

$$\mathcal{HS}_n = \left\{ \, (H,S) \;\middle|\; H,S \in \mathbb{C}^{n\times n},\;\; H^* = H,\; S^T = S \, \right\},$$

and for $(H,S)\in\mathcal{HS}_n$,

$$\begin{array}{rcll} i_{\geq}(H,S) & = & \max\{\, k\in\mathbb{N}_0 & \mid \text{There exists a } k\text{-dimensional subspace } \mathcal{X} \text{ of } \mathbb{C}^n \\ & & & \quad \text{such that } x^*Hx \geq |x^TSx| \text{ for all } x\in\mathcal{X} \,\}, \\ i_{\geq}(H) & = & \max\{\, k\in\mathbb{N}_0 & \mid \text{There exists a } k\text{-dimensional subspace } \mathcal{X} \text{ of } \mathbb{C}^n \\ & & & \quad \text{such that } x^*Hx \geq 0 \text{ for all } x\in\mathcal{X} \,\}. \end{array}$$

In the sequel $\lfloor\cdot\rfloor$ denotes the entire function, i.e. for all $x\in\mathbb{R}$, $\lfloor x\rfloor = \max\{\, k\in\mathbb{Z} \mid k\leq x \,\}$.

A.1 Lower bounds for $i_{\geq}(H,S)$ and $\tau_k(M)$

The aim of this section is to derive the following estimates.

Proposition A.1.1 *For all $(H,S)\in\mathcal{HS}_n$, $\quad i_{\geq}(H,S) \geq \lfloor i_{\geq}(H)/2\rfloor$.*

Proposition A.1.2 *For every $M\in\mathbb{C}^{q\times l}$ and $k\in\mathbb{N}$, $\quad \tau_k(M) \geq \sigma_{2k}(M)$.*

The proof is based on the lemma below.

Lemma A.1.3 *Let $S\in\mathbb{C}^{n\times n}$ be a symmetric matrix. Then each subspace $\mathcal{X}$ of $\mathbb{C}^n$ contains a subspace $\mathcal{X}'$ of dimension $\dim\mathcal{X}' \geq \lfloor\dim\mathcal{X}/2\rfloor$ such that $x^TSx = 0$ for all $x\in\mathcal{X}'$.*

Proof: The mapping $\mathcal{X}\times\mathcal{X} \ni (x,y) \mapsto x^TSy$ is a symmetric bilinear form on $\mathcal{X}$ of rank r, say. As is well known (see e.g. [32]) $\mathcal{X}$ has a basis $\mathfrak{B} = (x_1,\ldots,x_k)$ such that

$$x_j^TSx_l = \begin{cases} 1 & \text{if } j = l \leq r, \\ 0 & \text{otherwise.} \end{cases}$$

For $j=1,\ldots,\lfloor r/2\rfloor$ set $x_j' := x_j + i\,x_{\lfloor r/2\rfloor+j}$. Then $\mathcal{X}' := \operatorname{span}\{x_1',\ldots,x_{\lfloor r/2\rfloor}', x_{r+1},\ldots,x_k\}$ has the required properties. $\square$

Proof of Propositions A.1.1 Let $(H,S)\in\mathcal{HS}_n$ and let $\mathcal{X}\subseteq\mathbb{C}^n$ be a subspace such that $\dim\mathcal{X} = i_{\geq}(H)$ and $x^*Hx\geq 0$ for all $x\in\mathcal{X}$. By the lemma above $\mathcal{X}$ contains a subspace $\mathcal{X}'$ of dimension $\dim\mathcal{X}'\geq\lfloor\dim\mathcal{X}/2\rfloor$ such that $x^TSx=0$ for every $x\in\mathcal{X}'$. The latter implies $x^*Hx\geq|x^TSx|$ for $x\in\mathcal{X}'$. Hence $i_{\geq}(H,S)\geq\dim\mathcal{X}'\geq\lfloor i_{\geq}(H)/2\rfloor$. $\square$

Proof of Proposition A.1.2 Let $M\in\mathbb{C}^{q\times l}$ and $k\in\mathbb{N}$. If $2k>\min\{l,q\}$ then $\sigma_{2k}(M)=0$, and the statement of Proposition A.1.2 is trivial. Assume that $2k\leq\min\{l,q\}$. Choose pairwise orthogonal and normalized right singular vectors $v_1,\ldots,v_{2k}$ to the $2\,k$ largest singular values of M, i.e. vectors such that

$$M^*M\,v_j=\sigma_j^2(M)\,v_j,\qquad v_i^*v_j=\begin{cases}1 & \text{if } i=j,\\ 0 & \text{otherwise,}\end{cases}\qquad i,j\in\underline{2k}.$$

Then we have for every $x\in\mathcal{X}:=\operatorname{span}\{v_1,\ldots,v_{2\,k}\}$ that $x^*(M^*M-\sigma_{2\,k}^2(M)I_l)x\geq 0$. Thus $i_{\geq}(M^*M-\sigma_{2\,k}^2(M)I_l)\geq 2k$. Now, Proposition A.1.1 yields,

$$i_{\geq}(\,M^*M-\sigma_{2\,k}^2(M)I_l,\;M^TM-\sigma_{2\,k}^2(M)I_l\,)\geq k.$$

Hence, $\tau_k(M)\geq\sigma_{2k}(M)$ by Theorem 6.1.7. $\square$

A.2 The canonical form of hermitian-symmetric matrix pairs

In this section we discuss the canonical form for hermitian-symmetric matrix pairs under the simultaneous congruence transformations

$$(H,S)\longmapsto(\,P^*HP\,,\;P^TSP\,),\qquad P\in\mathbb{C}^{n\times n}\text{ invertible.}$$

The geometric meaning of such transformations is the following. Let V be an n-dimensional complex vector space endowed with a hermitian form $h:V\times V\to\mathbb{C}$ and a symmetric bilinear form $s:V\times V\to\mathbb{C}$. Let $\mathcal{B}$ be a basis of V and let $[v]_{\mathcal{B}}\in\mathbb{C}^n$ denote the vector of the coordinates of $v\in V$ with respect to $\mathcal{B}$. Then (as is well known) there exists a pair $(H,S)\in\mathcal{HS}_n$ such that

$$h(v,w)=[v]_{\mathcal{B}}^*\,H\,[w]_{\mathcal{B}},\qquad s(v,w)=[v]_{\mathcal{B}}^T\,S\,[w]_{\mathcal{B}}\qquad\text{for all } v,w\in V.$$

Let $\mathcal{B}'$ be a second basis of V and suppose that the change of coordinates $[v]_{\mathcal{B}'}\mapsto[v]_{\mathcal{B}}$ is given by $[v]_{\mathcal{B}}=P[v]_{\mathcal{B}'}$ for all $v\in V$, where P is an invertible $n\times n$ matrix. Then

$$h(v,w)=[v]_{\mathcal{B}'}^*P^*HP\,[w]_{\mathcal{B}'},\qquad s(v,w)=[v]_{\mathcal{B}'}^TP^TSP\,[w]_{\mathcal{B}'}\qquad\text{for all } v,w\in V.$$

Thus the pairs (H,S) and $(\,P^*HP\,,\;P^TSP\,)$ represent the same pair of forms, (h,s), in different coordinates.

In order to understand the canonical form which will be given below one first has to understand another issue, namely the basics of the theory of antilinear operators. An antilinear operator between complex vector spaces V_1 and V_2 is a map $f:V_1\to V_2$ with the properties

$$f(\lambda v)=\overline{\lambda}f(v),\qquad f(v+w)=f(v)+f(w)\qquad\text{for all } v,w\in V_1,\ \lambda\in\mathbb{C}.$$

Let $C\in\mathbb{C}^{n\times n}$. Then the map

$$f_C : \mathbb{C}^n \to \mathbb{C}^n \qquad f_C(x) := C\,\overline{x}$$

is antilinear. Conversely, let $f : V \to V$ be an antilinear operator, where V is an n-dimensional complex vector space with basis $\mathcal{B}$. Then there exists a matrix $C\in\mathbb{C}^{n\times n}$ such that

$$[f(v)]_{\mathcal{B}} = C\overline{[v]_{\mathcal{B}}} = f_C([v]_{\mathcal{B}})$$

for all $v\in V$. Let $\mathcal{B}'$ be a second basis of V, and let P be the matrix with $[v]_{\mathcal{B}} = P[v]_{\mathcal{B}'}$ for all $v\in V$. Then we have

$$[f(v)]_{\mathcal{B}'} = P^{-1}C\overline{P}\,\overline{[v]_{\mathcal{B}'}}.$$

Thus the matrices C and $P^{-1}C\overline{P}$ represent the same antilinear operator with respect to different bases. Transformations of the form

$$C \longmapsto P^{-1}C\overline{P}, \qquad P\in\mathbb{C}^{n\times n} \text{ invertible}$$

are called consimilarity transformations (see [61, Section 4.6]). The square $f^2 = f\circ f$ of an antilinear operator $f : V \to V$ is a linear. For the squares of the operators f_C, $C\in\mathbb{C}^{n\times n}$, we have that

$$f_C^2(x) = f_C(f_C(x)) = C\overline{f_C(x)} = C\overline{C\overline{x}} = C\overline{C}x \qquad \text{for all } x\in\mathbb{C}^n.$$

Thus $C\overline{C}$ is the matrix belonging to the linear operator f_C^2. The theory of antilinear operators is well developed [35, 61, 57, 58, 59, 60, 66]. In particular there are several canonical forms for a matrix C under consimilarity transformations which can be obtained from the Jordan canonical form of $C\overline{C}$. The connection of hermitian-symmetric matrix pairs with antilinear operators and consimilarity transformations is the following.

Proposition A.2.1 *Let $(H,S)\in\mathcal{HS}_n$, where S is invertible, and let $C := S^{-1}\overline{H}$. Then*

$$x^*Hy = (C\overline{x})^T Sy \qquad \textit{for all } x,y\in\mathbb{C}^n.$$

Moreover, for all invertible $P\in\mathbb{C}^{n\times n}$,

$$P^*HP = (P^{-1}C\overline{P})^T(P^TSP).$$

Let $O_{H,S} := \left\{\, (P^*HP\,,\, P^TSP\,) \;\middle|\; P\in\mathbb{C}^{n\times n} \text{ invertible} \,\right\}$ denote the kongruence orbit of $(H,S)\in\mathcal{HS}_n$. The canonical form problem for hermitian-symmetric matrix pairs is, to find a set $\mathcal{F}\subset\bigcup_{n\in\mathbb{N}}\mathcal{HS}_n$ and a bijective map $\Phi : \{\, O_{H,S} \mid (H,S)\in\mathcal{HS}_n,\ n\in\mathbb{N}\,\} \to \mathcal{F}$ such that the elements of $\mathcal{F}$ have a structure, which is as simple as possible. $\Phi(O_{H,S})$ is then said to be a canonical form of (H,S). Proposition A.2.1 indicates that this problem is closely related to consimilarity transformations. A canonical form for hermitian-symmetric matrix pairs (H,S), where S is invertible, is given in the theorem below. In order to formulate it we introduce the following notation.

Notation A.2.2 *For $m\in\mathbb{N}$ and $A_j,B_j,C_j\in\mathbb{C}^{n_j\times n_j}$, $A,B,C\in\mathbb{C}^{n\times n}$ let*

$$\begin{aligned}
\bigoplus_{j=1}^{m}(A_j,B_j,C_j) &:= (\,\mathrm{diag}(A_1,\dots,A_m)\,,\,\mathrm{diag}(B_1,\dots,B_m)\,,\,\mathrm{diag}(C_1,\dots,C_m)\,)\\
\oplus^m(A,B,C) &:= (\mathrm{diag}(\underbrace{A,\dots,A}_{m-times}),\mathrm{diag}(\underbrace{B,\dots,B}_{m-times}),\mathrm{diag}(\underbrace{C_1,\dots,C}_{m-times})\,).
\end{aligned}$$

(Direct sum of matrix triples)

Moreover, let E_k denote the $k \times k$ permutation matrix with 1's on the antidiagonal,

$$E_k := \begin{bmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{bmatrix} \in \{0,1\}^{k\times k}, \qquad E_1 = [1],$$

and let $J_k(\mu)$ denote the Jordan block

$$J_k(\mu) := \begin{bmatrix} \mu & 1 & & \\ & \ddots & \ddots & \\ & & & 1 \\ & & & \mu \end{bmatrix} \in \mathbb{C}^{k\times k}.$$

Note that

$$J_k(\mu)^T E_k = E_k J_k(\mu) = \begin{bmatrix} 0 & & & \mu \\ & & & 1 \\ & \iddots & \iddots & \\ \mu & 1 & & 0 \end{bmatrix}.$$

The theorem below has been shown in [66, 63], see also [6].

Theorem A.2.3 *Let $(H,S) \in \mathcal{HS}_n$, where S is invertible, and let $C := S^{-1}\overline{H}$. Then the spectrum of $C\overline{C}$ is symmetric to the real axis and the negative eigenvalues of $C\overline{C}$ have even algebraic multiplicity. Let r denote the number of (pairwise different) nonnegative eigenvalues of $C\overline{C}$ and let s denote the number of that (pairwise different) eigenvalues of $C\overline{C}$ which are either negative or have positive imaginary part. Then there exists an invertible $P \in \mathbb{C}^{n\times n}$ such that*

$$\begin{aligned}
&(P^T S P\ ,\ P^{-1} C \overline{P}\ ,\ P^* H P\ ) \\
&= \bigoplus_{j=1}^{r} \bigoplus_{k=1}^{n} \oplus^{n^+_{jk}} \left(\ E_k,\ \ J_k(\lambda_j)\ ,\ \ E_k\, J_k(\lambda_j)\ \right) \\
&\oplus \\
&\bigoplus_{j=1}^{r} \bigoplus_{k=1}^{n} \oplus^{n^-_{jk}} \left(\ E_k,\ \ -J_k(\lambda_j)\ ,\ \ -E_k\, J_k(\lambda_j)\ \right) \\
&\oplus \\
&\bigoplus_{j=1}^{s} \bigoplus_{k=1}^{\lfloor \frac{n}{2} \rfloor} \oplus^{m_{jk}} \left(\begin{bmatrix} E_k & 0 \\ 0 & E_k \end{bmatrix},\ \begin{bmatrix} 0 & J_k(\nu_j) \\ J_k(\overline{\nu_j}) & 0 \end{bmatrix},\ \begin{bmatrix} 0 & E_k\, J_k(\overline{\nu_j}) \\ E_k\, J_k(\nu_j) & 0 \end{bmatrix} \right),
\end{aligned}$$

where

(a) $\lambda_1, \ldots, \lambda_r \geq 0$ *and* $\lambda_1^2, \ldots, \lambda_r^2$ *are the (pairwise different) nonnegative eigenvalues of* $C\overline{C}$,

(b) $\nu_1, \ldots, \nu_s \in \mathbb{C} \setminus \mathbb{R}$ *and* $\nu_1^2, \ldots, \nu_s^2$ *are the (pairwise different) eigenvalues of* $C\overline{C}$ *which are either negative or have positive imaginary part,*

(c) $n^-_{jk} = 0$ *if* $\lambda_j = 0$ *and* k *is odd.*

If one of the numbers $r, s, n_{jk}^{+}, n_{jk}^{-}, m_{jk} \in \mathbb{N}$ *is zero then the accompanying sum is to be omitted. In such a decomposition the multiplicities* $n_{jk}^{+}, n_{jk}^{-}, m_{jk}$ *of the blocks are uniquely determined by* H *and* S.

Remark A.2.4 *Another canonical form for hermitian-symmetric pairs* (H, S) *with* $\det(S) \neq 0$ *can be found in [59, 60]. A canonical form for the case* $\det(S) = 0$ *has been given in [27]. See also [66].*

We will not reproduce the lengthy proof of Theorem A.2.3 here. For our purpose, the proof of the BQR-theorem, we only need the canonical form in the generic case, whose proof is comparatively simple. The canonical form for the generic case is given in the next theorem. We need a further notation.

Notation A.2.5

$$\mathcal{HS}_n^+ \;:=\; \{\,(H,S)\in\mathcal{HS}_n \;\mid\; \det(S)\neq 0,\quad \textit{all eigenvalues of } C\overline{C} \textit{ are simple },\ \textit{where } C := S^{-1}\overline{H}\}.$$

We will show at the end of this section that $\mathcal{HS}_n^+$ is an open and dense subset of $\mathcal{HS}_n$.

Theorem A.2.6 *Let* $(H, S)\in\mathcal{HS}_n^+$ *and* $C = S^{-1}\overline{H}$. *Then* $C\overline{C}$ *has no negative eigenvalues, the nonreal eigenvalues of* $C\overline{C}$ *occur in conjugate complex pairs and there exists an invertible matrix* $P\in\mathbb{C}^{n\times n}$ *such that*

$$\begin{aligned}
P^{-1}C\overline{P} &= \operatorname{diag}(\lambda_1,\ldots,\lambda_r) \oplus \bigoplus_{k=1}^{s} \begin{bmatrix} 0 & \nu_k, \\ \overline{\nu_k} & 0 \end{bmatrix}, \\
P^T S P &= I_n, \\
P^* H P &= \operatorname{diag}(\lambda_1,\ldots,\lambda_r) \oplus \bigoplus_{k=1}^{s} \begin{bmatrix} 0 & \overline{\nu_k} \\ \nu_k & 0 \end{bmatrix},
\end{aligned}$$

where

(*a*) $\lambda_1,\ldots,\lambda_r\in\mathbb{R}$ *and* $\lambda_1^2,\ldots,\lambda_r^2$ *are the real eigenvalues of* $C\overline{C}$,

(*b*) $\Im\nu_1,\ldots,\Im\nu_s > 0$ *and* $\nu_1^2,\ldots,\nu_s^2,\overline{\nu_1}^2,\ldots,\overline{\nu_s}^2$ *are the nonreal eigenvalues of* $C\overline{C}$.

The columns of P *form a basis of eigenvectors of* $C\overline{C}$.

Proof: (following the proof in [9])

(*i*) Suppose $z\in\mathbb{C}^n\setminus\{0\}$ is an eigenvector of $C\overline{C}$ to the eigenvalue $\mu\in\mathbb{C}$, i.e. $C\overline{C}z = \mu z$. Then

$$C\overline{C}(C\overline{z}) = \overline{\overline{C}C\overline{C}z} = \overline{\overline{C}\mu z} = \overline{\mu}C\overline{z}.$$

Thus either $C\overline{z}$ is an eigenvector of $C\overline{C}$ to the eigenvalue $\overline{\mu}$ or $C\overline{z} = 0$. The latter can happen only if $\mu = 0$. Suppose now, that μ is real. Then the assumption that the eigenvalues of $C\overline{C}$ are simple implies that $C\overline{z}$ is a scalar multiple of z, say $C\overline{z} = \lambda z,\ \lambda\in\mathbb{C}$. Hence, $|\lambda|^2 x = \overline{\lambda}C\overline{x} = C\overline{\lambda x} = C\overline{C}x = \mu x$. Thus μ is nonnegative.

(*ii*) It is easily verified that the following relation holds for $C = S^{-1}\overline{H}$.

$$S(C\overline{C}) = \overline{C^T S C} = (C\overline{C})^T S. \tag{A.1}$$

Suppose that $z_1, z_2 \in \mathbb{C}^n\setminus\{0\}$ are eigenvectors of $C\overline{C}$, i.e. $C\overline{C}z_k = \mu_k z_k$ with $\mu_k\in\mathbb{C}$, $k = 1, 2$. Then the relations (A.1) yield

$$\mu_1\, z_1^T S z_2 = (C\overline{C} z_1)^T S z_2 = z_1^T (C\overline{C})^T S z_2 = z_1^T S (C\overline{C}) z_2 = \mu_2\, z_1^T S z_2.$$

This implies

$$z_1^T S z_2 = 0 \text{ if } \mu_1 \neq \mu_2. \tag{A.2}$$

(*iii*) Let $\mu_1, \ldots, \mu_r$ be the real eigenvalues of $C\overline{C}$ and let $\mu_{r+1}, \ldots \mu_{r+s}$ be the eigenvalues of $C\overline{C}$ with positive imaginary part. By (*i*) and our assumption that the eigenvalues of $C\overline{C}$ are algebraically simple, there exists an invertible matrix

$$Q = [u_1, \ldots, u_r, v_1, C\overline{v_1}, \ldots, v_s, C\overline{v_s}] \in \mathbb{C}^{n\times n}, \qquad r + 2s = n,$$

whose columns form a basis of eigenvectors of $C\overline{C}$, i.e.

$$\begin{aligned} C\overline{C} u_j &= \mu_j u_j \quad 1 \le j \le r, \\ C\overline{C} v_j &= \mu_{r+j} v_j \quad 1 \le j \le s, \\ C\overline{C}(C\overline{v_j}) &= \overline{\mu_{r+j}}\, C\overline{v_j} \quad 1 \le j \le s. \end{aligned}$$

Since $\mu_j \neq \mu_k \neq \overline{\mu_j}$ for $1 \le j < k \le r+s$ and $\mu_{r+j} \neq \overline{\mu_{r+j}}$ for $1 \le j \le s$ we have by (A.2) that the eigenvectors are pairwise orthogonal with respect to the symmetric bilinear form associated with S. Hence,

$$Q^T S Q = \operatorname{diag}\left(u_1^T S u_1, \ldots, u_r^T S u_r, v_1^T S v_1, (C\overline{v_1})^T S (C\overline{v_1}), \ldots, v_s^T S v_s, (C\overline{v_s})^T S (C\overline{v_s})\right). \tag{A.3}$$
(A.4)

Furthermore, (A.1) yields that for $1 \le j \le s$,

$$(C\overline{v_j})^T S (C\overline{v_j}) = \overline{v_j^T \overline{C^T S C} v_j} = \overline{v_j^T S (C\overline{C}) v_j} = \overline{\mu_{r+j}}\, \overline{v_j^T S v_j}.$$

(*iv*) Since S is invertible, all diagonal elements of $Q^T S Q$ are nonzero. Choose $\alpha_j, \beta_j, \nu_j \in \mathbb{C}\setminus\{0\}$ such that

$$\begin{aligned} \alpha_j^2 &= u_j^T S u_j \quad \text{for } 1 \le j \le r, \\ \beta_j^2 &= v_j^T S v_j \quad \text{for } 1 \le j \le s, \\ \nu_j^2 &= \mu_{r+j} \quad \text{for } 1 \le j \le s, \end{aligned}$$

and define

$$\begin{aligned} x_j &:= \alpha_j^{-1} u_j && \text{for } 1 \le j \le r, \\ y_j &:= \beta_j^{-1} v_j && \text{for } 1 \le j \le s, \\ z_j &:= \overline{\nu_j^{-1} \beta_j^{-1}}\, C\overline{v_j} && \text{for } 1 \le j \le s \\ &= \overline{\nu_j^{-1}}\, C\overline{y_j}, \end{aligned}$$

and

$$\begin{aligned} P &:= [x_1, \ldots, x_r, y_1, z_1, \ldots, y_s, z_s] \\ &= Q \operatorname{diag}\left(\alpha_1^{-1}, \ldots, \alpha_r^{-1}, \beta_1^{-1}, \overline{\nu_1^{-1}\beta_1^{-1}}, \ldots, \beta_s^{-1}, \overline{\nu_s^{-1}\beta_s^{-1}}\right). \end{aligned}$$

By $(A.4)$ we have

$$Q^T SQ = \mathrm{diag}(\alpha_1^2, \dots, \alpha_r^2,\, \beta_1^2,\, \overline{\nu_1^2\beta_1^2}, \dots, \beta_s,\, \overline{\nu_s^2\beta_s^2}\,).$$

Therefore,

$$P^T SP = I_n.$$

Moreover, the vectors y_j, z_j satisfy

$$C\overline{y_j} = \overline{\nu_j}\, z_j, \qquad C\overline{z_j} = \nu_j\, y_j. \tag{A.5}$$

Since the vectors x_j are eigenvectors of $C\overline{C}$ to real eigenvalues our considerations in (i) yield that

$$C\overline{x_j} \;=\; \lambda_j x_j \tag{A.6}$$

with $\lambda_j \in \mathbb{C}$, $|\lambda_j|^2 = \mu_j$. The relations (A.5) and (A.6) are equivalent to the matrix equation

$$P^{-1}C\overline{P} \;=\; \mathrm{diag}(\lambda_1, \dots, \lambda_r) \oplus \bigoplus_{k=1}^{s} \begin{bmatrix} 0 & \nu_k \\ \overline{\nu_k} & 0 \end{bmatrix}.$$

Consequently,

$$P^* HP = (P^{-1}C\overline{P})^T (P^T SP) = \mathrm{diag}(\lambda_1, \dots, \lambda_r) \oplus \bigoplus_{k=1}^{s} \begin{bmatrix} 0 & \overline{\nu_k} \\ \nu_k & 0 \end{bmatrix}.$$

Since P^*HP is hermitian we have that $\lambda_1, \dots, \lambda_r \in \mathbb{R}$. $\square$

For the proof of the BQR-theorem we also need the following fact.

Proposition A.2.7 *The set $\mathcal{HS}_n^+$ is an open and dense subset of $\mathcal{HS}_n$.*

Proof: For $A \in \mathbb{C}^{n\times n}$ let $\mathrm{adj}(A)$ denote the classical adjoint of $A \in \mathbb{C}^{n\times n}$. If A is invertible then $A^{-1} = \mathrm{adj}(A)/\det(A)$. Moreover let $\mathrm{dis}(A)$ denote the discriminant of the characteristic polynomial $\chi_A(s) = \det(sI_n - A)$. Then $\mathrm{dis}(A)$ being the resultant of χ_A and χ_A' is a polynomial in the entries of A and we have $\mathrm{dis}(A) \neq 0$ if and only if all eigenvalues of A are simple. Now define for all $(H, S) \in \mathcal{HS}_n$,

$$p(H, S) := \mathrm{dis}(\,\mathrm{adj}(S)\overline{H}\,\mathrm{adj}(\overline{S})\,H\,).$$

Then p is a polynomial in the real and complex parts of the entries of H and S. If $\det(S) \neq 0$ and $C = S^{-1}\overline{H}$ then $\det(S)C = \mathrm{adj}(S)\overline{H}$ and consequently $|\det(S)|^2 C\overline{C} = \mathrm{adj}(S)\overline{H}\,\mathrm{adj}(\overline{S})\,H$. Thus $\mathcal{HS}_n^+ = \{\, (H, S) \in \mathcal{HS}_n \mid \det(S) \neq 0,\ p(H, S) \neq 0 \,\}$. The complement of $\mathcal{HS}_n^+$

$$\mathcal{HS}_n \setminus \mathcal{HS}_n^+ = \{\, (H, S) \in \mathcal{HS}_n \mid \det(S)\, p(H, S) = 0 \,\}$$

is a real algebraic subset of $\mathcal{HS}_n$. So either $\mathcal{HS}_n^+$ is open and dense in $\mathcal{HS}_n$ or $\mathcal{HS}_n^+ = \emptyset$. The latter fails since $(\,I_n\,,\ \mathrm{diag}(1, \dots, n)\,) \in \mathcal{HS}_n^+$.
$\square$

A.3 Proof of the BQR-Theorem

In this section we derive the BQR-theorem (Theorem 6.1.11) which states that for each $k \in \mathbb{N}$ and each $(H,S) \in \mathcal{HS}_n$ the following statements (A_k) and (B_k) are equivalent.

(A_k) There exists a k-dimensional subspace $\mathcal{X}$ of $\mathbb{C}^n$ such that $x^*Hx \geq |x^TSx|$ for all $x \in \mathcal{X}$.

(B_k) For all $\alpha \in (-1,0)$: $\quad \lambda_{2k}\left(\begin{bmatrix} H & \alpha\overline{S} \\ \alpha S & \overline{H} \end{bmatrix}\right) \geq 0.$

Since

$$\begin{bmatrix} I_n & 0 \\ 0 & -I_n \end{bmatrix}^{-1} \begin{bmatrix} H & \alpha\overline{S} \\ \alpha S & \overline{H} \end{bmatrix} \begin{bmatrix} I_n & 0 \\ 0 & -I_n \end{bmatrix} = \begin{bmatrix} H & -\alpha\overline{S} \\ -\alpha S & \overline{H} \end{bmatrix}$$

we have for all $\alpha \in \mathbb{R}$,

$$\lambda_{2k}\left(\begin{bmatrix} H & \alpha\overline{S} \\ \alpha S & \overline{H} \end{bmatrix}\right) = \lambda_{2k}\left(\begin{bmatrix} H & -\alpha\overline{S} \\ -\alpha S & \overline{H} \end{bmatrix}\right)$$

Thus in (B_k) the interval $(-1,0)$ can be replaced by the interval $(0,1)$. Furthermore by continuity of the eigenvalues, the interval $(0,1)$ can be replaced by the closed interval $[0,1]$. In summary, condition (B_k) is equivalent to the following conditon (B'_k).

(B'_k) For all $\alpha \in [0,1]$: $\quad \lambda_{2k}\left(\begin{bmatrix} H & \alpha\overline{S} \\ \alpha S & \overline{H} \end{bmatrix}\right) \geq 0.$

The conditions (A_k) and (B'_k) can be rewritten in the following form.

(A_k) $i_{\geq}(H,S) \geq k$.

(B'_k) $\min_{\alpha \in [0,1]} i_{\geq}\left(\begin{bmatrix} H & \alpha\overline{S} \\ \alpha S & \overline{H} \end{bmatrix}\right) \geq 2k.$

Thus the theorem below is just a reformulation of the BQR-theorem.

Theorem A.3.1 *For every $(H,S) \in \mathcal{HS}_n$:*

$$i_{\geq}(H,S) = \left\lfloor \frac{1}{2} \min_{\alpha \in [0,1]} i_{\geq}\left(\begin{bmatrix} H & \alpha\overline{S} \\ \alpha S & \overline{H} \end{bmatrix}\right) \right\rfloor. \tag{A.7}$$

We split the proof of the BQR-theorem in two lemmas and a proposition.

Lemma A.3.2 *Let $(H,S) \in \mathcal{HS}_n$. Then, for every $k \in \underline{n}$, condition (A_k) implies (B'_k). Equivalently,*

$$i_{\geq}(H,S) \leq \left\lfloor \frac{1}{2} \min_{\alpha \in [0,1]} i_{\geq}\left(\begin{bmatrix} H & \alpha\overline{S} \\ \alpha S & \overline{H} \end{bmatrix}\right) \right\rfloor. \tag{A.8}$$

Proof: Let $\mathcal{X}$ be a k-dimensional subspace of $\mathbb{C}^n$ such that $x^*Hx \geq |x^TSx|$ for all $x \in \mathcal{X}$. Then we also have $x^*Hx \geq |x^T(\alpha S)x|$ for all $x \in \mathcal{X}$ and all $\alpha \in [0,1]$. Now Lemma 6.4.2 yields that

$$\begin{bmatrix} x \\ \overline{y} \end{bmatrix}^* \begin{bmatrix} H & \alpha\overline{S} \\ \alpha S & \overline{H} \end{bmatrix} \begin{bmatrix} x \\ \overline{y} \end{bmatrix} \geq 0 \quad \text{for all} \quad \begin{bmatrix} x \\ \overline{y} \end{bmatrix} \in \mathcal{X}' := \left\{ \begin{bmatrix} \xi \\ \overline{\eta} \end{bmatrix} \;\middle|\; \xi, \eta \in \mathcal{X} \right\}.$$

$\mathcal{X}'$ is a complex vector space of dimension $\dim \mathcal{X}' = 2k$. Thus (B'_k) holds. □

Lemma A.3.3 *Suppose there is a dense subset D_n of $\mathcal{HS}_n$ such that (B_k') implies (A_k) for all $(H,S)\in D_n$. Then (B_k') implies (A_k) for all $(H,S)\in\mathcal{HS}_n$.*

Proof: Let $(H,S)\in\mathcal{HS}_n$ and suppose that condition (B_k') holds for (H,S). Then for all $\alpha\in[0,1]$ and all $j\in\mathbb{N}$,

$$\lambda_{2k}\left(\begin{bmatrix} H+(1/j)I_n & \alpha\overline{S} \\ \alpha S & \overline{H+(1/j)I_n}\end{bmatrix}\right) = \lambda_{2k}\left(\begin{bmatrix} H & \alpha\overline{S} \\ \alpha S & \overline{H}\end{bmatrix}\right) + \frac{1}{j} > 0.$$

From the continuity of eigenvalues and the conpactness of the interval $[0,1]$ it follows that there is an $\epsilon\in(0,1/j)$ such that

$$\min_{\alpha\in[0,1]} \lambda_{2k}\left(\begin{bmatrix} H' & \alpha\overline{S'} \\ \alpha S' & \overline{H'}\end{bmatrix}\right) > 0$$

for all $(H',S')\in\mathcal{HS}_n$ satisfying $\|H+(1/j)I_n-H'\|_2+\|S-S'\|_2<\epsilon$. Since D_n is dense in $\mathcal{HS}_n$ there is $(H_j,S_j)\in D_n$ satisfying $\|H+(1/j)I_n-H_j\|_2+\|S-S_j\|_2<\epsilon$. Thus

$$\min_{\alpha\in[0,1]} \lambda_{2k}\left(\begin{bmatrix} H_j & \alpha\overline{S_j} \\ \alpha S_j & \overline{H_j}\end{bmatrix}\right) > 0.$$

Hence, condition (B_k') holds for $(H_j,S_j)\in D_n$. By assumption this implies that there exists a k-dimensional subspace $\mathcal{X}_j$ of $\mathbb{C}^n$ such that $x^*H_jx\geq|x^TS_jx|$ for all $x\in\mathcal{X}_j$. Let $X_j\in\mathbb{C}^{n\times k}$ be a matrix whose columns form an orthogonal basis of $\mathcal{X}_j$. Then $X_j^TX_j=I_k$, $\operatorname{range}X_j=\mathcal{X}_j$ and

$$(X_j\xi)^*H_j(X_j\xi)\geq|(X_j\xi)^TS_j(X_j\xi)| \quad \text{for all } \xi\in\mathbb{C}^k. \tag{A.9}$$

The set of $X\in\mathbb{C}^{n\times k}$ satisfying $X^*X=I_k$ is compact. Thus the sequence X_j has a convergent subsequence. Let $X_0\in\mathbb{C}^{n\times k}$ denote its limit. Then $X_0^*X_0=I_k$, i.e. the columns of X_0 form an orthogonal basis of the k -dimensional subspace $\mathcal{X}_0:=\operatorname{range}X_0$. Since the sequence (H_j,S_j) converges to (H,S) it follows from (A.9) that

$$(X_0\xi)^*H(X_0\xi)\geq|(X_0\xi)^TS(X_0\xi)| \quad \text{for all } \xi\in\mathbb{C}^k.$$

Equivalently, $x^*Hx\geq|x^TSx|$ for all $x\in\mathcal{X}_0$. Thus we have shown that for every pair $(H,S)\in\mathcal{HS}_n$, condition (B_k') implies (A_k). □

We are now going to show the equivalence of (A_k) and (B_k') for all $(H,S)\in\mathcal{HS}_n^+$. Since $\mathcal{HS}_n^+$ is dense in $\mathcal{HS}_n$ this concludes the proof of the BQR-theorem.

Proposition A.3.4 *Let $(H,S)\in\mathcal{HS}_n^+$. Then there exists an invertible matrix*

$$P=[x_1,\ldots,x_r,y_1,z_1\ldots,y_s,z_s]\in\mathbb{C}^{n\times n}, \qquad x_k,y_k,z_k\in\mathbb{C}^n,$$

such that

$$\begin{aligned} P^TSP &= I_n, \\ P^*HP &= \operatorname{diag}(\lambda_1,\ldots,\lambda_\rho,\lambda_{\rho+1},\ldots,\lambda_{\rho+\ell},\lambda_{\rho+\ell+1},\ldots,\lambda_r)\oplus\bigoplus_{k=1}^s\begin{bmatrix}0 & \overline{\nu_k}\\ \nu_k & 0\end{bmatrix}, \end{aligned} \tag{A.10}$$

where

- $1>\lambda_1\geq\ldots\geq\lambda_\rho\geq-1$,

- $\lambda_{\rho+j} \geq 1$ *for* $1 \leq j \leq \ell$,
- $\lambda_j < -1$ *for* $\rho+\ell+1 \leq j \leq r$
- $\nu_k \in \mathbb{C} \setminus \mathbb{R}$ *with* $\Im \nu_k > 0$ *for all* $k \in \underline{s}$.

Moreover, (A.7) holds for the pair (H,S) *with*

$$i_{\geq}(H,S) = \left\lfloor \frac{1}{2} \min_{\alpha\in[0,1]} i_{\geq}\left(\begin{bmatrix} H & \alpha\overline{S} \\ \alpha S & \overline{H}\end{bmatrix}\right) \right\rfloor = m+\ell+s,$$

where

$$m \;:=\; \max\{\ q\in\mathbb{N}_0 \ \mid\ 2q \leq \rho \text{ and } \lambda_j + \lambda_{2q+1-j} \geq 0 \text{ for all } j \in \underline{q}\ \}.$$

A subspace $\mathcal{X} \subseteq \mathbb{C}^n$ *of maximal dimension such that* $x^*Hx \geq |x^TSx|$ *for all* $x \in \mathcal{X}$ *is given by*

$$\mathcal{X} \;=\; \operatorname{span}\{\, x_k + i\, x_{2m+1-k} \mid k \in \underline{m}\,\} \oplus \operatorname{span}\{x_{\rho+1},\ldots,x_{\rho+\ell}\} \oplus \operatorname{span}\{\, y_k + i\, z_k \mid k\in\underline{s}\,\}.$$

Proof: The existence of the matrix P is guaranteed by Theorem A.2.6. Each $x \in \mathcal{X}$ can be written in the form

$$x = \sum_{k=1}^{m} \alpha_k (x_k + i\, x_{2m+1-k}) + \sum_{k=1}^{\ell} \beta_k\, x_{\rho+k} + \sum_{k=1}^{t} \gamma_k (y_k + i\, z_k) \qquad \alpha_k, \beta_k, \gamma_k \in \mathbb{C},$$

and (A.10) yields

$$\begin{aligned}
x^*Hx \;&=\; \sum_{k=1}^{m} |\alpha_k|^2 \begin{bmatrix}1\\ i\end{bmatrix}^* \begin{bmatrix}\lambda_k & 0 \\ 0 & \lambda_{2m+1-k}\end{bmatrix} \begin{bmatrix}1\\ i\end{bmatrix} + \sum_{k=1}^{\ell} |\beta_k|^2 \lambda_{\rho+k} + \sum_{k=1}^{s} |\gamma_k|^2 \begin{bmatrix}1\\ i\end{bmatrix}^* \begin{bmatrix}0 & \overline{\nu_k} \\ \nu_k & 0\end{bmatrix} \begin{bmatrix}1\\ i\end{bmatrix} \\
&=\; \sum_{k=1}^{m} |\alpha_k|^2 \underbrace{(\lambda_k + \lambda_{2m+1-k})}_{\geq 0} + \sum_{k=1}^{\ell} |\beta_k|^2 \underbrace{\lambda_{\rho+k}}_{\geq 1} + \sum_{k=1}^{s} |\gamma_k|^2\, 2\underbrace{\Im(\nu_k)}_{\geq 0} \\
&\geq\; \sum_{k=1}^{\ell} |\beta_k|^2 \geq \left| \sum_{k=1}^{\ell} \beta_\ell^2 \right|, \\
x^TSx \;&=\; \sum_{k=1}^{m} \alpha_k^2 \underbrace{\begin{bmatrix}1\\ i\end{bmatrix}^T \begin{bmatrix}1 & 0 \\ 0 & 1\end{bmatrix} \begin{bmatrix}1\\ i\end{bmatrix}}_{=0} + \sum_{k=1}^{\ell} \beta_k^2 + \sum_{k=1}^{s} \gamma_k^2 \underbrace{\begin{bmatrix}1\\ i\end{bmatrix}^T \begin{bmatrix}1 & 0 \\ 0 & 1\end{bmatrix} \begin{bmatrix}1\\ i\end{bmatrix}}_{=0} = \sum_{k=1}^{s} \beta_k^2.
\end{aligned}$$

Thus $x^*Hx \geq |x^TSx|$ for all $x \in \mathcal{X}$ and hence $i_{\geq}(H,S) \geq \dim \mathcal{X} = m+\ell+s$. By Lemma A.3.2 we have $\left\lfloor \frac{1}{2}\min_{\alpha\in[0,1]} i_{\geq}\left(\begin{bmatrix} H & \alpha\overline{S} \\ \alpha S & \overline{H}\end{bmatrix}\right)\right\rfloor$. Thus the proposition is proved if we can show that

$$m+\ell+s \geq \left\lfloor \frac{1}{2} \min_{\alpha\in[0,1]} i_{\geq}\left(\begin{bmatrix} H & \alpha\overline{S} \\ \alpha S & \overline{H}\end{bmatrix}\right) \right\rfloor. \tag{A.11}$$

We now compute $i_{\geq}\left(\begin{bmatrix} H & \alpha\overline{S} \\ \alpha S & \overline{H}\end{bmatrix}\right)$ for $\alpha \in \mathbb{R}$. First note that there is a permutation matrix $Q\in\{0,1\}^{2n\times 2n}$ such that for all $\alpha\in\mathbb{R}$,

$$\begin{aligned} Q^*\begin{bmatrix} P & 0 \\ 0 & \overline{P}\end{bmatrix}^* \begin{bmatrix} H & \alpha\overline{S} \\ \alpha S & \overline{H}\end{bmatrix}\begin{bmatrix} P & 0 \\ 0 & \overline{P}\end{bmatrix} Q &= Q^*\begin{bmatrix} P^*HP & \alpha I_n \\ \alpha I_n & \overline{P^*HP}\end{bmatrix} Q \\ &= \bigoplus_{k=1}^{r}\begin{bmatrix} \lambda_k & \alpha \\ \alpha & \lambda_k\end{bmatrix} \oplus \bigoplus_{k=1}^{s} \underbrace{\left[\begin{array}{cc|cc} 0 & \overline{\nu}_k & \alpha & 0 \\ \nu_k & 0 & 0 & \alpha \\ \hline \alpha & 0 & 0 & \nu_k \\ 0 & \alpha & \overline{\nu}_k & 0 \end{array}\right]}_{=:H_k(\alpha)}. \end{aligned}$$

Since a congruence transformation does not change the inertia of a hermitian matrix, it follows that

$$i_{\geq}\left(\begin{bmatrix} H & \alpha\,\overline{S} \\ \alpha\, S & \overline{H}\end{bmatrix}\right) = \underbrace{i_{\geq}\left(\bigoplus_{k=1}^{r}\begin{bmatrix} \lambda_k & \alpha \\ \alpha & \lambda_k\end{bmatrix}\right)}_{=:g(\alpha)} + \sum_{k=1}^{t} i_{\geq}\left(H_k(\alpha)\right).$$

We are now going to evaluate this sum.

1. A matrix of the form $\begin{bmatrix} \lambda & \alpha \\ \alpha & \lambda\end{bmatrix}$ has the eigenvalues $\lambda\pm\alpha$ since $\begin{bmatrix} \lambda & \alpha \\ \alpha & \lambda\end{bmatrix}\begin{bmatrix} 1 \\ \pm 1\end{bmatrix} = (\lambda\pm\alpha)\begin{bmatrix} 1 \\ \pm 1\end{bmatrix}$. Thus for all $\alpha\in\mathbb{R}$,

   $$g(\alpha) = \#\{\, k\in\underline{r} \mid \lambda_k - \alpha \geq 0 \,\} + \#\{\, k\in\underline{r} \mid \lambda_k + \alpha \geq 0 \,\}.$$

   Now we have to distinguish three cases: For $k\in\{\rho+1,\ldots,\rho+\ell\}$ we have $\lambda_k \geq 1$ and therefore $\lambda_k \pm \alpha \geq 0$ for all $\alpha\in[0,1]$. For $k \geq \rho+\ell+1$ we have $\lambda_k < -1$ and hence $\lambda_k\pm\alpha < 0$ for $\alpha\in[0,1]$. If $k\in\underline{\rho}$ then $\lambda_k\in[-1,1)$ and one of the eigenvalues $\lambda_k\pm\alpha$ changes its sign in the interval $[0,1]$. Thus we have for $\alpha\in[0,1]$, $g(\alpha) = 2\ell + f(\alpha)$, where

   $$f(\alpha) = \#\left\{\, k\in\underline{\rho} \mid \lambda_k - \alpha \geq 0 \,\right\} + \#\left\{\, k\in\underline{\rho} \mid \lambda_k + \alpha \geq 0 \,\right\}.$$

2. In order to determine $i_{\geq}(H_k(\alpha))$ we argue as follows. Let

   $$\mathcal{X}_+ := \left\{ \begin{bmatrix} z_1 \\ iz_1 \\ z_2 \\ -iz_2 \end{bmatrix} \,\middle|\, z_1, z_2\in\mathbb{C} \right\}, \qquad \mathcal{X}_- := \left\{ \begin{bmatrix} z_1 \\ -iz_1 \\ z_2 \\ iz_2 \end{bmatrix} \,\middle|\, z_1, z_2\in\mathbb{C} \right\}.$$

   Then $\dim\mathcal{X}_+ = \dim\mathcal{X}_- = 2$ and a straightforward calculation (using the fact that $\Im\nu_k > 0$) shows that $x^*H_k(\alpha)x > 0$ for every $x\in\mathcal{X}_+\setminus\{0\}$ and $x^*H_k(\alpha)x < 0$ for every $x\in\mathcal{X}_-\setminus\{0\}$, $\alpha\in\mathbb{R}$. We conclude that $H_k(\alpha)$ has two positive and two negative eigenvalues. Thus $i_{\geq}(H_k(\alpha)) = 2$ for all $\alpha\in\mathbb{R}$.

So far we have shown that

$$i_{\geq}\left(\begin{bmatrix} H & \alpha\,\overline{S} \\ \alpha\, S & \overline{H}\end{bmatrix}\right) = f(\alpha) + 2\ell + 2s \qquad \text{for all } \alpha\in[0,1],$$

Thus (A.11) is proved if we can find an $\alpha_0\in[0,1]$ such that $\lfloor (1/2) f(\alpha_0)\rfloor \leq m$, where

$$m = \max\{\; q\in\mathbb{N}_0 \mid 2q \leq \rho \text{ and } \lambda_k + \lambda_{2q+1-j} \geq 0 \text{ for all } j\in\underline{q} \;\}.$$

However, by definition of m one of the following conditions holds.

(*a*) $\rho \geq 2(m+1)$ and there exists a $j_0 \leq m+1$ such that $\lambda_{j_0} + \lambda_{2(m+1)+1-j_0} < 0$;

(*b*) $\rho = 2m$;

(*c*) $\rho = 2m+1$.

In Case (*a*) we can take $\alpha_0 := \frac{1}{2}(\lambda_{j_0} - \lambda_{2(m+1)+1-j_0}) \in [0,1]$ since then

$$\lambda_{j_0} - \alpha_0 = \lambda_{2(m+1)+1-j_0} + \alpha_0 = \frac{1}{2}(\lambda_{j_0} + \lambda_{2(m+1)+1-j_0}) < 0,$$

and consequently the eigenvalues

$$\begin{aligned} &\lambda_k - \alpha_0, \quad \text{where} \quad j_0 \leq k \leq r, \\ &\lambda_k + \alpha_0, \quad \text{where} \quad 2(m+1)+1-j_0 \leq k \leq r \end{aligned}$$

are negative. Thus

$$\begin{aligned} f(\alpha_0) &= \#\{\ k \in \underline{\rho} \mid \lambda_k - \alpha \geq 0\ \} + \#\{\ k \in \underline{\rho} \mid \lambda_k + \alpha \geq 0\ \} \\ &\leq (j_0 - 1) + (2(m+1) - j_0) = 2m+1 \end{aligned}$$

and hence $\lfloor (1/2) f(\alpha_0) \rfloor \leq m$. In the cases (*b*) and (*c*) we can take $\alpha_0 = 1$ since $\lambda_k - 1 < 0$ for all $k \in \underline{\rho}$ and therefore $f(1) \leq \rho \leq 2m+1$. $\square$

Bibliography

[1] Balas, G.J.; Doyle, J.C.; Glover, K; Packard, A.; Smith, R.: *μ-analysis and synthesis Toolbox (For use with MATLAB).* The MathWorks, Inc, 1998.

[2] Bauer, F.L.; Fike, C.T.: Norms and exclusion theorems. *Numer. Math.* 2:137-141, 1960.

[3] Baumgärtel, H.: *Analytic perturbation theory for matrices and operators.* Birkhäuser, 1985.

[4] Benedetti, R.; Risler, J.: *Real algebraic and semi-algebraic sets.* Actualiés Mathématiques. Hermann Éditeurs des Sciences et des Arts, Paris, 1990.

[5] Bekas, C.; Gallopoulos, E.: Parallel computation of pseudospectra by fast descent. *Parallel Comput.* 28(2): 223-242, 2002.

[6] Bernhardsson, B.: Simultaneous block-diagonalization of one hermitian and one symmetric form. *Internal Report TFRT-7520, Department of Automatic Control, Lund Institute of Technology*, 1994

[7] Bernhardsson, B.; Qiu, L.; Rantzer, A.: On real perturbation values, *Preprint, 6th ILAS conference.* Chemnitz, 1996.

[8] Bernhardsson, B.; Qiu, L.; Rantzer, A.: *A summary on the real perturbation values*, in: Hurwitz Centenial on Stability Theory. Birkhäuser, 1996.

[9] Bernhardsson, B.; Qiu, L.; Rantzer, A.: Real perturbation values and real quadratic forms in a complex vector space. *Linear Algebra Appl.* 270:131-154, 1998.

[10] Bhatia, R.: *Matrix Analysis.* Graduate Texts in Mathematics 169. Springer, 1996.

[11] Bochnak, J.; Coste, M.; Roy, M.-F.: *Real Algebraic Geometry.* Springer, 1998.

[12] Böttcher, A.; Silbermann, B.: *Introduction to large truncated Toeplitz matrices.* Universitext. Springer, 1999

[13] Böttcher, A.; Embree, M., Sokolov, V.I.: The Spectra of Large Toeplitz Band Matrices with a Randomly Perturbed Entry. *OUCL Numerical Analysis Group Technical Report 01/16*, 2001.

[14] Böttcher, A.; Grudsky. S.M.: Can spectral value sets of Toeplitz band matrices jump? *Linear Algebra Appl.* 351:99-116, 2002.

[15] Byers, R.. A bisection method for measuring the distance of a stable matrix to the unstable matrices. *SIAM J. Sci. Stat. Comput.* 9(5):875-881, 1988.

[16] Brauer, A.: Limits for the characteristic roots of a matrix. *II. Duke Math. J.* 14:21-26, 1947.

[17] Brieskorn, E.: *Lineare Algebra und Analytische Geometrie, Band 2.* Vieweg-Verlag, 1993.

[18] Brualdi, R.A.: Matrices, eigenvalues, and directed graphs. *Linear Multilinear Algebra* 11:143-165, 1982.

[19] Chaitin-Chatelin, F.; Harrabi, A.; Ilahi, A.: About Hölder condition numbers and the stratification diagram for defective eigenvalues. *Math. Comput. Simul.* 54(4):397-402, 2000.

[20] Chatelin, F.: *Eigenvalues of matrices.* Chichester: John Wiley & Sons, 1993.

[21] Chaitin-Chatelin, F.; Frayssé, V.: *Lectures on finite precision computations.* SIAM, Society for Industrial and Applied Mathematics, 1996.

[22] Chaitin-Chatelin, F.: Is nonnormality a serious computational difficulty in practice? *CERFACS Technical report TR/PA/96/33*, 1996.

[23] Dedieu, J.; Tisseur, F.: Perturbation Theory for Homogeneous Polynomial Eigenvalue Problems. *Linear Algebra Appl.* 358:71-74, 2003.

[24] Elsner, L.; Paardekooper, M.H.C.: On measures of nonnormality of matrices. *Linear Algebra Appl.* 92:107-124, 1987.

[25] Embree, M.; Trefethen, L.N.: Generalizing eigenvalue theorems to pseudospectra theorems. *SIAM J. Sci. Comp.* 23: 583-590, 2001.

[26] Embree, M., Trefethen, L.N.: The Pseudospectra gateway. Web site: http://www.comlab.ox.ac.uk/pseudospectra.

[27] Ermolaev, J. B.: The simultaneous reduction of symmetric and hermitian forms, *Izv. Vyss. Zaved. Mat.* 21:10-23, 1961.

[28] Fischer, G.: *Plane algebraic curves.* Student Mathematical Library. 15. Providence, RI: American Mathematical Society . 2001.

[29] Fitzgerald, C.H.; Horn, R. A.: On the structure of Hermitian-symmetric inequalities. *J. London Math. Soc.* II. Ser. 15:419-430, 1977.

[30] Gallestey, E.: *Theory and Numerics of Spectral Value sets.* Ph. D. thesis, University of Bremen, 1998.

[31] Gallestey, E.: Computing spectral value sets using the subharmonicity of the norm of rational matrices. *BIT* 38(1):22-33, 1998.

[32] Gantmacher, F.R.: *Matrix theory.* Chelsea Publishing Company, New York, 1977.

[33] Genin, Y.; Stefan, R.; Van Dooren, P.: Real and complex stability radii of polynomial matrices. *Linear Algebra Appl.* 352:381-410, 2002.

[34] Golub, G.H.; Van Loan, C.F.: *Matrix computations.* The Johns Hopkins University Press. XVI, 1983.

[35] Haantjes, J.: Klassifikation der antilinearen Transformationen. *Math. Annalen* 1936:98-106.

[36] Harrabi, A.: Défaut de Normalité, *CERFACS Technical Report TR/PA/98/36*, 1998.

[37] Harrabi, A.: *Pseudospectres d'Operateurs Intégraux et Différentiels: Application a la Physique Mathematique.* Thesis. Universite des Sciences Sociales de Toulouse. May 1998.

[38] Henrici, P.: Bounds for iterates, inverses, spectral variation and fields of values of non-normal matrices. *Numer. Math.* 4:24-40, 1962.

[39] Higham, N.J.: Estimating the matrix p-norm. *Numer. Math.* 62(4):539-556, 1992.

[40] Higham, N.J.; Tisseur, F.: More on pseudospectra for polynomial eigenvalue problems and applications in control theory. *Linear Algebra Appl.* 352:435-453, 2002.

[41] Higham, N.J.; Tisseur, F.: Structured pseudospectra for polynomial eigenvalue problems, with applications. *SIAM J. Matrix Anal. Appl.* 23(1):187-208, 2001.

[42] Hinrichsen, D.; Kelb, B.: Spectral value sets: A graphical tool for robustness analysis, *Syst. Control Lett.* 21:127-136, 1993.

[43] Hinrichsen, D.; Kelb, B.: Stability radii and spectral value sets for real matrix perturbations. *Systems and Networks: Mathematical Theory and Applications, Vol.II, Invited and Contributed Papers. Akademie-Verlag, Berlin*, 1994.

[44] Hinrichsen, D.; Pappas, G.: Robust stability of linear systems described by higher order dynamic equations. *IEEE Trans. Autom. Control* 38(9):1430-1435, 1993.

[45] Hinrichsen, D., Plischke, E., Wirth, F.: State Feedback Stabilization with Guaranteed Transient Bounds. *Proc. Symposium on Mathematical Theory of Networks and Systems MTNS-2002.* South Bend, IN, US, 2002, (CD-ROM), paper 2132.

[46] Hinrichsen, D; Plischke, E.; Pritchard, A.J.: Liapunov and Riccati equations for practical stability, in *Proc. European Control Conf. ECC-2001*, Porto, Portugal, (CD-ROM), pp. 2883–2888, 2001.

[47] Hinrichsen, D.; Pritchard, A.J.: On the transient behaviour of stable linear systems, in *Proceedings of MTNS-2000*, (CD-ROM), Perpignan, France, 2000.

[48] Hinrichsen, D.; Pritchard, A.J.: Robustness measures for linear systems with application to stability radii of Hurwitz and Schur polynomials. *Int. J. Control* 55(4): 809-844, 1992.

[49] Hinrichsen, D.; Pritchard, A.J.: Real and complex stability radii: a survey, in: D. Hinrichsen and B. Mårtensson (eds.) *Control of Uncertain Systems, Proc. International Workshop Bremen 1989*, Birkhäuser 1990, 119-162.

[50] Hinrichsen, D.; Pritchard, A.J.: A note on some differences between real and complex stability radii, *Systems and Control Letters* 14:401-409, 1990.

[51] Hinrichsen, D.; Pritchard, A.J.: On spectral variations under real matrix perturbations. *Numer. Math.* 60:509-524, 1992.

[52] Hinrichsen, D.; Pritchard, A.J.: Destabilization by output feedback. *Differential and Integral Equations* 5:357-386, 1992.

[53] Hinrichsen, D.; Pritchard, A.J.: *Dynamical Systems Theory.* Manuscript in preparation.

[54] Hinrichsen, D.; Pritchard, A.J.: Stability radii of linear systems. *Syst. Control Lett.* 7:1-10, 1986.

[55] Hinrichsen, D.; Son, Nguyen, K.: μ-analysis and robust stability of positive linear systems. *Appl. Math. Comput. Sci.* 8(2): 253-268, 1998.

[56] Hörmander, L.: *The analysis of linear partial differential operators, vol. 2.* Springer, 1983.

[57] Y. Hong: *Consimilarity: Theory and Applications,* Doctoral Dissertation, Johns Hopkins Univ., 1984.

[58] Y. Hong, R. A. Horn: A Canonical Form for Matrices under Consimilarity. *Linear Algebra and its Applications* 102:143-168, 1988.

[59] Y. Hong: A Hermitian Canonical Form for Complex Matrices under Consimilarity. *Linear Algebra and its Applications,* 113:1-19, 1990.

[60] Y. Hong: A canonical form for Hermitian matrices under complex orthogonal congruence. *SIAM J. Matrix Anal. Appl.* 10(2):233-243, 1989.

[61] Horn, R.A.; Johnson, C.R.: *Matrix analysis.* Cambridge University Press, 1985.

[62] Horn, R.A.; Johnson, C.R.: *Topics in matrix analysis.* Cambridge University Press, 1991.

[63] Hua, L.: Orthogonal classification of Hermitian matrices. *Trans. Am. Math. Soc.* 59:508-523, 1946.

[64] Karow, M.: Spectral value sets for real matrix perturbations. *Proc. ECC99,* Karlsruhe, 1999.

[65] Karow, M.: On real spectral value sets of real normal matrices. *ZAMM, Z. Angew. Math. Mech* 79, Suppl. 3, S909-S910, 1999.

[66] Karow, M.: *Hermitesch-symmetrische Ungleichungen und der reelle Stabilitätsradius.* Diploma thesis. Universität Bremen (1997)

[67] Kato, T.: *Perturbation theory for linear operators.* Springer, 1980.

[68] Lang, S.: *Algebra. 3rd revised ed.* Springer, 2002.

[69] Lawrence, C. L.T.; Tits, A. L.; Van Dooren, P.: A fast algorithm for the computation of an upper bound on the μ-norm. *Automatica* 36(3):449-456, 2000.

[70] A. B. Li Santi, R. C. Thompson: Simultaneous Reduction of a Complex Skew Matrix and a Hermitian Matrix. *Linear Algbra and its Appl.* 147:275-322, 1991.

[71] Lewkowicz, I.: When are the complex and the real stability radii equal? *IEEE Trans. Autom. Control* 37(6):880-883, 1992.

[72] Lojasiewicz, S.: Sur les ensembles semi-algébriques. *Sympos. Math., Roma 3, Probl. Evolut. Sist. solare, Nov. 1968 e Geometria.* 1969:233-239, 1970.

[73] Moro, J.; Burke, J.V.; Overton, M. L.: On the Lidskii-Vishik-Lyusternik perturbation theory for eigenvalues of matrices with arbitrary Jordan structure. *SIAM J. Matrix Anal. Appl.* 18(4):793-817, 1997.

[74] Opt'n Hof, F.: *Stabilitätsradien von Matrizen unter gebrochen linearen Störungen.* Diploma thesis. Universität Bremen, November 1997.

[75] Packard, A.; Doyle, J.: The complex structured singular value. *Automatica* 29(1):71-109, 1993.

[76] Paice, A.D.B., Wirth, F.R.: Analysis of the Local Robustness of Stability for Flows. *Mathematics of Control, Signals and Systems*, 11(4):289-302, 1998.

[77] Pritchard, A.J.: Transitory behaviour of uncertain systems, in F. Colonius et al., eds., *Advances in Mathematical Systems Theory*, pages 1–18, Boston, 2000. Birkhäuser.

[78] Qiu, L.; Bernhardsson, B.; Rantzer, A.; Davison, E.J.; Young, J.C.; Doyle, J.: A Formula for Computation of the Real Stability Radius. *Automatica* 31(6):879-890, 1995.

[79] Rohn, J.: Computing the norm $\|A\|_{\infty,1}$ is NP-hard. *Linear Multilinear Algebra* 47(3):195-204, 2000.

[80] Ruhe, A.: On the closeness of eigenvalues and singular values for almost normal matrices. *Linear Algebra Appl.* 11:87-94, 1975.

[81] Seidenberg, A.: A new decision method for elementary algebra. *Ann. of Math.* 60:365-374, 1954.

[82] Shafarevich, I.R.: *Basic Algebraic Geometry , vol. 1.* Berlin: Springer-Verlag, 1994

[83] Shapiro, H.: A survey of canonical forms and invariants for unitary similarity. *Linear Algebra Appl.* 147:101-167, 1991.

[84] Safonov, M.G.(ed.); Fan, K.-H.M.(ed.): Multivariable stability margin. *Int. J. Robust Nonlinear Control* 7(2):97-226, 1997.

[85] Sontag, E.D.: *Mathematical control theory. Deterministic finite dimensional systems.* Springer-Verlag, 1990.

[86] Stewart, G.W., Sun, J.: *Matrix Perturbation Theory.* Academic Press, San Diego, 1990.

[87] Tarski, A.:*A decision method for elementary algebra and geometry.* Prepared for publication by J.C.C. Mac Kinsey, Berkeley, 1951

[88] Thompson, R.C.: Pencils of Complex and Real Symmetric and Skew Matrices. *Linear Algebra and its Applications* 147:323-371, 1991.

[89] Toh, K.; Trefethen, L.N.: Pseudozeros of polynomials and pseudospectra of companion matrices. *Numerische Mathematik* 68:403-425 1994.

[90] Toumazou, V.: *Portaits Spectraux de Matrices: Un Outil d'Analyse de la Stabilité*, Ph. D. thesis, Université Henri Poincaré, Nancy-1, France, 1996.

[91] Trefethen, L.N.: Computation of pseudospectra. *Acta Numerica*, 8:247-295, 1999.

[92] Trefethen, L.N.: Pseudospectra of linear operators. *SIAM Rev.* 39(3):383-406, 1997.

[93] Trefethen, L.N.: Pseudospectra of matrices *In D. F. Griffiths and G. A. Watson, eds., Numerical Analysis 1991 (Dundee, 1991)*, 234-266. Longman Sci. Tech, Harlow, Essex, 1992

[94] Van Dooren, P.; Sreedhar, J., Tits, L.A.: A Fast Algorithm to Compute the Stuctured Hurwitz-Stability Radius. *34th CDC-conference*, 1995.

[95] Van Loan, C.: How near is a stable matrix to an unstable matrix?. *Contemporary Mathematics* 47:465-478, 1985

[96] Varga, R. S.: Gersgorin-type eigenvalue inclusion theorems and their sharpness. *ETNA, Electron. Trans. Numer. Anal.* 12:113-133, 2001.

[97] Varga, R. S.; Krautstengel, A.: On Gershgorin-type problems and ovals of cassini. *ETNA, Electron. Trans. Numer. Anal.* 8, 15-20, 1999.

[98] Zhou, K.; Doyle, J. C.; Glover, K.: *Robust and optimal control.* Upper Saddle River, NJ: Prentice Hall, 1996.