UNIVERSITÉ DE TOULOUSE

THÈSE

présentée en vue de l'obtention du

Diplôme de Doctorat

Spécialité: Mathématiques Appliquées

par

Morad Ahmadnasab

Université Toulouse 1 et CERFACS

Homotopic Deviation theory: A qualitative study

Théorie de la Déviation Homotopique: Etude qualitative

Défendue publiquement le 24 octobre 2007 à Toulouse devant le jury composé de:

Président:	Prof. Jacqueline Fleckinger	Université de Toulouse 1
Rapporteurs:	Prof. Martin B. van Gijzen	T. U. Delft, Netherlands
	D.R. Stéphane Gaubert	INRIA Rocquencourt
Directeur de thèse:	Prof. Françoise Chaitin-Chatelin	Université de Toulouse 1
Examinateurs:	Prof. Iain S. Duff	RAL, UK et CERFACS
	Prof. Jacques Giacomoni	Université de Pau
	Dr. Elisabeth Traviesas-Cassan	SOGETI High Tech.
	Prof. Jean-Claude Yakoubsohn	Université de Toulouse 3

Abstract

Given the square complex matrices A and E, we consider their coupling by the complex parameter t into $A(t) = A + tE$. The deviation matrix E is singular. The complex parameter t varies in the completed complex plane and its modulus can be unbounded.
This work on Homotopic Deviation theory has two components.
1. The purely algebraic aspect of the theory introduces new kinds of singularities such as frontier and critical points.
2. Computer experiments are used to perform a qualitative analysis of Homotopic Deviation in finite precision. For this aim, several graphical tools are developed.
As an application of this theory, the dependence of the structure of the regular pencil is analyzed by means of the notion of frontier points.
This work also performs a homotopic backward analysis and contrasts it with the classical normwise backward analysis. The thesis ends by an application of this theory to Arnoldi's method.

Résumé

Soient deux matrices carrées complexes A et E données. Nous considérons leur couplage par le paramètre complexe t sous la forme $A(t) = A + tE$. La matrice de déviation E est singulière. Le paramètre complexe t varie dans le plan completé par le point à l'infini.
Ce travail sur la Déviation Homotopique a deux composantes.
1. L'aspect purement algébrique de la théorie introduit de nouveaux types de singularités tels que les points-frontière et les points critiques.
2. Des expériences sur ordinateur sont utilisées pour exécuter une analyse qualitative en précision finie. Dans ce but, plusieurs outils graphiques sont développés.
Une application de la théorie est faite pour caractériser la structure d'un faisceau régulier à l'aide des points-frontière.
Ce travail réalise aussi une analyse inverse homotopique et a compare avec l'analyse inverse classique (de type normwise). La thèse se termine par une application de la théorie à la méthode d'Arnoldi.

Keywords: Resolvent matrix, singular value decomposition, Jordan form, frontier point, critical point, limit point, Lidskii's theory, regular matrix pencil, normwise backward analysis, homotopic backward analysis, invariant eigenvalue, evolving eigenvalue, spectral portrait, frontier portrait, Krylov subspace methods, Arnoldi method.

Acknowledgments

A journey is easier when you travel together. Interdependence is certainly more valuable than independence. This thesis is the result of four years of work whereby I have been accompanied and supported by many people. It is a pleasant aspect that I have now the opportunity to express my gratitude to all of them.

The first person I would like to thank is my supervisor Madame Françoise Chaitin-Chatelin, Professor of Mathematics at the Université Toulouse 1. I have been in her group, the Qualitative Computing Group in the Parallel Algorithms Team, CERFACS since August 2003. During these years I have known Madame Françoise Chaitin-Chatelin as a sympathetic and principle-centered person. Her high enthusiasm and integral view of research and her dedication for providing "only high-quality work and not less", has made a deep impression on me. I owe her deep gratitude for having taught to me this way of doing research. I am really glad that I have come to get to know Professor Françoise Chaitin-Chatelin in my life.

I am extremely grateful to Professor Martin B. van Gijzen, T. U. Delft, Netherlands and Professor Stéphane Gaubert, Research Director of INRIA Rocquencourt and Professor at the Ecole Polytechnique, France who accepted to act as referees for my thesis. It was an honour for me to benefit from their feedback and suggestions on my research work.

I wish to express my sincere gratitude to Professor Iain S. Duff, RAL, UK and leader of the Parallel Algo Team at CERFACS for his continued support and his professional advices during my study and for having me honoured by his presence in the jury.

I wish to thank Madame Jacqueline Fleckinger, Professor and Director of the CEREMATH group at the Université Toulouse 1, to honour me by accepting to chair the thesis Committee.

My sincere thanks go to Professor Jean-Claude Yakoubsohn, Université of Toulouse 3, who accepted to take part in the jury. I appreciate his perceptive questions and suggestions about the structure of singular pencils and also about the possible application of HD to PDE.

I gladly acknowledge my debt to Professor Jacques Giacomoni, Université of Pau, and to Dr. Elisabeth Traviesas-Cassan, from SOGETI High Tech. who accepted to take part in the jury.

Grateful acknowledgments are made to Professor Fermin S. V. Bazán, Universidade Federal de Santa Catarina, Florianópolis, Brazil who taught me many things about Linear Algebra and directed my attention to the rich algebraic structure of the communication matrix M_z.

I would like to thank sincerely Professor Luc Giraud, INPT and Dr. Serge Gratton, CNES, Senior Researcher at CERFACS for their help and support during these years.

I wish to thank Dr. Xavier Vasseur, Senior Researcher at CERFACS for his professional help.

I would like to thank all the members of the Parallel Algorithms Team, Brigitte Yzel, Nicole Boutet, Marc Baboulin, Bora Ucar, Azzam Haidar, Mélodie Mouffe, Milagros Garcia, Jean Tshimanga Ilunga, Xavier Pinel, Tzvetomila Slavova, Anke Troeltzsch, and also some earlier visitors, Post-Doctoral fellows and PhD students, Abderrazak Ilahi, Bruno Carpentieri, Songklod Riyavong, Emeric Martin, Nasreddine Megrez and Fabian Bastin for helping me and sharing my enjoyment for this research.

My sincere gratitude goes to the CSG group, administrative staff and the direction of CERFACS for their professional and friendly support.

I would like to thank the direction and administrative members of the Ecole Doctorale MPSE, Université Toulouse 1, especially Madame Aude Schloesing for their help during these years.

My special gratitude is due to my parents and parents-in-law, my brothers, my sisters, their families, my wife's brothers and our relatives for their loving support. I owe my loving thanks to my wife Farnoosh and to my son Saeid. Without their encouragement and understanding it would have been impossible for me to finish this work.

I gratefully acknowledge the University of Kurdistan in Sanandaj and the Ministry of Science, Research and Technology of Iran for providing my PhD scholarship.

Morad Ahmadnasab,

October 24, 2007
Toulouse, France

Contents

General presentation

Homotopic Deviation theory, HD, is concerned with the properties of a linear coupling by a *complex* parameter t of two square matrices A and E of order n, yielding $A(t) = A + tE$. $A = A(0)$ is the original matrix and E is the deviation matrix, with $1 \leq \operatorname{rank} E = r \leq n$. Typically $r \ll n$, but $|t|$ can be unbounded: $t \in \hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$.

Of special interest are the properties of the resolvent $R(t,z) = (A(t) - zI)^{-1}$ when it exists, as well as the spectrum $t \mapsto \sigma(A(t))$, for $(t,z) \in \hat{\mathbb{C}} \times \mathbb{C}$.

The study starts with the formal factorization

$$A + tE - zI = (I + tE(A - zI)^{-1})(A - zI).$$

The singular value decomposition of E allows us to write $E = UV^H$, with $U, V \in \mathbb{C}^{n \times r}$, so that

$$-E(A - zI)^{-1}U = U(V^H(zI - A)^{-1}U) = UM_z,$$

for $z \in re(A) = \mathbb{C} \backslash \sigma(A)$. The algebraic properties of the rational matrix map $z \mapsto M_z \in \mathbb{C}^{r \times r}$, for $z \in re(A)$ are key concepts in the theory. For example, $z \in \mathbb{C}$ is an eigenvalue of $A(t)$ iff $t\mu_z = 1$, where μ_z is an eigenvalue of M_z and $t \in \hat{\mathbb{C}}$.

It is common wisdom in perturbation theory that, for $z \in re(A)$, the resolvent $R(t,z)$ is analytic for $|t| < \frac{1}{\rho(M_z)}$. But one has *more* at any z in $re(A)$ such that M_z is invertible: $\lim_{|t| \to \infty} R(t,z) = R(\infty, z)$ ($\neq 0$ for $r < n$) exists, and $R(t,z)$ is analytic in t around ∞ (that is analytic in $s = 1/t$ around 0) for $|t| > \rho(M_z^{-1})$. Points z in $re(A)$ such that M_z is singular are called *frontier points*, which form the frontier set $F(A,E) = \{z \in re(A),\ \operatorname{rank} M_z < r\}$. At a frontier point z, there is only *one* possibility of analyticity for $R(t,z)$: analyticity around 0. At a regular point $z \in re(A) \backslash F(A,E)$, the two possibilities *coexist*: analyticity around 0 and ∞.

When the frontier set is discrete, it is easily characterized from the data A, U, V as the (at most $n - r$) roots of $\hat{\pi}(z) = \det \hat{A}(z)$ not in $\sigma(A)$, where $\hat{A}(z)$ is the augmented matrix

$$\hat{A}(z) = \begin{bmatrix} zI - A & -U \\ V^H & 0 \end{bmatrix}$$

of order $n + r$. The simplicity of this characterization is remarkable.
When they exist, the points z in $F(A,E)$ such that $\rho(M_z) = 0$ are called *critical* points. At such points z, $R(t,z)$ has a *finite* representation: it is a polynomial in t of degree $\leq r$.

The companion study of $t \mapsto \sigma(A(t))$, and in particular of $\lim_{|t|\to\infty} \sigma(A(t))$ reveals also very interesting properties when $r < n$. We define $\lim_{|t|\to\infty} \sigma(A(t)) = \{\infty, \mathrm{Lim}\}$: the set Lim consists of the limit points $\lim_{|t|\to} \lambda_i(t)$, of the eigenvalues $\lambda_i(t) \in \sigma(A(t))$ which are in $\mathbb{C}$, at finite distance. When $r = n$, then $\mathrm{Lim} = \emptyset$: all eigenvalues escape to ∞. But when $1 \leq r < n$, Lim can be nonempty. When $0 \in \sigma(E)$ is semi-simple, Lim consists of $n-r$ points which are simply characterized as the eigenvalues of $\Pi = PAP_{\restriction \mathrm{Ker} E}$, P being the eigenprojection for E on the eigenspace $\mathrm{Ker}\, E$.
When $0 \in \sigma(E)$ is defective, the theory is more difficult. The arithmetic *mean* of m eigenvalues of $A(t)$ converges to the point $\frac{1}{m} \mathrm{tr}\,(PAP_{\restriction M})$ which is at finite distance, where m is the algebraic multiplicity, and P becomes the *spectral* projection on the invariant subspace $M = \mathrm{Ker}\, E^m$ associated with $0 \in \sigma(E)$. The analysis of *individual* eigenvalues requires to extend the Lidskii-Puiseux perturbation theory. A complete characterization of Lim, the set of individual limits, is not always known. Therefore, the theoretical analysis presented in Part I remains incomplete.

To compensate for the possible lack of theoretical prediction about Lim, visualization tools based on the graphical representation of the spectral field $t \mapsto \sigma(A(t))$, as well as of the maps $z \mapsto \rho(M_z)$, $\rho(M_z^{-1})$, are developed in Part II. These tools allow us to perform a qualitative analysis of Homotopic Deviation. In particular, the homotopic backward analysis is contrasted with the normwise backward analysis classically used in Numerical Analysis to assess the validity of computer simulations. In this latter approach, the structure of the modification ΔA of A is arbitrary with a constraint on $\|\Delta A\|$. Such a broad context does not allow us to go further about $A + \Delta A - zI$ than the simple discrimination between existence/nonexistence of the inverse, which leads to the concept of normwise spectral portrait: $z \mapsto 1/\|(A - zI)^{-1}\|$. As useful as it is, this portrait does not measure up to the spectral information provided by the homotopic portraits associated with M_z.
The homotopic notions of analyticity in t at ∞, of frontier and critical points that can be associated with $R(t, z)$ have no counterpart in normwise analysis. This illustrates dramatically the computational wealth of HD.

The last chapter of the thesis is devoted to an attempt to use HD to further our understanding of the success of the basic incomplete Arnoldi algorithm to compute eigenvalues in finite precision.

To conclude we remark that, as computer simulation becomes ubiquitous in science and engineering, there is an increasing concern about issues such as reliability, safety and robustness for non linear structural dynamics. It becomes crucial to understand the effects that uncertainties may have on non linear models for natural and man-made systems which are used in Engineering, as well as in physical and biological Sciences. We view the theory of HD which is to be presented (for the first time in

a single written document) as a necessary and very preliminary first step towards a satisfactory understanding of the issues at stake. This work presents the mathematical foundations. Much, much more research is needed to transform basic homotopic tools into useful validation concepts for nonlinear model theories and simulations.

Part I

HD in exact arithmetic: the theory

Chapter 1

Background in Matrix Algebra

1.1 Introduction

Let $\mathbb{C}^n$ represent the space of column vectors x with complex components ζ_j, $j = 1, \cdots, n$. Then x^H is the row vector with components $\bar{\zeta_j}$. The *Euclidean scalar product* on $\mathbb{C}^n$ is given by $< x, y >= y^H x$. The vectors x and y are said to be **orthogonal** when $< x, y >= 0$. Let $\{x_j : j = 1, \cdots, n\}$ be a basis of $\mathbb{C}^n$, that is a set of n linearly independent vectors. The basis is **orthonormal** if and only if $< x_i, x_j >= \delta_{ij}$ for $i, j = 1, \cdots, n$ where δ_{ij} is the *Kronecker symbol*, equal to 1 if $i = j$ and 0 if $i \neq j$.
Let $\{a_j : j = 1, \cdots, r\}$ be a set of r vectors of $\mathbb{C}^n$. The rectangular matrix of size $n \times r$ whose columns are the vectors $a_1, \cdots, a_r$ is denoted by $A = [a_1, \cdots, a_r] \in \mathbb{C}^{n \times r}$.

If $a = (a_{ij})$ for $i = 1, \cdots, n$ and $j = 1, \cdots, r$, then $A = (a_{ij}) \in \mathbb{C}^{n \times r}$. The *range* of the matrix $A \in \mathbb{C}^{n \times r}$ is the subspace

$$\text{Im } A = \{Ax \ \mid \ x \in \mathbb{C}^r\} \subset \mathbb{C}^n,$$

generated by the r column vectors of A. The *rank* of a matrix is equal to the dimension of the range of A, that is $\dim(\text{Im } A) = \text{rank } A$. The *kernel* (or *null* space) of A is defined by

$$\text{Ker } A = \{x \in \mathbb{C}^r \mid Ax = 0\} \subset \mathbb{C}^r.$$

A matrix belonging to $\mathbb{C}^{n \times n}$ is said to be **square matrix**. The **transpose** of a matrix $A \in \mathbb{C}^{n \times r}$ is a matrix $B \in \mathbb{C}^{r \times n}$ whose elements are defined by $b_{ij} = a_{ji}$ for $i = 1, \cdots, r$ and $j = 1, \cdots, n$. The transpose of a matrix A is denoted by A^T. The **transpose conjugate** matrix of matrix A is denoted by A^H and is defined by $A^H = \overline{A}^T = \overline{A^T}$, where the bar denotes the (element-wise) complex conjugation. If $A = A^H$, then A is **Hermitian**. By definition, a hermitian matrix must be square. If a real matrix $A \in \mathbb{R}^{n \times n}$ is hermitian, then it is said to be **symmetric**.

The $n \times n$ *identity matrix* I_n is defined by the column partitioning $I_n = I = \begin{bmatrix} e_1 & \cdots & e_n \end{bmatrix}$ where the $n \times 1$ column vector e_k is the k th *canonical* basis vector whose unique nonzero entry is its k th entry which is equal to 1. The identity matrix satisfies the equality $AI = IA = A$ for every square matrix A of order n. The *inverse* of a square matrix, when it exists, is a matrix C such that $CA = AC = I$. The inverse of A is denoted by A^{-1}. If A^{-1} exists, then A said to be *nonsingular* or *invertible*, or *regular*. Otherwise, it is called *singular*.
Let A and B be matrices of compatible dimensions, then

$$(AB)^H = B^H A^H. \tag{1.1.1}$$

Similarly, the product of two invertible square matrices A and B, satisfies

$$(AB)^{-1} = B^{-1} A^{-1}. \tag{1.1.2}$$

The **Sherman-Morrison** formula [43] states that

$$(A + uv^H)^{-1} = A^{-1} - \frac{(A^{-1}u)(v^H A^{-1})}{1 + v^H A^{-1} u}, \tag{1.1.3}$$

provided that A is invertible and $v^H A^{-1} u =< A^{-1}u, v > \neq -1$.

The **Sherman-Morrison-Woodbury** formula [38] gives the expression for the inverse of $(A + UV^H)$ where $A \in \mathbb{C}^{n \times n}$ and $U, V \in \mathbb{C}^{n \times r}$:

$$(A + UV^H)^{-1} = A^{-1} - A^{-1}U(I_r + V^H A^{-1} U)^{-1} V^H A^{-1}, \tag{1.1.4}$$

provided that A of order n and $(I_r + V^H A^{-1} U)$ of order r are invertible. The formula (1.1.4) shows that a rank r deviation added to a matrix entails a rank r deviation in the inverse.

1.2 Norms

The notions of size and distance for a vector space are described by norms. A **norm** is a function $\| \cdot \| : \mathbb{C}^n \to \mathbb{R}$ that assigns a real-valued length to each vector. It is called a norm on $\mathbb{C}^n$ and is denoted by $\| \cdot \|_{\mathbb{C}^n}$. In order to conform to a reasonable notion of length, a norm must satisfy the following three conditions. For all vectors x and y and for all scalars $\alpha \in \mathbb{C}$,

(1) $\|x\| \geq 0$, and $\|x\| = 0$ if and only if $x = 0$,

(2) $\|x + y\| \leq \|x\| + \|y\|$,

(3) $\|\alpha x\| = |\alpha| \, \|x\|$.

An important class of vector norms, consists of the p-norms. We cite the most frequently used ones, corresponding to $p = 1, 2, \infty$ respectively:

$\|x\|_1 = \sum_{j=1}^n |\zeta_j|$,

$\|x\|_2 = (\sum_{j=1}^n |\zeta_j|^2)^{1/2}$,

$\|x\|_\infty = \max_{1\le j\le n} |\zeta_j|$.

The norm $\|x\|_2$ is the Euclidean norm, deriving from the complex scalar product: $\|x\|_2^2 =< x, x >= x^H x$.

By considering a vector norm on $\mathbb{C}^{nr}$, one can define a norm for $A \in \mathbb{C}^{n\times r}$. To take advantage of the structure $\mathcal{L}(\mathbb{C}^r, \mathbb{C}^n)$, one considers arbitrary vector norms $\|\cdot\|_{\mathbb{C}^r}$ and $\|\cdot\|_{\mathbb{C}^n}$ given on $\mathbb{C}^r$ and $\mathbb{C}^n$ respectively. The *induced norm* (or *subordinate norm*) on matrices in $\mathbb{C}^{n\times r}$ is derived from *vector* norms in $\mathbb{C}^r$ and $\mathbb{C}^n$ as follows:

$$\|A\| = \max_{0\neq x\in\mathbb{C}^r} \frac{\|Ax\|_{\mathbb{C}^n}}{\|x\|_{\mathbb{C}^r}}, \quad A \in \mathbb{C}^{n\times r}.$$

When the norm $\|\cdot\|_1$ is used for both $\mathbb{C}^r$ and $\mathbb{C}^n$, then

$$\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^n |a_{ij}|.$$

We have

$$\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^n |a_{ij}|,$$

when the norm $\|\cdot\|_\infty$ is used for both $\mathbb{C}^r$ and $\mathbb{C}^n$.
Let $A \in \mathbb{C}^{n\times r}$ and $B \in \mathbb{C}^{r\times q}$. For any $x \in \mathbb{C}^q$ we have $\|Bx\| \le \|B\|\|x\|$ [40], hence

$$\|ABx\| \le \|A\|\|Bx\| \le \|A\|\|B\|\|x\|.$$

This implies that an induced norm satisfies the following *submultiplicative* property

$$\|AB\| \le \|A\|\|B\|.$$

Matrix norms do not have to be induced by vector norms. A *matrix norm* is a particular vector norm which has the *submultiplicative* property.

Not all norms for A satisfy the submultiplicative property. For instance, $\|A\|_\Delta = \max|a_{ij}|$ defines a matrix norm but for $A = B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, $\|AB\|_\Delta > \|A\|_\Delta\|B\|_\Delta$.

A matrix norm of practical importance which is not induced by a vector norm is the *Hilbert-Schmidt* or *Frobenius* norm, defined for $A \in \mathbb{C}^{n\times r}$ by

$$\|A\|_F = \left(\sum_{i=1}^n \sum_{j=1}^r |a_{ij}|^2\right)^{1/2} = (trA^H A)^{1/2}.$$

An $m \times m$ square matrix Q is called **unitary** if $Q^H Q = QQ^H = I$. This is equivalent to say that an $m \times m$ square matrix $Q \in \mathbb{C}^{m\times m}$ is called unitary if its columns forms an orthonormal basis of $\mathbb{C}^m$.

An important property of unitary matrices is that they preserve Euclidean norms, because inner products are preserved. Via (1.1.1), for unitary matrix $Q \in \mathbb{C}^{m\times m}$ and vectors $x, y \in \mathbb{C}^m$,

$$(Qx)^H(Qy) = x^H y,$$

which simply shows that $\|Qx\|_2 = \|x\|_2$, and also $\|Q\|_2 = 1$.
Matrix 2-norm and Frobenius norm are invariant under multiplication by unitary matrices. In fact, for any $A \in \mathbb{C}^{m\times n}$ and unitary matrix $Q \in \mathbb{C}^{m\times m}$, it is easy to show [52] that

$$\|QA\|_2 = \|A\|_2, \tag{1.2.1}$$

and

$$\|QA\|_F = \|A\|_F. \tag{1.2.2}$$

1.3 Eigenvalues of matrices

Let A be a real or complex square matrix of order n. The **eigenvalue problem** is:

$$find\ \lambda \in \mathbb{C}\ and\ 0 \neq x \in \mathbb{C}^n\ such\ that\ Ax = \lambda x. \tag{1.3.1}$$

The scalar λ in (1.3.1) is called an *eigenvalue* of A and $0 \neq x \in \mathbb{C}^n$ is a *right eigenvector* associated with λ. The complex number λ is an eigenvalue of the matrix A *iff* it is a zero of the so-called *characteristic polynomial* $\pi(z)$:

$$\pi(z) = \det(zI - A), \tag{1.3.2}$$

the determinant of $zI - A$. The polynomial $\pi(z)$ has n zeros in $\mathbb{C}$, not necessarily distinct. The set of such zeros forms the *spectrum* of A and is denoted by $\sigma(A)$. This means that $\sigma(A) = \{\lambda \in \mathbb{C} : \pi(\lambda) = 0\}$.

Proposition 1.3.1 *If λ is an eigenvalue of $A \in \mathbb{C}^{n\times n}$ then $\bar{\lambda}$ is an eigenvalue of A^H. An eigenvector y of A^H associated with the eigenvalue $\bar{\lambda}$ is called left eigenvector of A.*

The eigenvalue λ, the right and left eigenvectors, x and y, satisfy the relations

$$Ax = \lambda x\ ,\quad y^H A = \lambda y^H,$$

for $x, y \neq 0$.

The *spectral radius* of matrix A is a real non negative number defined by

$$\rho(A) = \max\{|\lambda|; \lambda \in \sigma(A)\}.$$

Let $\|\cdot\|$ denote any norm on $\mathbb{C}^n$ and also the induced matrix norm on $\mathbb{C}^{n\times n}$. Then we have

$$\rho(A) \leq \|A\|.$$

The sum of all diagonal elements of matrix $A \in \mathbb{C}^{n\times n}$ is called the *trace* of matrix A.

$$\mathrm{tr}(A) = \Sigma_{i=1}^{n} a_{ii}.$$

If $\lambda_1, \cdots, \lambda_n$ are eigenvalues of matrix $A \in \mathbb{C}^{n\times n}$ (not necessarily distinct), then

$$\mathrm{tr}(A) = \sum_{i=1}^{n} \lambda_i, \tag{1.3.3}$$

and

$$\det(A) = \Pi_{i=1}^{n} \lambda_i. \tag{1.3.4}$$

Let us denote the set of distinct eigenvalues of A by

$$\{\lambda_1, \cdots, \lambda_p \ ; \ p \leq n\},$$

then

a) An eigenvalue λ of A is said to have *algebraic multiplicity* m if it is a root of multiplicity m of the characteristic polynomial, which means

$$\pi(z) = (z-\lambda)^m \pi_1(z); \quad \pi_1(\lambda) \neq 0.$$

b) The *geometric multiplicity* of $\lambda \in \sigma(A)$, denoted by g, is the number of linearly independent eigenvectors which correspond to λ, that is

$$g = \dim \mathrm{Ker}(A - \lambda I).$$

It is clear that $1 \leq g \leq m$.

c) An eigenvalue λ of algebraic multiplicity 1 is called *simple*; otherwise it is said to be *multiple*.

d) An eigenvalue of multiplicity $m > 1$ is said to be *semi-simple* if it admits m linearly independent eigenvectors, that is $g = m$. Otherwise it is said to be *defective*.

e) A matrix is said to be *derogatory* if the geometric multiplicity of at least one of its eigenvalues is larger than one.

A subspace S is said to be *invariant* under a square matrix A if $AS \subset S$. In particular, for any eigenvalue λ of A the subspace $\mathrm{Ker}\,(A - \lambda I)$ is invariant under A. The subspace $\mathrm{Ker}\,(A-\lambda I)$ is called the *eigenspace* associated with λ: it consists of all the eigenvectors of A associated with λ, plus the vector 0.

1.3.1 The Hadamard-Gershgorin Theorem

In some situations one wishes to have a rough (global) idea of where the eigenvalues lie in the complex plane, by directly exploiting some knowledge on the entries of the matrix A. We already know a simple localization result that uses any matrix norm, since we have

$$|\lambda_i| \leq \|A\|,$$

i.e., any eigenvalue belongs to the *disk* centered at the origin and of radius $\|A\|$. A more precise localization result is provided by the following theorem often attributed to Gershgorin, although it was already known to Hadamard.

Theorem 1.3.2 *Any eigenvalue λ of a matrix A is located in at least one of the n closed disks of the complex plane centered at a_{ii} and having the radius*

$$\sum_{j=1,\ j\neq i}^{j=n} |a_{ij}|.$$

In other words,

$$\forall\ \lambda \in \sigma(A),\quad \exists i\ \ \textit{such that}\ \ |\lambda - a_{ii}| \leq \sum_{j=1,\ j\neq i}^{j=n} |a_{ij}|.$$

Let us denote the n closed disks of the complex plane defined in Theorem 1.3.2 by $D_i = \{z \in \mathbb{C} :\ |z - a_{ii}| \leq \sum_{j=1,\ j\neq i}^{j=n} |a_{ij}|\}$, $i = 1, \ldots, n$. It is shown that if the Gershgorin disk D_i is isolated from the other disks, then it contains precisely one of A's eigenvalues [55].

1.4 The adjoint matrix

An important scalar-valued function associated with the elements of a square matrix is the *determinant*. According to *Cramer's rule*, solving linear systems can be realized theoretically by means of determinants.

Let $A = [a_{ij}]_{i,j=1}^{n}$ be an $n \times n$ matrix. A *minor* of order $n-1$ of A is defined to be the determinant of a submatrix of A obtained by striking out one row and one column from A. The minor obtained by striking out the ith row and jth column is denoted by M_{ij} $(1 \leq i, j \leq n)$.

For an arbitrary $n \times n$ matrix A and any i, j with $1 \leq i, j \leq n$, one has

$$\det A = a_{i1}A_{i1} + a_{i2}A_{i2} + \cdots + a_{in}A_{in} \tag{1.4.1}$$

or, similarly,

$$\det A = a_{1j}A_{1j} + a_{2j}A_{2j} + \cdots + a_{nj}A_{nj} \tag{1.4.2}$$

where $A_{pq} = (-1)^{p+q} M_{pq}$. The numbers

$$A_{pq} \ (1 \le p, q \le n) \tag{1.4.3}$$

are called the *cofactors* of the elements a_{pq}, and therefore, formulas (1.4.1) and (1.4.2) are referred to as, respectively, *row* and *column cofactor expansions* of $\det A$.

The *adjoint* of matrix A, written by $\operatorname{adj} A$, is defined to be the transposed matrix of cofactors of A [37, 43]. Thus

$$\mathrm{adj}A = ([A_{ij}]_{i,j=1}^{n})^T. \tag{1.4.4}$$

Also $\operatorname{adj}(\alpha A) = \alpha^{n-1} \operatorname{adj} A$ for $\alpha \in \mathbb{C}$.

Example 1.4.1 *Let* $A = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$. *Then*

$$\mathrm{adj}A = \begin{bmatrix} \begin{vmatrix} 1 & 0 \\ 1 & 1 \end{vmatrix} & -\begin{vmatrix} 0 & 0 \\ 1 & 1 \end{vmatrix} & \begin{vmatrix} 0 & 1 \\ 1 & 1 \end{vmatrix} \\ -\begin{vmatrix} 2 & 1 \\ 1 & 1 \end{vmatrix} & \begin{vmatrix} 1 & 1 \\ 1 & 1 \end{vmatrix} & -\begin{vmatrix} 1 & 2 \\ 1 & 1 \end{vmatrix} \\ \begin{vmatrix} 2 & 1 \\ 1 & 0 \end{vmatrix} & -\begin{vmatrix} 1 & 1 \\ 0 & 0 \end{vmatrix} & \begin{vmatrix} 1 & 2 \\ 0 & 1 \end{vmatrix} \end{bmatrix}^T = \begin{bmatrix} 1 & -1 & -1 \\ 0 & 0 & 0 \\ -1 & 1 & 1 \end{bmatrix},$$

where $\begin{vmatrix} a & b \\ c & d \end{vmatrix} = \det \begin{bmatrix} a & b \\ c & d \end{bmatrix}$.

△

Theorem 1.4.1 *[43] For any nonsingular* $n \times n$ *matrix* A,

$$A^{-1} = \frac{1}{\det A} \mathrm{adj}A. \tag{1.4.5}$$

An immediate corollary of Theorem 1.4.1 is that for $z \notin \sigma(A)$, one has

$$(zI - A)^{-1} = \frac{1}{\pi(z)} \mathrm{adj}(zI - A). \tag{1.4.6}$$

A nonzero scalar polynomial $p(\lambda)$ is defined as an *annihilating polynomial* of the matrix A if $p(A) = 0$. There are many annihilating polynomials for every matrix A. One of them has lowest degree [43]. The monic annihilating polynomial of the least possible degree is called *minimal polynomial* for A and is denoted by $m(z)$.

But the adjoint matrix of $zI - A$ can be derived by an effective formula [37] which uses the characteristic polynomial

$$\pi(z) = \det(zI - A) = z^n - p_1 z^{n-1} - p_2 z^{n-2} - \cdots - p_n, \quad \text{for } z \in \mathbb{C}.$$

where $p_1 = \operatorname{tr} A$ and $p_n = (-1)^{n-1} \det A$.

It is shown [37] that the adjoint matrix of $zI - A$ is

$$\operatorname{adj}(zI - A) = z^{n-1} I + B_1 z^{n-2} + B_2 z^{n-3} + \cdots + B_{n-1}, \tag{1.4.7}$$

where

$$B_1 = A - p_1 I, \quad B_2 = A^2 - p_1 A - p_2 I, \ldots$$

and, in general,

$$B_k = A^k - p_1 A^{k-1} - p_2 A^{k-2} - \cdots - p_k I \quad (k = 1, 2, \ldots, n-1). \tag{1.4.8}$$

The matrices $B_1, \ldots, B_{n-1}$ can be computed in succession, by the recurrence relation

$$B_k = AB_{k-1} - p_k I \quad (k = 1, 2, \ldots, n-1; B_0 = I).$$

Moreover,

$$B_1 = A - (\operatorname{tr} A) I, \quad B_{n-1} = (-1)^{n-1} \operatorname{adj} A, \quad \text{and} \quad AB_{n-1} - p_n I = 0, \tag{1.4.9}$$

with $p_n = (-1)^{n-1} \det(A)$. When A is nonsingular, then $p_n \neq 0$, and it follows from (1.4.9) that B_{n-1} has rank n and we get back (1.4.5), that is

$$A^{-1} = \frac{1}{p_n} B_{n-1} = \frac{1}{\det A} \operatorname{adj} A.$$

When A is singular, $AB_{n-1} = 0$: the n columns of B_{n-1} belong to $\operatorname{Ker} A \neq \{0\}$ and $\operatorname{rank} B_{n-1} \leq \dim \operatorname{Ker} A$, which represents the geometric multiplicity of $0 \in \sigma(A)$.

Let $A \in \mathbb{C}^{n \times n}$ and let $B(z) = \operatorname{adj}(zI - A)$. As noted in (1.4.7), $B(z)$ is a monic matrix polynomial over $\mathbb{C}$ of order n and degree $n - 1$. Let $c(z)$ denote the (monic) greatest common divisor of the elements of $B(z)$. Then it is shown [43] that $\pi(z)$ is divisible by $m(z)$ and $\pi(z)/m(z)$ is the polynomial $c(z)$ just defined: $\pi(z) = c(z)m(z)$, for $m(z) =$ minimal polynomial $= \Pi_{i=1}^{d} (z - \lambda_i)^{l_i}$ with d distinct eigenvalues $\lambda_i \in \sigma(A)$.

Now degree $c(z) \geq 0$, and $c(z) \equiv 1 \Leftrightarrow$ degree $c(z) = 0 \Leftrightarrow$ A is non derogatory $\Leftrightarrow$ $\pi(z) = m(z)$.

When A is derogatory, one gets the *reduced adjoint*

$$C(z) = \frac{1}{c(z)} B(z),$$

for the matrix A. This causes a simplification for the resolvent of derogatory matrices.

1.5 Projections

In this section, we shall introduce the fundamental tool consisting of **projection matrices**. A *projection* is a square matrix P that transforms $\mathbb{C}^n$ into a subspace of itself. Such a matrix is *idempotent*, i.e. $P^2 = P$.
When P is a projection, then $(I-P)$ is a projection too and $\mathrm{Ker}\,(P) = \mathrm{Im}\,(I-P)$ such that $\mathrm{Ker}\,(P) \cap \mathrm{Im}\,(P) = \{0\}$. In addition, every element x of $\mathbb{C}^n$ can be written uniquely as $x = Px + (I-P)x$. As a result, the space $\mathbb{C}^n$ can be decomposed as the direct sum

$$\mathbb{C}^n = \mathrm{Ker}(P) \oplus \mathrm{Im}(P).$$

Conversely, every pair of subspaces S_1 and S_2 that form a direct sum of $\mathbb{C}^n$ define a unique projection P such that $\mathrm{Im}\,(P) = S_1$ and $\mathrm{Ker}\,(P) = S_2$ [34]. The projection P is said to be the *orthogonal projection* onto S_1, when the subspace S_2 is the orthogonal complement of S_1, i.e., when

$$\mathrm{Ker}(P) = \mathrm{Im}(P)^{\perp}.$$

Proposition 1.5.1 *A projection matrix P is orthogonal if and only if it is Hermitian, that is $P^H = P = P^2$.*

1.6 Equivalence transformation on $A \in \mathbb{C}^{m\times n}$

We recall that the *rank* of a matrix is *invariant* under an equivalence transformation.

1.6.1 Singular Value Decomposition

The **singular value decomposition** (SVD) is an equivalence transformation on $A \in \mathbb{C}^{m\times n}$. We assume without loss of generality that $m \geq n$. The *singular values* of the $m \times n$ rectangular matrix A are the non-negative square roots of the eigenvalues of the square matrix $A^H A$ of order n. The $n \times n$ square matrix $\hat{A} = A^H A$ is Hermitian and positive semidefinite, that is for every $0 \neq x \in \mathbb{R}^n$, $x^H \hat{A} x \geq 0$. This means that the singular values of matrix A (the eigenvalues of $\hat{A}$) are real and nonnegative and we can write them in decreasing order of magnitude $0 \leq \sigma_n \leq \cdots \leq \sigma_1$. The $m \times m$ matrix AA^H is also Hermitian positive semidefinite. Its largest n eigenvalues are identical to those of $A^H A$, and the others are zero.

The n eigenvectors of $A^H A$ are called *right singular vectors* for A. We denote them by $v_1, \cdots, v_n$, where v_i is the eigenvector for the eigenvalue σ_i^2. The m eigenvectors of AA^H are called *left singular vectors* which are denoted by $u_1, \cdots, u_m$, where u_1 through u_n are eigenvectors for eigenvalues σ_1^2 through σ_n^2, and u_{n+1} through u_m are eigenvectors for the zero eigenvalues of AA^H.

Theorem 1.6.1 *[38] Let $A \in \mathbb{C}^{m\times n}$ and* $\mathrm{rank}\, A = r \leq min(m,n)$. *There exist unitary matrices U_1 and V_1 of orders m and n respectively such that $U_1^H A V_1 =$*

$diag(\sigma_i) = \Sigma$ is the following $m \times n$ matrix

$$\Sigma = \begin{pmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix},$$

whose only nonzero elements are $\sigma_1, \sigma_2, \cdots, \sigma_r > 0$ on the diagonal where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, for $r = \operatorname{rank} A = \operatorname{rank} AA^H = \operatorname{rank} A^H A$.

By comparing columns in the equations $AV_1 = U_1\Sigma$ and $A^H U_1 = V_1\Sigma^H$, we can see that $Av_i = \Sigma_i u_i$ and $A^H u_i = \Sigma_i v_i$ for $i = 1, \cdots, r$. The left singular vectors $u_1, \cdots, u_m$ (resp. the right singular vectors $v_1, \cdots, v_n$) are used to define the $m \times m$ (resp. $n \times n$) matrix U_1 (resp. V_1) in Theorem 1.6.1.

Under the assumptions of Theorem 1.6.1, we have $A = U_1 \Sigma V_1^H$. Now let us denote the first r columns of the matrix $U_1\Sigma$ by U and the first r columns of the matrix V_1 by V. Then we get

$$A = UV^H. \tag{1.6.1}$$

Using the definitions of the matrices U and V above, every matrix $A \in \mathbb{C}^{m \times n}$ of rank r can be written under the form (1.6.1). **This decomposition plays a key role in HD (with $m = n$).**

The 2-norm and Frobenius norm of A can be expressed in terms of singular values: for $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$, where r is the rank of A,

$$\|A\|_F^2 = \sigma_1^2 + \cdots + \sigma_r^2,$$

$$\|A\|_2 = \sigma_1,$$

$$\min_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sigma_r.$$

1.6.2 The Schur complement matrix

In linear algebra, the *Schur complement* (named after Issai Schur) of an invertible block matrix A within a larger matrix plays an important role. It is defined as follows. Suppose that the 4 blocks A, B, C, D are respectively of size $p \times p, p \times q, q \times p$ and $q \times q$, and that A is *invertible*. Let

$$N = \begin{bmatrix} A & B \\ C & D \end{bmatrix}, \tag{1.6.2}$$

be so that N is a $(p+q)\times(p+q)$ matrix. The *Schur complement* of the block A in the matrix N [43] is the $q\times q$ matrix

$$S = D - CA^{-1}B. \tag{1.6.3}$$

It satisfies

$$\det N = \det \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \det S \, \det A, \tag{1.6.4}$$

which is in accord with the formula for the determinant of a 2×2 matrix with the additional assumption that $\det A \neq 0$. Therefore $\det N = 0 \Leftrightarrow \det S = 0$. We make three remarks:

1. The Schur complement (1.6.3) arises as the result of performing a block *Gaussian elimination* by multiplying the matrix N from the right with the "upper triangular" block matrix

$$U = \begin{bmatrix} A^{-1} & -A^{-1}B \\ 0 & I_q \end{bmatrix}, \tag{1.6.5}$$

which yields

$$NU = \begin{bmatrix} I_p & 0 \\ CA^{-1} & S \end{bmatrix}. \tag{1.6.6}$$

2. One can also block-diagonalize N as follows.

$$\begin{bmatrix} I & 0 \\ CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & S \end{bmatrix} \begin{bmatrix} I & A^{-1}B \\ 0 & I \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}. \tag{1.6.7}$$

The matrix N has the *equivalent* block diagonal form $\begin{bmatrix} A & 0 \\ 0 & S \end{bmatrix}$.

3. For the case $p = 1$, $A = \alpha$ is a scalar, D is a $q\times q$ matrix and both $B^T = b$ and $C = c$ are $q\times 1$ matrices. It is shown ([43], page 65) that

$$\det \begin{bmatrix} \alpha & b^T \\ c & D \end{bmatrix} = \alpha \, \det D - b^T \, \mathrm{adj} D \, c. \tag{1.6.8}$$

When $\det D \neq 0$, (1.6.8) can be written as $\det D(\alpha - b^T D^{-1} c)$, where the scalar $\alpha - b^T D^{-1} c$ is the 1×1 Schur complement of D. The formula (1.6.8) is valid even if $\alpha = 0$, or D is not invertible. This identity will be important for our analysis of Arnoldi's method.

1.7 Similarity transformation on $A \in \mathbb{C}^{n\times n}$

In many applications, the reduction of a square matrix into a simpler form by similarity is useful. We shall be concerned with diagonal, Schur and Jordan forms. Here *similarity* means a transformation that preserves the eigenvalues of a matrix.

Definition 1.7.1 *Two matrices A and B are said to be* similar *if there is a non-singular matrix X such that*

$$A = XBX^{-1}.$$

The mapping $B \mapsto A$ is a *similarity transformation*, which preserves the eigenvalues of the matrix A. An eigenvector u_B of matrix B is transformed into the eigenvector $u_A = Xu_B$ of matrix A.

1.7.1 Diagonalizability

The simplest desired form in which a matrix can be transformed is the **diagonal** form. In fact, the matrix A is diagonalizable *iff* it is similar to a diagonal matrix. This transformation is not always possible. More precisely, the matrix A is *diagonalizable iff* it possesses n linearly independent eigenvectors x_i ($i = 1, \cdots, n$), that is *iff* it can be decomposed into the form

$$A = XDX^{-1}, \tag{1.7.1}$$

where the ith column of X (resp. the ith column of X^{-1}) is the right eigenvector x_i (resp. the left eigenvector x_i^H) associated with the eigenvalue λ_i [34].
The decomposition (1.7.1) exists for every matrix with distinct eigenvalues. But not all matrices with multiple eigenvalues are similar to a diagonal matrix.

Proposition 1.7.2 *[34] A matrix A is diagonalizable* iff *its eigenvalues are semi-simple. In this case, the matrix A is called semi-simple.*

When A is not diagonalizable, it is called defective. In practice, even if A is diagonalizable, the matrix X, though invertible, may be ill-conditioned with respect to inversion. This could make $X^{-1}AX$ difficult to compute in finite precision.

1.7.2 Unitary Diagonalization

It may happen that not only does an $n \times n$ matrix A have n linearly independent eigenvectors, but these can be chosen to be orthogonal. In such a case, A is *unitarily diagonalizable*, that is , there exists a unitary matrix Q such that

$$A = Q\Lambda Q^H. \tag{1.7.2}$$

This is both an eigenvalue decomposition and a singular value decomposition, except for the matter of complex signs ($\sigma = |\lambda|$ for $\lambda \in \mathbb{C}$) of the entries of Λ. By definition, we say that a matrix A is *normal* iff $AA^H = A^HA$. The class of matrices that are unitarily diagonalizable has the following elegant characterization.

Theorem 1.7.3 *A matrix is unitarily diagonalizable if and only if it is normal.*

Hermitian matrices are therefore unitarily diagonalizable.

Theorem 1.7.4 *A Hermitian matrix is unitarily diagonalizable, and its eigenvalues are real.*

Another example for a unitarily diagonalizable matrix is given by unitary matrices. They have eigenvalues on the unit circle centered at 0 in the complex plane.

Theorems 1.7.3 and 1.7.4 follow from Theorem 1.7.5 below.

1.7.3 The Schur form

Any matrix A, is unitarily similar to an upper triangular matrix which is the **Schur form** for A. This is what the following existence theorem asserts.

Theorem 1.7.5 *[34] For every matrix A, there exists a unitary matrix Q such that $S = Q^H AQ$ is an upper triangular matrix whose diagonal elements are the eigenvalues $\lambda_1, \cdots, \lambda_n$ in an arbitrary order.*

The chosen order for the eigenvalues $\{\lambda_i\}$ determines the Schur basis Q up to a unitary block-diagonal matrix. The matrix A in Theorem 1.7.5 is normal *iff* the matrix S is diagonal. We get the theorems 1.7.3 and 1.7.4 as corollaries.

Similarity transformation by unitary matrices is optimal from a computational point of view. The so-called algorithm QR, provides the Schur form after an infinite number of steps. In practice, it is the most powerful and reliable software to compute eigenvalues and roots of polynomials of degree less than a few 10^3.

1.7.4 The Jordan (canonical) form

The **Jordan form** is an upper bidiagonal matrix with only *ones* or *zeros* on the first super diagonal. This reduction is always possible as proved by C. Jordan. When A is in Jordan form, the invertible Jordan basis matrix X replaces the unitary Schur matrix Q, and the bidiagonal matrix $J = D + U$ replaces the upper triangular matrix $S = D + N$ in the Schur form. Here D is a diagonal matrix of eigenvalues, N is a triangular nilpotent matrix, and U is a matrix with 1 or 0 on its first super-diagonal and 0 elsewhere. The following theorem establishes the Jordan form of an arbitrary matrix.

Theorem 1.7.6 *[34] Let A be a matrix of order n with distinct eigenvalues $\lambda_1, \cdots, \lambda_p$ ($p \leq n$). Then there exists an invertible matrix X such that*

$$X^{-1}AX = diag(J_{ij}) = J, \tag{1.7.3}$$

where

$$J_{ij} = \lambda_i I + U_{ij}$$

and U_{ij} *is a matrix of order* k_{ij} *of the form*

$$U_{ij} = \begin{bmatrix} 0 & I_{k_{ij}-1} \\ & \\ 0 & 0 \end{bmatrix} \quad (j = 1, \cdots, g_i),$$

that is there are g_i *blocks* U_{ij} *corresponding to a particular* λ_i *.*

The set of *Jordan blocks* J_{ij} $(j = 1, \cdots, g_i)$ associated with λ_i constitutes the *Jordan box* associated with λ_i. The order of the Jordan box J_i is

$$m_i = k_{i1} + \cdots + k_{ig_i},$$

and it contains g_i blocks, so

$$g_i \leq m_i.$$

Here, we list some important consequences of Theorem 1.7.6.

a) For every integer l and each eigenvalue λ_i, one has the inclusion

$$\mathrm{Ker}(A - \lambda_i I)^{l+1} \supset \mathrm{Ker}(A - \lambda_i I)^l \quad \text{in} \quad \mathbb{C}^n.$$

b) The above property implies that there is a least integer $l_i \geq 1$ such that

$$\mathrm{Ker}(A - \lambda_i I)^{l_i+1} = \mathrm{Ker}(A - \lambda_i I)^{l_i},$$

and in fact $\mathrm{Ker}\,(A - \lambda_i I)^l = \mathrm{Ker}\,(A - \lambda_i I)^{l_i}$ for all $l \geq l_i$. The integer l_i is called the *index* (or *ascent*) of λ_i.

c) The subspace

$$M_i = \mathrm{Ker}(A - \lambda_i I)^i \tag{1.7.4}$$

is invariant under A. Moreover,

$$\mathbb{C}^n = \oplus_{i=1}^p M_i,$$

and we have

$$\dim(M_i) = m_i,$$

which is the algebraic multiplicity of λ_i.

d) Since $\mathbb{C}^n$ is the direct sum of the subspaces $M_i,\ i = 1, \cdots, p$, each vector $x \in \mathbb{C}^n$ can be written in a unique way as

$$x = x_1 + x_2 + \cdots + x_i + \cdots + x_p,$$

where x_i is a member of the subspace M_i. The linear transformation defined by

$$P_i : x \to x_i$$

is a projection onto M_i along the direct sum of the subspaces $M_j,\ j \neq i$. The family of projections $P_i,\ i = 1, \cdots, p$ satisfies the following properties,

$$\mathrm{Im}(P_i) = M_i \tag{1.7.5}$$

$$P_i P_j = P_j P_i = 0,\ if\ i \neq j \tag{1.7.6}$$

$$\sum_{i=1}^{p} P_i = I. \tag{1.7.7}$$

Any family of projections that satisfies the above three properties is uniquely determined and is associated with the decomposition of $\mathbb{C}^n$ into the direct sum of the images of the $P_i's$ [49].

The matrix representation J of A in the new basis described in (1.7.3) can be expanded as follow,

$$X^{-1}AX = J = \begin{bmatrix} J_1 & & & & & \\ & J_2 & & & & \\ & & \ddots & & & \\ & & & J_i & & \\ & & & & \ddots & \\ & & & & & J_p \end{bmatrix} \tag{1.7.8}$$

where each J_i corresponds to the subspace M_i associated with the eigenvalue λ_i. The size of J_i is m_i and its structure is as follows,

$$J_i = \begin{bmatrix} J_{i1} & & & \\ & J_{i2} & & \\ & & \ddots & \\ & & & J_{ig_i} \end{bmatrix}$$

with

$$J_{ik} = \begin{bmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \lambda_i & 1 \\ & & & \lambda_i \end{bmatrix}, \quad k = 1, \ldots, g_i.$$

Each of the blocks J_{ik} corresponds to a different eigenvector associated with the eigenvalue λ_i.

Example 1.7.1 *When the size of every Jordan block J_{ik} in each Jordan box J_i corresponding to the matrix $A \in \mathbb{C}^{n\times n}$ is one, the matrix A is diagonalisable. Such a matrix is semi-simple ($g_i = m_i$ for all eigenvalues).*

Example 1.7.2 *For the matrix A defined by*

$$A = \begin{bmatrix} 1 & 0 & 2 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix},$$

the Jordan form, J, and the similarity transformation matrix, X, are

$$J = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \; X = \begin{bmatrix} 2 & 2 & -2 & 0 & 0 \\ 0 & 1 & -1 & -1 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

respectively such that $X^{-1}AX = J = \begin{bmatrix} J_1 & \\ & J_2 \end{bmatrix}$ *where* $J_1 = [0]_{1\times 1}$,

$J_2 = \begin{bmatrix} J_{21} & & \\ & J_{22} & \\ & & J_{23} \end{bmatrix}$, *for* $J_{21} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ *and* $J_{22} = J_{23} = [1]_{1\times 1}$.

The complete Jordan decomposition of A is

$$J = diag\left[[0], \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, [1], [1]\right].$$

△

1.8 Spectral decomposition

Let $\lambda_1, \cdots, \lambda_p$ be the distinct eigenvalues of A. The *spectral projection* associated with λ_i of algebraic multiplicity m_i is the projection P_i on the invariant subspace $M_i = \mathrm{Ker}\,(A - \lambda_i I)^{m_i}$ parallel to $\bigoplus_{j\neq i} M_j$. The following spectral decomposition exists for every matrix $A \in \mathbb{C}^{n\times n}$.

Theorem 1.8.1 *[34] Any matrix A possesses a spectral decomposition of the form*

$$A = \sum_{i=1}^{p} (\lambda_i P_i + D_i), \quad D_i^{l_i} = 0, \tag{1.8.1}$$

where $D_i = (A - \lambda_i I)P_i$ and l_i is the index of λ_i.

By theorem 1.7.6, we have $A = XJX^{-1}$ where J is a block-diagonal matrix consisting of p Jordan boxes $B_1, \cdots, B_p$. The box B_i is an $m_i \times m_i$ matrix of the form

$$B_i = \lambda_i I_{m_i} + N_i$$

where N_i is a matrix whose only non-zero elements appear on the first superdiagonal and can be taken to be equal to unity.

Let X_i (resp. $X_{i^*}^H$) be the matrix constructed by the m_i columns of X (resp. rows of X^{-1}) which are associated with λ_i. The column vectors of X_i are a basis for M_i which possesses as the adjoint basis the corresponding row vector of X^{-1}, that is $X_{i^*}^H X_i = I_{m_i}$. The matrix

$$P_i = X_i X_{i^*}^H$$

represents the projection on M_i parallel to $\bigoplus_{j \neq i} M_i = \{x \in \mathbb{C}^n : \ X_i^H x = 0\}$.

If A is diagonalizable, then for each eigenvalue λ_i the invariant subspace M_i is identical to the eigenspace $\mathrm{Ker}\,(A - \lambda_i I)$, that is $l_i = 1$, $D_i = 0$. In this case it is said the spectral projection P_i reduces to the *eigenprojection.*

Corollary 1.8.2 *[34] The spectral decomposition of a diagonalizable matrix A is of the form*

$$A = \sum_{i=1}^{p} \lambda_i P_i, \tag{1.8.2}$$

where P_i is the eigenprojection associated with λ_i.

Proposition 1.8.3 *[34] The transpose conjugate matrix A^H admits the following Jordan decomposition*

$$A^H = \sum_{i=1}^{p} (\bar{\lambda}_i P_i^H + D_i^H). \tag{1.8.3}$$

The Proposition 1.8.3 shows that λ_i and $\bar{\lambda}_i$ have the same multiplicities and indices in A and A^H respectively.

1.9 Matrix polynomials and pencils

1.9.1 Matrix polynomials, PEVP, and PEP

We consider, the matrix polynomial

$$P(z) = z^m A_m + z^{m-1} A_{m-1} + \cdots + A_0, \tag{1.9.1}$$

where $A_k \in \mathbb{C}^{n\times n}$, $k = 0, \cdots, m$ and $z \in \mathbb{C}$. The number m is called the *degree* of the matrix polynomial, provided $A_m \neq 0$. The number n is called the *order* of the matrix polynomial [37].
Next we consider the scalar polynomial of degree $\leq nm$

$$\det P(z) = \det(P(z)) = \sum_{j=0}^{nm} \zeta_j z^j \tag{1.9.2}$$

with $\zeta_{nm} = \det(A_m)$ and $\zeta_0 = \det(A_0)$. We write

$$Z = Z(\det P(z)) = \{z \in \mathbb{C}; \det P(z) = 0\}. \tag{1.9.3}$$

When $\det P(z)$ is not identically zero, it is said that $P(z)$ is *regular* [43]. In this case $\det P(z)$ has at most nm roots in $\mathbb{C}$.

When $\det P(z) \equiv 0$, then the matrix polynomial $P(z)$ is called *singular* and $Z(\det P(z)) = \mathbb{C}$.

Classically, two problems are associated with $P(z)$ in (1.9.1) for $m \geq 1$:

$$find\ \lambda \in \mathbb{C}\ such\ that\ \det P(\lambda) = 0, \tag{1.9.4}$$

$$find\ \lambda \in \mathbb{C}\ and\ X \in \mathbb{C}^{n\times nm}\ such\ that\ P(\lambda)X = 0. \tag{1.9.5}$$

We refer to (1.9.4) as a *polynomial eigenvalue problem*, PEVP, and to (1.9.5) as a *polynomial eigenproblem*, PEP.
These two problems are related but not equivalent. For instance, for $m = 1$, where $X \in \mathbb{C}^{n\times n}$ and $A_1 = I_n$ (where we have an *ordinary eigenvalue problem*), the condition numbers of the eigenvalues and eigenvectors may be decoupled [34, 23].
In the generic situation, one has:

Lemma 1.9.1

(a). $\operatorname{card}(Z(\det P(z))) = nm$ *iff* $\zeta_{nm} \neq 0$, *that is* A_m *invertible.*

(b). $0 \notin Z(\det P(z))$ *iff* $\zeta_0 \neq 0$, *that is* A_0 *invertible.*

Proof. Clear by (1.9.2) □

Exceptional cases correspond to less than nm roots (iff $\zeta_{nm} = 0$), or to at least one zero root (iff $\zeta_0 = 0$).

1.9.2 Matrix pencils, GEVP, and GEP

A matrix pencil $A - zB$ is a particular case of matrix polynomial (1.9.1) for which we have $m = 1$, $A_1 = -B$ and $A_0 = A$ in (1.9.1). A matrix pencil $(A, B) = \{A - zB : z \in \mathbb{C}\}$ is *regular* if $\det(A - zB)$ is not identically zero [37, 34]. Here, we restrict our study to regular matrix pencils.

A *generalized eigenvalue problem*, GEVP, consists in finding the eigenvalues of the pencil $A - zB$ defined by

$$sp(A, B) = \{\lambda \in \mathbb{C} \ : \ \det(A - \lambda B) = 0\}, \tag{1.9.6}$$

where

$$z \mapsto \det(A - zB) \tag{1.9.7}$$

is a polynomial of degree n iff B is *nonsingular*. This means that GEVP has n finite eigenvalues iff $\operatorname{rank} B = n$. These are the roots of the scalar polynomial (1.9.7).

When B is nonsingular then $sp(A, B) = sp(B^{-1}A, I) = sp(B^{-1}A)$. This suggests a method for solving (1.9.6) when B is regular:

- Solve $BC = A$ for C using (say) Gaussian elimination with pivoting.
- Use the QR algorithm to compute the eigenvalues of C.

Note that when B is ill-conditioned, a popular alternative approach to the $A - \lambda B$ problem is the QZ method [38, 10].

The *generalized eigenproblem*, GEP, consists in finding $\lambda \in sp(A, B)$ and $0 \neq x \in \mathbb{C}^n$ an associated eigenvector such that

$$Ax = \lambda Bx \qquad x \neq 0. \tag{1.9.8}$$

The eigenvectors generate a basis $X \in \mathbb{C}^{n \times n}$ iff they are independent.

When B is rank deficient, then $sp(A, B)$ may be finite, empty, or infinite. This is shown by the example 1.9.1.

Example 1.9.1

(i) For $A = \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix}$ *and* $B = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$, *we have* $sp(A, B) = \{1\}$.

This means that $sp(A, B)$ *is a finite set.*

(ii) For $A = \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix}$ *and* $B = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, *we have* $sp(A, B) = \emptyset$.

(iii) For $A = \begin{bmatrix} \alpha & 0 \\ 0 & 0 \end{bmatrix}$ *and* $B = \begin{bmatrix} \beta & 0 \\ 0 & 0 \end{bmatrix}$, *we have* $sp(A, B) = \mathbb{C}$.

This means that $sp(A, B)$ *is an infinite set. In this case* $\operatorname{Ker} A \cap \operatorname{Ker} B \neq \{0\}$.

△

1.9.3 Linearization of a PEP by augmentation

Lemma 1.9.2 *The PEP in (1.9.5) is equivalent to a generalized eigenproblem, GEP, of the form*

$$\mathcal{A}\xi = \lambda\mathcal{B}\xi, \quad \xi \in \mathbb{C}^{nm} \tag{1.9.9}$$

where $\mathcal{A}$ *(resp.* $\mathcal{B}$*) is in block-companion (resp. block-diagonal) form of order* nm *with blocks of order* n*.*

Proof. Let I_n be the identity matrix of order n. Using the following block structures

$$\mathcal{A} = \begin{pmatrix} 0 & I_n & 0 & \cdots & 0 \\ 0 & 0 & I_n & \ddots & \vdots \\ \vdots & & & \ddots & 0 \\ & & & & I_n \\ -A_0 & -A_1 & -A_2 & \cdots & -A_{m-1} \end{pmatrix}, \tag{1.9.10}$$

and

$$\mathcal{B} = \begin{pmatrix} I_n & & & & \\ & I_n & & & \\ & & \ddots & & \\ & & & I_n & \\ & & & & A_m \end{pmatrix}, \tag{1.9.11}$$

it is easy to check that

$$\xi = \begin{pmatrix} x \\ \lambda x \\ \vdots \\ \lambda^{m-2}x \\ \lambda^{m-1}x \end{pmatrix}, \tag{1.9.12}$$

with $0 \neq x \in \mathbb{C}^n$, satisfies (1.9.9). Therefore $\operatorname{rank}\mathcal{B} = \operatorname{rank} A_m$.

□

Chapter 2

Theory of Homotopic Deviation, I

2.1 Introduction

The coupling of the square matrices A and E by the complex parameter t into $A(t) = A + tE$, $t \in \mathbb{C}$ is **Inexact Computing**. The matrix A becomes $A(t)$ by addition of the matrix tE which has a *fixed* structure E as t varies in $\mathbb{C}$. The matrix E is the *deviation* matrix.

Let $B = A + E$. The term *homotopy method* has been given in numerical analysis to the study of the family of matrices $A(t) = A + tE$ when the parameter t is restricted to be *real* in $[0, 1]$, so that $A(0) = A$ and $A(1) = B$.

Because the eigenvalues of $A(t)$ are *complex*, the restriction $t \in [0, 1]$ is too limiting: it is necessary to consider t in $\mathbb{C}$. The analytic properties of $t \mapsto R(t, z) = (A(t) - zI)^{-1}$ and $t \mapsto \sigma(A(t))$ for $t \in \mathbb{C}$ with $|t|$ small enough, have been used by various authors to relate the eigenvalues of $A = A(0)$ and $B = A(1)$, in particular to get upper bounds on the distance between their spectra. See for example Kato (1965 , [42]), or Chatelin (1983 , [32], 1993 , [34]).

More recently, with the advent of easy-to-use graphical software in the $1990s$, it became possible to explore the properties of the non linear spectral map $t \in \mathbb{C} \mapsto \sigma(A(t)) = \{\lambda(t) \text{ eigenvalue of } A(t)\} \in \mathbb{C}^n$, where the color is used to parameterize the variation of $t = he^{i\theta}$, with either $h = |t|$ or θ fixed. See the PhD thesis of E. Traviesas in 2000 [51], and [27, 29]. It is important to observe that in [27, 51], the variation of $|t|$ is *bounded*.

A crucial observation was made during the Summer 2002 by Prof. M. B. van Gijzen (University of Delft) upon his arrival at Cerfacs as a Senior Researcher. He realized that it made sense to look at the limits of $R(t, z)$ and $\sigma(A(t))$ as $|t| \to \infty$. See [16, 30]. The reason for that will become clear later.

To make explicit the difference between classical analytic perturbation theory ($|t|$ small enough or t in $\mathbb{C}$) and the new viewpoint where $|t|$ is unbounded in $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$, the name *Homotopic Deviation theory* has been coined.

Inexact Computing offers a computational approach for the study of the parameter dependence:

$$(t, z) \in \hat{\mathbb{C}} \times \mathbb{C} \to (A + tE - zI)^{-1} = R(t, z)$$

based on the factorization

$$A(t, z) = A + tE - zI = (I + tE(A - zI)^{-1})(A - zI). \tag{2.1.1}$$

In $A(t, z)$, the parameter t multiplies E (possibly rank deficient) whereas z multiplies I (nonsingular and semi-simple).
Analytic perturbation theory [11, 32, 42, 48] is the favourite tool to study the phenomena of Inexact Computing locally for $|t|$ small enough, however there may be *non local effects*. To study such a possibility, the theory of **Homotopic Deviation** was developed [16, 18, 21, 31].
The framework of Homotopic Deviation allows us to perform a **Backward Analysis** for computational methods which are **Inexact** [21, 4]. Homotopic Deviation is also of interest for engineering when the *complex* parameter t has a physical meaning and can be naturally unbounded [30, 31].
Various approaches have been used to study $A(t, z)$, ranging from analytic/algebraic spectral theory [8, 32, 33, 34, 23, 30] to linear control system theory [43, 35]. The theory presented here is *Homotopic Deviation* [16, 18, 30, 21, 5, 4] which specifically looks beyond analyticity for $|t|$ large and unbounded: $t \in \hat{\mathbb{C}}$.

2.2 Presentation of the Homotopic Deviation theory (HD)

Given the matrices A and E in $\mathbb{C}^{n\times n}$, the family

$$A(t) = A + tE, \tag{2.2.1}$$

for $t \in \mathbb{C}$ represents the coupling between A and E by the complex parameter t. We denote the *spectrum* of A by $\sigma(A)$ and the *resolvent* set of A by $re(A) = \mathbb{C}\backslash\sigma(A)$.

By definition [33, 34], the matrix

$$R(t, z) = (A + tE - zI)^{-1}$$

is called the *resolvent matrix*, because it allows to solve the associated linear system

$$(A + tE - zI)x(t, z) = y. \tag{2.2.2}$$

From the point of view of solving the system (2.2.2) when *non local* effects of singularities are of importance, the following two categories of questions are considered:

Q1) existence and analyticity of the resolvent

$$t \in \hat{\mathbb{C}} \to R(t,z) = (A(t) - zI)^{-1},\ z \in \mathbb{C}, \tag{2.2.3}$$

and, if existence, its limit as $|t| \to \infty$.

Q2) limit of the spectrum

$$\lim_{|t|\to\infty} \sigma(A(t)), \tag{2.2.4}$$

that is the limit, as $|t| \to \infty$, of the spectrum of $A(t)$ (or of the singularity set of $R(t,z)$).

In section 2.3, we will consider the questions of existence and analyticity of the resolvent $R(t,z)$, the limits $\lim_{|t|\to\infty} R(t,z)$ and $\lim_{|t|\to\infty} \sigma(A(t))$ for the case where the matrix E in (2.2.1) is regular. It will appear that the case of interest corresponds to the deviation matrix E being *singular*, which yields a much richer situation. In section 2.4, we will discuss the existence and analyticity of resolvent matrix $R(t,z)$ when the matrix E is singular. Questions about the limit of the spectrum of $A(t)$ as $|t| \to \infty$ when $0 \in \sigma(E)$ is semi-simple will be answered in section 2.5.

Notation 2.2.1 The sets considered in HD consist of roots of polynomials. Therefore points are counted *with* their algebraic multiplicity, unless otherwise stated.

$A \subset B \iff$ *if* $x \in A$ *then* $x \in B$ *and the algebraic multiplicity of* x *relative to* A *is not larger than its algebraic multiplicity relative to* B.

Observe that this **differs** from the usual definition in *set theory.* When we occasionally use this classical notion, we denote it $\subset_s$.

2.3 HD when the deviation matrix E is regular

Let us denote $re(A) = \mathbb{C}\backslash\sigma(A)$. For $z \in re(A)$, one can write,

$$R(t,z) = (A + tE - zI)^{-1} = (A - zI)^{-1}(I + tE(A - zI)^{-1})^{-1}. \tag{2.3.1}$$

By substituting the notation

$$F_z = -E(A - zI)^{-1} = E(zI - A)^{-1},\ z \in re(A), \tag{2.3.2}$$

into (2.3.1), we get

$$R(t,z) = R(0,z)(I - tF_z)^{-1}, \tag{2.3.3}$$

which exists for $t \neq \frac{1}{\mu_z}$, $\mu_z \in \sigma(F_z)\backslash\{0\}$. $R(t,z)$ is computable as

$$R(t,z) = R(0,z)\sum_{k=0}^{\infty}(tF_z)^k \tag{2.3.4}$$

for $|t| < \frac{1}{\rho(F_z)}$, and $\rho(F_z) = \max\{|\mu_z|,\ \mu_z \in \sigma(F_z)\}$.

Lemma 2.3.1 *[16] The point* $z \in re(A)$ *is an eigenvalue of* $A+tE$ iff *there exists an eigenvalue* $\mu_z \neq 0$ *of* F_z *such that* $t\mu_z = +1$.

Proof. Let $y \neq 0$ be an eigenvector of F_z associated with $\mu_z \neq 0$. Then

$$-E(A-zI)^{-1}y = \mu_z y, \quad y \neq 0 \quad \Leftrightarrow -Eu = \mu_z(A-zI)u,$$

for $u = (A-zI)^{-1}y \neq 0$. It means that,

$$-E(A-zI)^{-1}y = \mu_z y, \quad y \neq 0 \quad \Leftrightarrow (A + \frac{1}{\mu_z}E)u = zu.$$

□

We assume that the rank of matrix E in (2.2.1) is equal to n , then $0 \notin \sigma(F_z)$. Let the n eigenvalues of F_z be denoted by $\mu_{iz}, i = 1, \cdots, n$. Therefore $R(t,z)$ is defined for all $t \in \mathbb{C}$, $t \neq t_i = \frac{1}{\mu_{iz}}, i = 1, \cdots, n$. Consequently any $z \in re(A)$ is an eigenvalue of the n matrices $A(t_i), i =, \cdots, n$. What happens for $R(t,z)$ in the limit when $|t| \to \infty$?

Proposition 2.3.2 *[16] When the matrix* E *is regular, then for* $z \in re(A)$

$$\lim_{|t|\to\infty} R(t,z) = 0.$$

Proof. We set $s = \frac{1}{t}$, $t \neq 0$. Then, one has

$$I - tF_z = (sF_z^{-1} - I)tF_z,$$

and

$$(I - tF_z)^{-1} = -sF_z^{-1}(I - sF_z^{-1})^{-1} \to 0 \ \ as \ \ s \to 0$$

hence

$$\lim_{|t|\to\infty} R(t,z) = \lim_{|t|\to\infty} R(0,z)(I - tF_z)^{-1} = 0.$$

□

The relation $A+tE = t(E+\frac{1}{t}A) = \frac{1}{s}(E+sA)$, with $s = \frac{1}{t}$, shows that $A(t) = A+tE$ and $E(s) = E + sA$ share the same eigen/Jordan vector structure for $s, t \neq 0$ [18]. Below, we use the relationship between the spectra $\sigma(A(t))$ and $\sigma(E(s))$ to find $\lim_{|t|\to\infty} \sigma(A(t))$ when the matrix E is regular.

The definition of the characteristic polynomial for both $A(t)$ and $E(s)$ can be used to show the relationship between their eigenvalues.

Lemma 2.3.3 *For every* $i \in \{1, \cdots, n\}$, $\lambda_i(t) = \frac{\nu_i(s)}{s}$, *where* $\lambda_i(t) \in \sigma(A + tE)$, $\nu_i(s) \in \sigma(E + sA)$ *and* $s = \frac{1}{t}$, $s, t \neq 0$.

Proof. For $t \neq 0$, one has

$$\sigma(A+tE) = \{\lambda(t) : \det(A+tE-\lambda(t)I) = 0\}, \tag{2.3.5}$$

which for $s = \frac{1}{t}$, is equal to

$$\sigma(A+tE) = \{\lambda(t) : \frac{\det(E+sA-s\lambda(t)I)}{s^n} = 0\}. \tag{2.3.6}$$

Since $t \neq 0$, therefore using (2.3.5) and (2.3.6), it is clear that $\nu_i(s) = s\lambda_i(t)$ for $i = 1, \cdots, n$.

□

Using the above lemma, the limit of every $\lambda(t) \in \sigma(A+tE)$, as $|t| \to \infty$ can be computed as follows

$$\lim_{|t|\to\infty} \lambda(t) = \lim_{|s|\to 0} \frac{\nu(s)}{s}, \tag{2.3.7}$$

where $\nu(s)$ belongs to $\sigma(E+sA)$.

Proposition 2.3.4 *[16] When the matrix E is regular, then for every $\lambda_i(t) \in \sigma(A(t))$, $i = 1, \ldots, n$,*

$$\lim_{|t|\to\infty} \lambda_i(t) = \infty.$$

Proof. For $s = \frac{1}{t}$, $t \neq 0$, we have

$$A(t) = A + tE = t(sA+E) = \frac{1}{s}(E+sA).$$

The lemma 2.3.3 says that an eigenvalue $\lambda(t)$ of $A(t)$ is such that

$$\lambda(t) = \frac{\nu(s)}{s},$$

for $\nu(s) \in \sigma(E+sA)$. Clearly, by continuity,

$$\nu(s) \to \nu \in \sigma(E) \quad as \quad s \to 0, \quad s \neq 0.$$

But here $\nu \neq 0$ which implies $|\lambda(t)| \to \infty$.

□

2.4 The matrix E is singular : $1 \leq r < n$

2.4.1 Existence of $R(t,z)$, $t \in \hat{\mathbb{C}}$

In this section, we shall consider the family $A(t)$ with a singular matrix $E \neq 0$. We recall that any matrix $E \neq 0$ of rank $r \leq n$ can be written under the form

$$E = UV^H, \tag{2.4.1}$$

where $U, V \in \mathbb{C}^{n\times r}$ of rank r represent a basis for Im E, Im E^H respectively. We assume from now on in this chapter that $1 \leq r < n$. dim Ker $E = g$, $1 \leq g \leq n-1$; $g = n - r$ is the *geometric multiplicity* of $0 \in \sigma(E)$.

For a singular matrix E with rank r, the $n \times n$ matrix F_z in (2.3.2) has rank r, so at most r eigenvalues $\mu_{iz}, i = 1, \cdots, r$ are nonzero. These are the r eigenvalues of the $r \times r$ matrix M_z defined by

$$M_z = V^H(zI - A)^{-1}U \in \mathbb{C}^{r\times r}, \tag{2.4.2}$$

for $z \in re(A)$.

By applying (1.1.4) for

$$(I - tF_z)^{-1}$$

in (2.3.3), one has

$$R(t,z) = R(0,z)[I_n - tU(I_r - tM_z)^{-1}V^H R(0,z)], \tag{2.4.3}$$

which exists for $z \in re(A)$, $t \neq \frac{1}{\mu_z}$, where $0 \neq \mu_z \in \sigma(M_z)$. This means that $R(t,z)$ is not defined for $z \in re(A)$ when $t \in \mathbb{C}$ satisfies $t\mu_z = 1$, $0 \neq \mu_z \in \sigma(M_z)$. If M_z is *regular*, this condition is equivalent to $t \in \sigma(M_z^{-1})$.

The fact that $z \in re(A)$ is an eigenvalue of $A + tE$ iff $t\mu_z = 1$ is of fundamental importance. It means that any z in $re(A)$ is an *inexact* eigenvalue for A at homotopic distance $|t| = 1/|\mu_z|$ which can be unbounded: if $\mu_z = 0$ then $|t| = \infty$. This is the reason why it makes sense to look at the limits as $|t| \to \infty$, as was remarked by Prof. M. B. van Gijzen. We therefore assume that $t \in \hat{\mathbb{C}}$. And z is the *exact* eigenvalue of r matrices $A(t_i) = A + t_i E$ with $t_i = \frac{1}{\mu_{iz}} \in \mathbb{C}$, $\mu_{iz} \neq 0$, $i = 1, \cdots, r$, when M_z is of rank r.

When $r = 1$, U and V are the vectors u, v. There is a unique homotopic distance $|t| = 1/|\mu_z|$ where $\mu_z = v^H(zI - A)^{-1}u$. But when $r > 1$, the homotopic distance is not uniquely defined. There are at most r homotopic distances $|t_i| = 1/|\mu_{iz}|$, $\mu_{iz} \neq 0$.

The $r \times r$ matrix M_z defined in (2.4.2), *for $z \in re(A)$, plays a crucial role in* HD.

This role will become clear as we progress.

2.4.2 Frontier points in $F(A,E) \subset re(A)$

Definition 2.4.1 *We call frontier points the elements in the frontier set*

$$F(A,E) = \{z \in re(A);\ \text{rank}M_z < r\},$$

for which M_z is rank *deficient. A point z in $re(A)$ which is not in $F(A,E)$ is generic.*

The term *frontier point* comes from the following

Proposition 2.4.2 *For $z \in re(A)\backslash F(A,E)$ the matrix pencil $(A - zI) + tE$ has exactly r finite eigenvalues and $n - r = g$ infinite ones.*

Proof. See [20] Lemma A1. and Lemma A2. and my talk at ICIAM07 [3]. □

When $z \in F(A,E)$, there are less than r *finite* eigenvalues for the pencil. The exact number depends on the location of z in $F(A,E)$ as we shall see in Chapter 4.

$F(A,E)$ is a set in $re(A)$ which can be empty (for $r = n$ for example), discrete with finite cardinal, or continuous equal to $re(A)$. This will be proved in due course. Proposition 2.4.4 below will justify even more the term "frontier point".

Definition 2.4.3 *A point z in $F(A,E)$ is critical when $\rho(M_z) = 0$. The set of critical points is denoted by $C(A,E)$.*

At a critical point, M_z is nilpotent ($M_z^\delta = 0$ with $M_z^{\delta-1} \neq 0$, $1 \leq \delta \leq r$). Therefore

$$(I_r - tM_z)^{-1} = \sum_{k=0}^{\delta-1} (tM_z)^k, \tag{2.4.4}$$

and $t \mapsto R(t,z)$ is a polynomial of degree δ at a critical point. The right hand side of (2.4.4) can be computed for any t in δ steps. δ is the size of the largest Jordan block (that is the ascent or index) of $0 \in \sigma(M_z)$ [33, 34]. It is algorithmically important to look at the question of whether M_z can be nilpotent for some $z \in re(A)$. We leave the investigation of this question for the chapter 3. For now, we only remark that $F(A,E) = C(A,E)$ when $r = 1$.

2.4.3 $R(\infty,z) = \lim_{|t|\to\infty} R(t,z)$ for $z \in re(A)\backslash F(A,E)$

Let the $r \times r$ matrix M_z be regular for $z \in re(A)\backslash F(A,E)$. We can order the magnitude of the eigenvalues of M_z decreasingly such that, with an additional subscript index running from 1 to r, we have

$$|\mu_{1z}| \geq |\mu_{2z}| \geq \cdots \geq |\mu_{rz}| > 0, \tag{2.4.5}$$

or

$$|t_{1z}| \leq |t_{2z}| \leq \cdots \leq |t_{rz}| < \infty, \tag{2.4.6}$$

where $\mu_{iz} = 1/t_{iz}$ is the ith eigenvalue of M_z in terms of magnitude. We suppose that $|t| > \frac{1}{\min_{1\leq i\leq r}|\mu_{iz}|} = 1/|\mu_{rz}| = \rho(M_z^{-1})$ for M_z of rank r.

Proposition 2.4.4 *[21] For $1 \leq r < n$ and z given in $re(A)\backslash F(A,E)$, $\lim_{|t|\to\infty} R(t,z)$ exists, and is denoted by $R(\infty,z)$. Its representation in closed form is given by*

$$R(\infty,z) = R(0,z)[I_n + UM_z^{-1}V^H R(0,z)].$$

Proof. By assumption, M_z^{-1} exists. $I_r - tM_z = (sM_z^{-1} - I_r)tM_z$,

$$(I_r - tM_z)^{-1} = -sM_z^{-1}(I_r - sM_z^{-1})^{-1},$$

Now

$$-tU(I_r - tM_z)^{-1} = UM_z^{-1}(I_r - sM_z^{-1})^{-1} \to UM_z^{-1}$$

when

$$|t| \to \infty \ (\ or\ |s| \to 0).$$

The rest follows from (2.4.3).

□

Suppose that $\operatorname{rank} U = \operatorname{rank} V = r < n$. Then the $r \times r$ matrix $V^H U$ (resp. $M_z = V^H(zI - A)^{-1}U$) is regular iff 0 is a semi-simple eigenvalue for E (resp. F_z) of multiplicity $g = n - r$ [16]. In proposition 2.4.4, $P_{rz} = I_n + UM_z^{-1}V^H R(0,z)$ is the eigenprojection for F_z associated with the semi-simple eigenvalue $0 \in \sigma(F_z)$ of multiplicity $g = n - r$.

When M_z^{-1} exists, the asymptotic resolvent $R(\infty,z)$ exists and is computable in closed form as $R(0,z)P_{rz}$, using M_z^{-1}.

2.4.4 Analyticity of $R(t,z)$ around 0 and ∞ for $z \in re(A)\backslash F(A,E)$

We assume in this section that M_z is invertible, i.e. $z \in re(A)\backslash F(A,E)$. Proposition 2.4.4 shows the dual role played by the two quantities $|t_{1z}| = \frac{1}{\max_{1\leq i\leq r}|\mu_{iz}|} = \frac{1}{\rho(M_z)} = \frac{1}{|\mu_{1z}|}$ and $|t_{rz}| = \frac{1}{\min_{1\leq i\leq r}|\mu_{iz}|} = \rho(M_z^{-1}) = \frac{1}{|\mu_{rz}|}$ defined in (2.4.5) and (2.4.6) when M_z^{-1} exists. More precisely, one can say for $z \in re(A)\backslash F(A,E)$ the following two analytic developments hold for $R(t,z)$.

(i) $|t_{1z}|$ defines the largest analyticity disk for $R(t,z)$. It rules the convergence of the initial analytic development for $|t|$ around 0

$$R(t,z) = R(0,z)[I_n - tU\sum_{k=0}^{\infty}(tM_z)^k V^H R(0,z)] \tag{2.4.7}$$

based on M_z and valid for $|t|<|t_{1z}|$ (around $t=0$).

The series expansion (2.4.7) becomes *finite* when M_z is *nilpotent* ($\rho(M_z)=0$), that is when z is critical, if this happens.

(ii) $|t_{rz}|$ defines the smallest value for $|t|$ beyond which $R(t,z)$ is analytic in $s=\frac{1}{t}$. This is analyticity in t around ∞. It rules the convergence of the asymptotic analytic development in $s=1/t$

$$R(t,z)=R(0,z)[I_n+UM_z^{-1}\sum_{k=0}^{\infty}(sM_z^{-1})^kV^HR(0,z)] \tag{2.4.8}$$

$$=R(\infty,z)+R(0,z)UM_z^{-1}\sum_{k=1}^{\infty}(tM_z)^{-k}V^HR(0,z),$$

based on M_z^{-1} and valid for $|t|>|t_{rz}|$, (around $|t|=\infty$, that is $s=0$).

Observe that M_z^{-1} cannot be nilpotent (because it is invertible).

2.5 $\lim_{|t|\to\infty}\sigma(A(t))$ for a singular deviation E with 0 as a semi-simple eigenvalue

2.5.1 The assumption (Σ)

When E is singular, it is possible that certain eigenvalues $\lambda(t)$ in $\sigma(A(t))$ do not diverge to ∞, as $|t|\to\infty$ as is the rule when E is regular. We call Lim the set of such limit points at finite distance in $\mathbb{C}$:

$$\lim_{|t|\to\infty}\sigma(A(t))=\{\text{Lim},\infty\}. \tag{2.5.1}$$

We denote the cardinality of Lim by $l_*=\text{card Lim}$ (counting multiplicities). Clearly $n\geq l_*\geq 0$ and $l_*=0$ when $\text{Lim}=\emptyset$. The points in Lim which are *not* eigenvalues are called limit points. The others are limit eigenvalues.

Notation 2.5.1 *The assumption* $0\in\sigma(E)$ *is semi-simple is denoted by* (Σ), *i.e.* $0\in\sigma(E)$ *semi-simple* $\iff$ $(\Sigma)\iff\det(V^HU)\neq 0$.

We assume that (Σ) holds throughout the rest of the chapter. We denote the *geometric(=algebraic)* multiplicity of $0\in\sigma(E)$ by $g=n-r$ for $n=\text{order}\,A$ and $r=\text{rank}\,E$. Chapter 3 will be devoted to the general case where $0\in\sigma(E)$ can be defective. It will require a more in-depth analysis.

Lemma 2.5.2 *[18]* $\mathbb{C}^n=\text{Ker}\ E\oplus\text{Im}\ E\iff\text{rank}\ G=\text{rank}\ (V^HU)=r\iff(\Sigma)$.

Remark. When we have both $\mathrm{Ker}\, E \cap \mathrm{Im}\, E = \{0\}$ and $(\mathrm{Ker} E)^{\perp} = \mathrm{Im} E = \mathrm{Im} E^H$, the direct sum of lemma 2.5.2 becomes orthogonal. Also in this case, the bases U and V in $\mathrm{Im}\, E$ satisfay $U = VB$ for $B \in \mathbb{C}^{r\times r}$ of rank r. It is possible to choose $B = I_r$, then $E = UU^H$ is Hermitian, semi positive definite: (Σ) is satisfied.

Since we assume that (Σ) holds, the invariant subspace for E associated with 0 is the eigenspace $\mathrm{Ker}\, E$. The associated eigenprojection is $P = I - UG^{-1}V^H$ which projects onto $\mathrm{Ker}\, E$ along $\mathrm{Im}\, E$.

2.5.2 Characterization of Lim

Proposition 2.5.3 *[16, 30] Under the assumption* (Σ)*, there exist* g *eigenvalues* $\lambda(t)$ *such that* $\lim_{|t|\to\infty} \lambda(t) = \xi$*, with* $\xi \in \sigma(\Pi)$ *where* Π *is the* $g \times g$ *matrix representing* PAP *restricted to* $\mathrm{Ker}\, E$*. Therefore* $\mathrm{Lim} = \sigma(\Pi)$*, and* $l_* = n - r = g$*.*

Proof. We recall that $\lambda(t) = \frac{\nu(s)}{s}$, where $\nu(s) \in \sigma(E + sA)$. By assumption, $\nu = 0$ is semi-simple. Therefore the series expansion $\nu(s) = \xi s + \mathrm{O}(s^{\alpha})$, $\alpha > 1$ is valid around $\nu = 0$ of multiplicity g for s small enough as is classical [32], [33], [42]. □

Among the n eigenvalues of $A(t)$, r escape to infinity, while $g = n - r$ remain at finite distance as $|t| \to \infty$. That is $1 \le l_* = n - r = g \le n - 1$. Their limits are the g Ritz values for A, associated with the projection P on $\mathrm{Ker}\, E$.

2.5.3 Limit points and their relationship with $F(A, E)$ and $C(A, E)$ under (Σ)

Observing the evolution $t \mapsto \lambda(t)$ leads to the distinction for the eigenvalues in $\sigma(A) = \sigma^i \cup \sigma^e$ between *invariant* eigenvalues $\lambda \in \sigma^i$ ($\lambda(t) = \lambda$ for any $t \in \mathbb{C}$) and *evolving* eigenvalues $\lambda \in \sigma^e$ ($\lambda(t) \neq \lambda$ for almost all $t \in \mathbb{C}$). Clearly, $\mathrm{Lim} = \sigma^i \cup \mathrm{Lim}^{\,e}$, where $\mathrm{Lim}^{\,e}$ is the set of limits at finite distance of evolving eigenvalues.

Since $\mathrm{Lim} = \sigma(\Pi)$, it is possible that $\sigma(\Pi) \cap \sigma(A) \neq \emptyset$. When this happens, $\lambda \in \sigma(\Pi) \cap \sigma(A)$ is an eigenvalue of A as well as a limit of $\lambda(t)$ as $|t| \to \infty$: it is a *limit eigenvalue*.

Definition 2.5.4 *The limit points of* (A, E) *are defined by* $\Lambda(A, E) = \mathrm{Lim} \cap re(A)$.

We have proved that $\Lambda(A, E) = \sigma(\Pi) \cap re(A)$: the limit points are the (at most $g = n - r$) *kernel* points in $\sigma(\Pi) \cap re(A)$, (see the general definition 3.9.8 in Chapter 3).

Under (Σ), the matrix E cannot be nilpotent because 0 is semi-simple and $E \neq 0$. $\mathrm{card}\, C(A, E) \le \mathrm{card}\ \mathrm{Lim} = g \Longrightarrow C(A, E)$ is discrete. Now we show that under (Σ), $F(A, E)$ is necessarily discrete containing a finite number of points in $re(A)$.

Proposition 2.5.5 *[18] Under* (Σ), $F(A,E)$ *is necessarily discrete and*

$$0 \le \text{card } F(A,E) \le (n-1)r, \quad \textit{and} \quad 0 \le \text{card } C(A,E) \le n-r.$$

Proof. Let $\pi(z) = \det(zI - A)$ denote the characteristic polynomial for the matrix A. One can formally write

$$M_z = \frac{1}{\pi(z)} V^H \text{adj}(zI - A)U = \frac{1}{\pi(z)} Q(z) \tag{2.5.2}$$

where $Q(z) = V^H \text{adj}\,(zI - A)U$ is a matrix polynomial of order r and degree $\le n-1$, defined for $z \in \mathbb{C}$. The matrix coefficient for z^{n-1} is $G = V^H U$ which is regular when $0 \in \sigma(E)$ is semi-simple (assumption (Σ)): $Q(z)$ has degree $n-1$. For z in $re(A)$, the values z for which at least one $\mu_z \in \sigma(M_z)$ is zero are the roots of $\det Q(z)$ [5, 18]. This is a scalar polynomial equation of degree $(n-1)r$ under $(\Sigma) \Leftrightarrow V^H U$ has rank r. This means that $\det Q(z)$ has at most $(n-1)r$ roots in $re(A)$, which are the elements of $F(A,E)$.

We have seen that for $r = n$, $F(A,E) = \emptyset$ since $0 \notin \sigma(F_z) = \sigma(M_z)$. When $r < n$, $F(A,E) = \emptyset$ iff $0 \in \sigma(F_z)$ is semi-simple, which is equivalent to $0 \notin \sigma(M_z)$, for any $z \in re(A)$. From the proposition 2.5.3, $\text{Lim} = \sigma(\Pi)$. And from the proposition 2.5.6 below, one has $\text{card}\, C(A,E) \le \text{card}\, \Lambda(A,E) \le g = n - r$.

□

This result will be strengthened in Chapter 3 (Proposition 3.5.1). We know that the 2 sets $F(A,E)$ and $C(A,E)$ are discrete under (Σ). We now prove the following more general property, that we shall need in Chapter 3.

Proposition 2.5.6 *In general, with no assumption on* $0 \in \sigma(E)$, *the following implication holds*

$$\text{card}\, F(A,E) < \infty \implies \Lambda(A,E) \subseteq F(A,E).$$

Proof. We suppose that $F(A,E)$ is discrete. According to the lemma 2.3.1, $z' \in re(A)$ belongs to $\sigma(A(t)) \Leftrightarrow$ there exists an eigenvalue $\mu_{z'} \neq 0$ of M_z such that $t = 1/\mu_{z'} \in \mathbb{C}$. Now, $z' \in \Lambda(A,E) \Leftrightarrow |t|$ is unbounded $\Leftrightarrow \lim_{z\to z'} \mu_z = 0$. By continuity of μ_z as $z \to z' \in re(A)$, one has $\mu_{z'} = 0$ and $z' \in F(A,E)$. This shows that $\Lambda(A,E) \subseteq F(A,E)$.

We observe that in general $C(A,E) \subseteq F(A,E)$. When $r = 1$, we have $C(A,E) = F(A,E)$.

□

The proposition 2.5.6 states that if ξ is in $\Lambda(A,E)$ then $\xi \in F(A,E)$. This means that ξ is an *inexact* eigenvalue of A at "infinite" homotopic distance $|t| = \infty$. This is one more reason to consider $t \in \hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$,

as advocated by M. B. van Gijzen. This Proposition is valid in full generality: 0 need not be semi-simple in $\sigma(E)$. What follows assumes again that (Σ) holds.

Under (Σ), $\mathrm{Lim} = \sigma(\Pi) \neq \emptyset$. Can we have $\mathrm{card}\, F(A,E) = \emptyset$? This would imply $\Lambda(A,E) = \emptyset \Longleftrightarrow \sigma(\Pi) \subset \sigma(A)$. An example is given below for $r = 1$.

2.5.4 The critical and frontier sets for $r = 1$

When $r = 1$, M_z is the scalar $\mu_z = v^H(zI - A)^{-1}u$ and the frontier points are critical. Moreover $R(t,z)$ is a polynomial in t of degree 1 for $z \in F(A,E) = C(A,E)$: $R(t,z) = R(0,z)[I_n + tER(0,z)]$ when $\mu_z = 0$. In this case, one has

$$C(A,E) = \Lambda(A,E) = \sigma(\Pi) \cap re(A) = F(A,E). \tag{2.5.3}$$

There is a possibility that $C(A,E) = F(A,E) = \emptyset$ as shown below.

Example 2.5.1 *Let consider $A = \lambda I$. This yields $M_z = \frac{1}{z-\lambda} V^H U = \frac{1}{z-\lambda} G$ for $z \in re(A)$. When $0 \in \sigma(E)$ is semi-simple, M_z is regular for any $z \neq \lambda$. Therefore $C(A,E) = F(A,E) = \emptyset$ and one has $\emptyset \neq \sigma(\Pi) \subset \sigma(A) = \{\lambda\}$. More precisely, $\sigma(\Pi) = \{(\lambda^1)^g\}$ and $\sigma(A) = \{(\lambda^1)^n\}$.*

$\triangle$

For $r = 1$, $l_* = \mathrm{card}\ \mathrm{Lim}$ takes its *maximum* value $l_* = g = n - 1$, and

$$0 \leq \mathrm{card}\ F(A,E) = \mathrm{card}\ C(A,E) \leq n - 1. \tag{2.5.4}$$

We present below an important application of the case $r = 1$ (that is, when the matrix deviation E has rank $r = 1$). It is concerned with problems known with **uncertainty on the data**.

2.6 An example of rank $1-$deviation: the normwise backward analysis for $Ax = b$

We review the fundamental example of a rank $1-$deviation matrix provided by the *normwise backward analysis* for the linear system $Ax = b$. Let be given any $\tilde{x} \in \mathbb{C}^n$. The associated residual, $r = b - A\tilde{x}$, is assumed to be nonzero. There is an infinity of matrices ΔA such that $(A + \Delta A)\tilde{x} = b$. Among them, the rank $1-$ matrix E below is special. For simplicity, we assume that the matrix norm is induced by $\|\cdot\|_2$. (see [23]).

Let $E = \frac{1}{\tilde{x}^H\tilde{x}} r\tilde{x}^H$, then $E\tilde{x} = r$ and $(A + E)\tilde{x} = A\tilde{x} + b - A\tilde{x} = b$: $\tilde{x}$ is the exact solution for $(A + E)x = b$. Then

$$\|E\| = \max_{x \neq 0} \|\frac{r}{\tilde{x}^H\tilde{x}}\tilde{x}^H x\| = \frac{\|r\|}{\|\tilde{x}\|} \max_{x \neq 0} \frac{|\tilde{x}^H x|}{\|\tilde{x}\|\|x\|} = \frac{\|r\|}{\|\tilde{x}\|},$$

which is the normalized residual norm at $\tilde{x}$. One has the following result for A, b, and $\tilde{x}$ of fundamental practical importance.

Proposition 2.6.1 *Given any $\tilde{x} \in \mathbb{C}^n$. Then for the associated residual $r = b - A\tilde{x}$, the rank $1-$ matrix $E = \frac{1}{\tilde{x}^H\tilde{x}} r\tilde{x}^H$ realizes*

$$\frac{\|r\|}{\|\tilde{x}\|} = \min\{\|\Delta A\|, \text{ for } \Delta A \in \mathbb{C}^{n\times n} \text{ such that } (A + \Delta A)\tilde{x} = b\}.$$

Proof. Let $(A + \Delta A)\tilde{x} = b$ for some $\Delta A \in \mathbb{C}^{n\times n}$. This means that $A\tilde{x} + \Delta A\tilde{x} = (b - r) + \Delta A\tilde{x} = b$ which evidently yields the equality

$$\Delta A\ \tilde{x} = r. \tag{2.6.1}$$

Using (2.6.1) one has $\|\Delta A\|\|\tilde{x}\| \geq \|r\|$ or $\|\Delta A\| \geq \frac{\|r\|}{\|\tilde{x}\|}$. This shows that $\forall\ \Delta A$, $\|\Delta A\| \geq \|E\|$. Therefore

$$\min\ \|\Delta A\| = \frac{\|r\|}{\|\tilde{x}\|}$$

for $\Delta A \in \mathbb{C}^{n\times n}$ such that $(A + \Delta A)\tilde{x} = b$ is achieved by $E = \frac{1}{\tilde{x}^H\tilde{x}} r\tilde{x}^H$.

□

The quantity $\|E\| = \frac{\|r\|}{\|\tilde{x}\|}$ is the *normwise backward error* associated with $\tilde{x}$ and A, b: it gives the minimum size perturbation ΔA such that $\tilde{x}$ is an exact solution for $(A + \Delta A)\tilde{x} = b$. The matrix $A + E$ is a rank $1-$ modification of A.

The relative version is

$$\min\{\frac{\|\Delta A\|}{\|A\|}, \text{ for } \Delta A \in \mathbb{C}^{n\times n} \text{ such that } (A + \Delta A)\tilde{x} = b\} = \frac{\|r\|}{\|A\|\|\tilde{x}\|}. \tag{2.6.2}$$

This is the version used to assess results computed in finite precision: optimally, the backward error should be of the order of machine precision ($\sim 10^{-16}$). However, if A is known only with limited accuracy α, then the backward error should be $\leq \alpha$. Any $\tilde{x}$ for which $\frac{\|A\tilde{x}-b\|}{\|\tilde{x}\|} < \alpha$ can be accepted as a solution of $Ax = b$ with the level α of uncertainty on the data in A.

In this *normwise* backward analysis, the structure of ΔA is arbitrary in $\mathbb{C}^{n\times n}$ (n^2 parameters). Only its norm matters. We shall present, in chapter 5, the *homotopic backward analysis*, where $\Delta A = tE$ has a *fixed* structure E, and only t varies in $\mathbb{C}$ (1 parameter).

However, the application of the ideas above is much broader than finite precision computation. It concerns all *Experimental Sciences* where data are only known with uncertainty. Indeed, the specificity of Experimental Sciences (as opposed to Mathematics) is that the data are collected from measurements hence they have no absolute certainty. Therefore a kind of normwise backward analysis is required to assess the validity of modeling results against Nature's results. This is an important aspect of the engineering computation activity.

Chapter 3

Theory of Homotopic Deviation, II

3.1 Introduction

In chapter 2, the essential of Homotopic Deviation theory was presented under the simplifying assumption that $0 \in \sigma(E)$ is semi-simple. In this chapter, when we assume that $0 \in \sigma(E)$ is defective, this is symbolically denoted as (Δ).

We look at the 2 following questions successively when (Σ) does not necessarily hold. What can be said about the sets

(i) $C(A,E)$, $\Lambda(A,E)$, $F(A,E)$ in $re(A)$?

(ii) Lim in $\mathbb{C}$?

Answers to the first question are based on an analysis of the algebraic structure of $z \mapsto M_z$, $z \in re(A)$ which is more thorough than what was needed in Chapter 2 under (Σ).

3.2 The algebraic structure of $M_z,\ z \in re(A)$

The analysis of chapter 2 has already shown that the $r \times r$ matrix M_z plays a key role in HD because of the relation $t\mu_z = 1$. The role is computationally important when r is small compared to n: the necessary information for HD lies in $\sigma(M_z)$. Under the simplifying assumption (Σ) we have been able to develop the theory with a very limited knowledge about $M_z = V^H(zI - A)^{-1}U$. To venture into the general case, where (Σ) does not necessarily hold, a deeper knowledge of M_z is required.

During his post-doctoral visit at Cerfacs (March 2004 to February 2005), Prof. F. S. V. Bazán from the Universidade Federal de Santa Catarina, Florianópolis, Brazil, directed my attention to the rich algebraic structure of M_z [5, 12]. I acknowledge many illuminating discussions on HD with Prof. Bazán during the Summer and Fall of 2004.

3.2.1 M_z as a particular transfer function in Linear System theory

Consider the linear dynamical system

$$\mathcal{S}: \quad \begin{cases} \dot{x} = Ax + Bu \\ y = Cx \end{cases}$$

where $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{n\times r}$, $C \in \mathbb{C}^{r\times n}$ are invariant matrices, $x \in \mathbb{C}^n$, $u, y, \in \mathbb{C}^r$ are vector functions of the time. This state-space description can be rewritten in the frequency domain $\mathbb{C}$ as

$$\begin{aligned} y(s) &= C(sI - A)^{-1}Bu(s), \quad s \notin \sigma(A) \\ &= H(s)u(s) \end{aligned}$$

The transfer matrix $H(s) \in \mathbb{C}^{r\times r}$ identifies with M_z for $s = z$ and $C = V^H$, $B = U$ [43, 35].

3.2.2 A Schur complement interpretation

We recall that for a singular matrix E with rank $r < n$, the $n \times n$ matrix F_z in (2.3.2) has rank r, so at most r eigenvalues μ_{iz}, $i = 1, \cdots, r$ are nonzero. These are the r eigenvalues of the $r \times r$ matrix M_z. Here, we use the following Schur complement interpretations to derive important properties of the $r \times r$ matrix M_z which should be considered instead of the $n \times n$ matrix F_z [5, 4].

Set $N = n + r$. One has the following equivalence formulae in $\mathbb{C}^{N\times N}$, for z given in $re(A)$, and $t \in \mathbb{C}$ [5]:

$$\left[\begin{array}{c|c} I_n & 0 \\ \hline -V^H R & I_r \end{array}\right] \left[\begin{array}{c|c} A - zI_n & -tU \\ \hline V^H & I_r \end{array}\right] \left[\begin{array}{c|c} I_n & tRU \\ \hline 0 & I_r \end{array}\right] = \left[\begin{array}{c|c} A - zI_n & 0 \\ \hline 0 & I_r - tM_z \end{array}\right] \tag{3.2.1}$$

where $R = (A - zI_n)^{-1}$ for $z \notin \sigma(A)$

$$\left[\begin{array}{c|c} I_r & 0 \\ \hline -tU & I_n \end{array}\right] \left[\begin{array}{c|c} I_r & 0 \\ \hline 0 & A - zI_n + tE \end{array}\right] \left[\begin{array}{c|c} I_r & V^H \\ \hline 0 & I_n \end{array}\right] = \left[\begin{array}{c|c} I_r & V^H \\ \hline -tU & A - zI_n \end{array}\right]. \tag{3.2.2}$$

The rank of the augmented matrix, for $z \in re(A)$, remains constant equal to $N = n + r$ as long as rank $(I_r - tM_z) = r$ (resp. rank $(A - zI + tE) = n$) in case (3.2.1) (resp. case (3.2.2)).

Lemma 3.2.1 *[5] Let $A \in \mathbb{C}^{n\times n}$ and $z \in re(A)$, then*

$$\det(A + tE - zI_n) = \det(A - zI_n)\ \det(I_r - tM_z). \tag{3.2.3}$$

Proof. For $1 \leq r \leq n$, $n+1 \leq N \leq 2n$, the augmented matrix

$$\hat{B}(z,t) = \left[\begin{array}{c|c} A - zI_n & -tU \\ \hline V^H & I_r \end{array}\right]$$

has the equivalent block diagonal forms (3.2.1) and (3.2.2). Their determinants are equal to $\det(\hat{B}(z,t)) = \det(A + tE - zI)$.

1. First, $\left[\begin{array}{c|c} A - zI_n & 0 \\ \hline 0 & I_r - tM_z \end{array}\right]$,
   where its second diagonal block has rank r as long as $t\mu_z \neq 1$ for $\mu_z \in \sigma(M_z)$.
2. Second, $\left[\begin{array}{c|c} A - zI_n + tE & 0 \\ \hline 0 & I_r \end{array}\right]$,
   where its first diagonal block has rank n as long as z is not an eigenvalue of $A(t)$.

Thus for $z \notin \sigma(A)$, the equality (3.2.3) is valid.

□

This lemma shows the role of M_z of order $r \leq n$ in the study of $R(t,z)$ of order n.

3.2.3 M_z as a matrix rational function of z in $re(A)$

Let $A, E \in \mathbb{C}^{n\times n}$ with $\operatorname{rank} E = r < n$. $\pi(z) = \det(zI - A)$ denotes the characteristic polynomial for the matrix A. One can formally write

$$M_z = \frac{1}{\pi(z)} V^H \operatorname{adj}(zI - A) U = \frac{1}{\pi(z)} Q(z) \tag{3.2.4}$$

where $Q(z) = V^H \operatorname{adj}(zI - A)U$ is a matrix polynomial of order r and degree $\leq n-1$, defined for $z \in \mathbb{C}$. The matrix coefficient for z^{n-1} is $G = V^H U$ which is regular when $0 \in \sigma(E)$ is semi-simple (assumption (Σ)). For z in $re(A)$, the values z for which at least one $\mu_z \in \sigma(M_z)$ is zero are the roots of $\det Q(z)$ [5, 18]. This is a scalar polynomial equation of degree $(n-1)r$ under $(\Sigma) \Leftrightarrow V^H U$ has rank r. This means that $\det Q(z)$, when $\not\equiv 0$, has at most $(n-1)r$ roots in $re(A)$, which are the elements of $F(A,E)$.

3.3 Singularities of M_z for z in $re(A)$

The *frontier points* in $F(A,E)$ have been defined in chapter 2, (see [18, 21]) as the points $z \in re(A)$ such that $0 \in \sigma(M_z)$: at such points, $R(\infty, z) = \lim_{|t|\to\infty} R(t,z)$ does not exist. How are these points defined in terms of $Q(z)$?

3.3.1 Eigenvalues μ_z of M_z, $z \in re(A)$

We define $d_i(z)$, $i = 1, \ldots, r$, as the r functions of $z \in \mathbb{C}$ which satisfy

$$\det(Q(z) - d(z)I_r) = 0, \tag{3.3.1}$$

for any arbitrary but fixed z in $\mathbb{C}$.
The r eigenvalues μ_{iz}, $i = 1, \ldots, r$, of M_z are defined for $z \in re(A)$ by

$$\mu_{iz} = \frac{d_i(z)}{\pi(z)}. \tag{3.3.2}$$

For $z \in re(A)$, $\mu_z = 0$ iff $d(z) = 0$ since $\pi(z) \neq 0$. Therefore, $\det M_z = 0$ iff $\det Q(z) = 0$ for $z \in re(A)$.

3.3.2 Zeroes of $\det Q(z)$ in $\mathbb{C}$ and in $re(A)$: a preliminary study

For $z \in \mathbb{C}$, the set of zeroes of $\det Q(z)$ is defined as

$$Zer = Zer(\det Q) = \{z \in \mathbb{C} : \ \det Q(z) = 0\} = \{z \in \mathbb{C} : \ \exists i \text{ such that } d_i(z) = 0\}.$$

This means that $Zer = Zer(\det Q) = \bigcup_{i=1}^{r} Zer(d_i(z))$ where

$$Zer(d_i(z)) = \{z \in \mathbb{C} : d_i(z) = 0\},$$

for $i = 1, \cdots, r$, which is equivalent to write,

$$\det Q(z) = d_1(z)d_2(z)\cdots d_r(z). \tag{3.3.3}$$

Also we have $\operatorname{card} Zer = \sum_{i=1}^{r} \operatorname{card} Zer(d_i(z))$, where the roots in $Zer(d_i(z))$ are counted with their algebraic multiplicities, $i = 1, \ldots, r$.

When $\det Q(z) \not\equiv 0$, $\det Q(z)$ is a scalar polynomial of degree $\leq (n-1)r$. When $G = V^H U$ has rank r, the $(n-1)r$ points in Zer can be interpreted as the $(n-1)r$ eigenvalues of the block companion matrix of order $(n-1)r$

$$\begin{bmatrix} 0 & & & -C_0 \\ I_r & \ddots & & \vdots \\ & \ddots & 0 & -C_{n-3} \\ 0 & & I_r & -C_{n-2} \end{bmatrix},$$

where $Q(z) = V^H \operatorname{adj}(zI - A)U = V^H U\,(I_r z^{n-1} + C_{n-2}z^{n-2} + \cdots + C_0)$ [43]. These $(n-1)r$ eigenvalues will be derived from Proposition 3.4.1 below.

When $0 \in \sigma(E)$ is defective, $V^H U$ is singular and the degree of $\det Q(z)$ is $< (n-1)r$.

Lemma 3.3.1 *The singularities of* M_z *in* $re(A)$ *are the frontier points in the discrete set* $F(A,E) = Zer \cap re(A)$ *when* $\det Q(z) \not\equiv 0$.

Proof. The singularities of M_z are the roots of the equation $\det M_z = 0$ where

$$\det M_z = \frac{1}{\pi(z)^r} \det Q(z) = 0. \tag{3.3.4}$$

Now, the relations (3.3.3) and (3.3.2) show that when $\det Q(z) \not\equiv 0$, $F(A,E) = Zer \cap re(A)$.

□

When $\det Q(z) \equiv 0$, $Zer = \mathbb{C}$ and $F(A,E) = re(A)$ is continuous.

3.4 The factorization of det $Q(z)$

Considering the importance of frontier points for localizing the limit points, in $\Lambda(A,E)$, the following factorization of $\det Q(z)$ provides a remarkable simplification of the theory. All what we should seek is a scalar polynomial in z of degree $\leq n-r$ which easily derived from the data A, U and V .

Let us consider the following augmented matrix of order $n+r$

$$\hat{A}(z) = \left[\begin{array}{c|c} zI_n - A & -U \\ \hline V^H & 0 \end{array}\right]. \tag{3.4.1}$$

It is interesting to contrast $\hat{A}(z)$ with $\hat{B}(z,t)$ already introduced. Via (1.6.3), the Schur complement of $zI_n - A$ in the augmented matrix $\hat{A}(z)$ for $z \in re(A)$ is

$$0 - V^H(zI - A)^{-1}(-U) = M_z. \tag{3.4.2}$$

Using Laplace formula we see that $\det \hat{A}(z) = \hat{\pi}(z)$ is a polynomial in z of degree $\hat{d} \leq n - r = g$, $\hat{d} = n - r$ iff $V^H U$ is invertible.

Using the relation (1.6.4), one has

$$\hat{\pi}(z) = \det \hat{A}(z) = \pi(z) \det M_z \tag{3.4.3}$$

for $z \in re(A)$.

Proposition 3.4.1 *[20] For* $r \geq 1$,

$$\det Q(z) = (\pi(z))^{r-1} \hat{\pi}(z), \tag{3.4.4}$$

where $\hat{\pi}(z) = \det \hat{A}(z)$ *is the scalar polynomial in* z *of degree* $\leq n - r$ *satisfying (3.4.3). Hence for* $z \in re(A)$, $\det M_z = \frac{\hat{\pi}(z)}{\pi(z)} = \Pi_{i=1}^r \mu_{iz}$.

Proof. Using the relation (3.2.4), we get

$$\det M_z = \frac{1}{\pi(z)^r}\ \det Q(z). \tag{3.4.5}$$

This, together with (3.4.3), shows that $\det Q(z) = (\pi(z))^{r-1}\hat{\pi}(z)$ for $z \in re(A)$. This means that for every z in the open set $re(A)$, we have an identity between two polynomials in z. Therefore, the identity is true for any $z \in \mathbb{C}$. Observe that $\det Q \equiv 0 \Leftrightarrow \hat{\pi}(z) \equiv 0$.

□

The coefficient of $z^{(n-1)r}$ in $\det Q(z)$ is $\det(V^H U)$; this is the coefficient of z^{n-r} in $\hat{\pi}(z)$. We know that

$$\deg \det Q(z) = (n-1)r \Longleftrightarrow (\Sigma). \tag{3.4.6}$$

Therefore

$$\deg \hat{\pi}(z) = (n-1)r - n(r-1) = n - r \Longleftrightarrow (\Sigma).$$

When (Σ) does not hold, $\deg \det Q(z) < (n-1)r \Leftrightarrow \deg \hat{\pi}(z) < n-r$.

For $r = 1$, $\det Q(z) = \hat{\pi}(z)$ has degree $\leq n-1$ (easy to check by direct computation of $v^H \operatorname{adj}(zI - A)u$).

3.5 Analysis of $F(A,E)$, $\Lambda(A,E)$, $C(A,E)$ in $re(A)$

In this section, we make no assumption on $0 \in \sigma(E)$. We analyse the frontier points in $F(A,E)$, the critical points in $C(A,E)$, the limit points in $\Lambda(A,E)$ and the relationships between these 3 sets in $\mathbb{C}$.

3.5.1 The zeroes of $\det Q(z)$

Using the proposition 3.4.1, one has

$$Zer = Zer(\det Q(z)) = (\sigma(A))^{r-1} \cup \hat{Z} \tag{3.5.1}$$

where $\hat{Z} = Zer(\hat{\pi}(z)) = \{z \in \mathbb{C} : \hat{\pi}(z) = 0\}$. And $Zer = \hat{Z} \subset \mathbb{C} \Leftrightarrow r = 1$.

According to Proposition 3.4.1, one has the remarkably simple characterization for the frontier set

$$F(A,E) = re(A) \cap \hat{Z} = re(A) \cap Zer. \tag{3.5.2}$$

The set Zer is replaced by $\hat{Z} \subset Zer$ to get the frontier set $F(A,E)$ for $r > 1$.

Under (Δ) and when $\hat{\pi}(z) \not\equiv 0$, $\operatorname{card} \hat{Z} < n - r$, but when $\hat{\pi}(z) \equiv 0$, $Zer = \hat{Z} = \mathbb{C}$.

3.5.2 $F(A,E)$ is discrete when $\hat{\pi}(z) \not\equiv 0$

The proposition 3.4.1 shows that when $\hat{\pi}(z) \not\equiv 0$, then $F(A,E)$ is a discrete set which is derived from (3.5.2).

Proposition 3.5.1 *[18] When* $\hat{\pi}(z) \not\equiv 0$, *then* $\operatorname{card} C(A,E) \leq \operatorname{card} F(A,E) \leq n-r$, *and*

$$\{C(A,E),\ \Lambda(A,E)\} \subseteq F(A,E) \tag{3.5.3}$$

with equality $C(A,E) = F(A,E)$ *when* $r = 1$.

Proof. Using the relation (3.5.2), one has $F(A,E) \subseteq \hat{Z} = Zer(\hat{\pi}(z))$. Therefore, when $\hat{\pi}(z) \not\equiv 0$, $F(A,E)$ is discrete and

$$\operatorname{card} C(A,E) \leq \operatorname{card} F(A,E) \leq \operatorname{card} \hat{Z} \leq n-r.$$

The rest is a result of proposition 2.5.6.

□

Under (Δ), $\operatorname{card} F(A,E) < n-r$: there are at most $g-1 = n-r-1$ frontier points in $re(A)$ where $R(\infty,z)$ does not exist.

3.5.3 $Zer = \hat{Z} = \mathbb{C}$ when $\hat{\pi}(z) \equiv 0$

We suppose now that $\hat{\pi}(z) \equiv 0$. Therefore $F(A,E) = re(A)$ is continuous by (3.5.2). But the critical set $C(A,E)$ can be either discrete or continuous.

Theorem 3.5.2 *[18] When the critical set* $C(A,E)$ *is continuous, then* $F(A,E) = C(A,E) = re(A)$, *and* $\operatorname{Lim} = \sigma(A)$: $\sigma(A(t)) = \sigma(A)$ *for* $t \in \mathbb{C}$. *A necessary condition is that* E *is nilpotent.*

More precisely, we can show the following:

Theorem 3.5.3 *[5] When* $C(A,E)$ *contains at least* n *distinct points, then* $C(A,E) = re(A)$ *is continuous and* $\forall t \in \mathbb{C}$, $\sigma(A(t)) = \sigma(A)$.

Proof. For $z \in C(A,E)$, M_z is nilpotent or zero and

$$\det(A + tE - zI) = \det(A - zI) \neq 0,$$

for $t \in \mathbb{C}$. In other words,

$$\Pi_{i=1}^{n}(\lambda_i(t) - z) = \Pi_{i=1}^{n}(\lambda_i - z)$$

for $t \in \mathbb{C}$ with $\lambda_i(0) = \lambda_i$. This can be written as

$$z^n + \Sigma_{i=1}^{n} a_i(t) z^{n-i} = z^n + \Sigma_{i=1}^{n} a_i z^{n-i},$$

where $a_i(0) = a_i$. Therefore

$$\Sigma_{i=1}^{n}(a_i(t) - a_i)z^{n-i} = 0. \tag{3.5.4}$$

If we assume that $C(A,E)$ contains n *distinct* points z_j, $j = 1,\ldots,n$, then for t given in $\mathbb{C}$, (3.5.4) is a linear system in the unknowns $a_i(t) - a_i$, $i = 1,\ldots,n$, corresponding to the $n \times n$ Vandermonde matrix based on the n distinct points z_j. Therefore, the unique solution to (3.5.4) is $a_i(t) = a_i$ for $t \in \mathbb{C}$. Hence $\lambda_i(t) = \lambda_i$, $i = 1,\ldots,n$, for any t and $C(A,E) = re(A)$ ([18], Corollary 5.2). It follows that, when $C(A,E)$ is discrete, it contains at most $n-1$ distinct points. When $C(A,E)$ contains at least n distinct point, $C(A,E) = re(A)$ and $\sigma(A(t))$ is invariant under t: no evolution takes place for $\lambda(t) \equiv \lambda$, $t \in \mathbb{C}$.

□

Under the assumption of the Theorem 3.5.3, one has $\mathrm{Lim} = \sigma(A)$, $l_* = \mathrm{card}\,\mathrm{Lim} = n$ and $\Lambda(A,E) = \emptyset$ and there is no evolution for the eigenvalues in $\sigma(A(t))$.

When $\hat{\pi}(z) \equiv 0$, $F(A,E)$ is continuous and if $C(A,E)$ is finite and contains at most $n-1$ distinct points, then $C(A,E)$ is discrete and the eigenvalues of $\sigma(A(t))$ evolve. When $\hat{\pi}(z) \not\equiv 0$, the fact that $\mathrm{card}\,C(A,E) \leq n - r \leq n-1$ yields that the property $\forall t \in \mathbb{C}$, $\sigma(A(t)) = \sigma(A)$ is impossible: $\sigma(A(t))$ always evolve.

3.6 The case (Σ) revisited

Thanks to the remarkable factorization

$$\mathrm{det}Q(z) = (\pi(z))^{r-1}\hat{\pi}(z)$$

we are able to further the special study of HD which was presented in Chapter 2 under (Σ). We have seen that $\hat{\pi}(z)$ has exactly the degree $n - r$ under (Σ). Therefore Proposition 2.5.5 can be strengthened into the Proposition 3.5.1 above. Hence $\mathrm{card}\ C(A,E) \leq \mathrm{card}\ \Lambda(A,E) \leq \mathrm{card}\ F(A,E) \leq n - r$.

Lemma 3.6.1 *Under* (Σ)*, if* $\mathrm{card}\ \Lambda(A,E) = n-r = g$*, then* $\Lambda(A,E) = F(A,E)$*. Moreover*

$$\sigma(A) \cap \sigma(\Pi) = \sigma(A) \cap \hat{Z} = \emptyset$$

and

$$\mathrm{Lim} = \sigma(\Pi) = \hat{Z} = F(A,E) \subset re(A)$$

.

Proof. Direct consequence of the inclusion $\Lambda(A,E) \subseteq F(A,E)$ and $\deg\ \hat{\pi}(z) = n - r = g \Leftrightarrow (\Sigma)$.

□

Theorem 3.6.2 *Under the assumptions of Lemma 3.6.1, one has the identity*

$$\frac{1}{\det G}\hat{\pi}(z) = \frac{1}{\det G}\det\hat{A}(z) \equiv \det(zI_g - \Pi) \tag{3.6.1}$$

where $G = V^H U$.

Proof. The polynomials $\frac{1}{\det G}\hat{\pi}(z)$ and $\det(zI_g - \Pi)$ are monic polynomials with same degree $g = n - r$ and same roots. Therefore they are equal for all $z \in \mathbb{C}$.

□

Lemma 3.6.3 *Under* (Σ) , card $\Lambda(A, E) = g = n - r$ *iff* $\sigma(A) \cap \sigma(\Pi) = \emptyset$. *This implies that* $\sigma(A) \cap \hat{Z} = \emptyset$.

Proof. Clear since under (Σ) , $\Lambda(A, E) = \sigma(\Pi) \cap re(A)$.

□

The algebraic identity (3.6.1) is remarkable. It is valid under (Σ) iff A and $\Pi = PAP_{\restriction \mathrm{Ker}E}$ have distinct eigenvalues. It implies that no eigenvalue of A can be a frontier point. One observes that the assumption $0 \in \sigma(E)$ semi-simple is *essential* in the proof.

3.7 Analysis of Lim in $\mathbb{C}$ under (Δ)

Under the condition (Δ) , one has $0 \leq l_* \leq m \leq n$ where m denotes the algebraic multiplicity of $0 \in \sigma(E)$, $g = n - r < m$.

3.7.1 Comparison between $\Lambda(A, E)$ and Lim

By definition, Lim is the set of all finite limits of $\lim_{|t|\to\infty} \sigma(A(t))$. And $\Lambda(A, E) = \mathrm{Lim} \cap re(A)$ contains finite limits which belongs to $re(A)$: $\Lambda(A, E)$ as a subset of Lim which differs from Lim in the eigenvalues of A which are limit eigenvalues if they exist.

Let $d = l_* - \mathrm{card}\,\Lambda(A, E)$ be the number of limit eigenvalues. Then

(i) when $C(A, E)$ is discrete $\Longrightarrow$ $d \leq m \leq n$,

(ii) when $C(A, E) = re(A)$ $\Longrightarrow$ $d = n$.

We shall go back to the possibility $\lambda \in \mathrm{Lim}$ in Chapter 7 .

3.7.2 Double inclusion for Lim

In general when $\hat{\pi}(z) \not\equiv 0$ ($F(A,E)$ is discrete), one can determine two different sets L_1 and L_2 in $\mathbb{C}$ such that

$$L_1 \subseteq \text{ Lim } \subseteq_s L_2 \tag{3.7.1}$$

The right inclusion in (3.7.1) is *setwise* $\subseteq_s$.

Because $\Lambda(A,E) \subseteq F(A,E)$, $\text{Lim} \subseteq_s \Lambda(A,E) \cup \sigma(A) \subseteq_s \hat{Z} \cup \sigma(A) \subseteq Zer$ for $r \geq 2$. Therefore $L_2 = \hat{Z} \cup \sigma(A)$.

As it was shown in Proposition 3.4.1, the determination of $\hat{\pi}(z)$ (hence of L_2) is easy from the knowledge of the matrix $\hat{A}(z)$ of order $n+r$.

When $\hat{\pi}(z) \equiv 0$, then $L_2 = Zer = \hat{Z} = \mathbb{C}$. When $C(A,E)$ is also continuous then $\Lambda(A,E) = \emptyset$ and $L_1 = \text{Lim} = \sigma(A)$.

3.7.3 About the subset L_1

The left inclusion in (3.7.1) is *algebrawise* $\subseteq$. The algebraic determination of the set L_1 is much more difficult than that of L_2. The question is intimately connected with the theory of the convergence $\nu(s) \to 0$ defective, as $s \to 0$, for $\nu(s) \in \sigma(E(s))$.

We review this theory first. The main references are [48, 45, 11, 46].

3.8 A survey of perturbation theory for the eigenvalues of a matrix in Jordan form

Let us rewrite

$$A(t) = A + tE = t(E + sA) = tE(s), \quad s = \frac{1}{t},$$

for $t \neq 0$. Therefore

$$\lambda(t) \to \xi \in \text{ Lim} \quad \text{as} \quad |t| \to \infty, \quad \lambda(t) \in \sigma(A(t)) \iff$$

$$\nu(s) = \xi s + o(s) \quad \text{as} \quad s \to 0, \quad \nu(s) \in \sigma(E(s)).$$

This clearly requires that $\nu(s) \to 0$ as $s \to 0$, hence $\nu(0) = 0 \in \sigma(E)$. Any eigenvalue $\lambda(t)$ which does not escape to ∞, has as its limit, the coefficient ξ (0 or not) for s in the asymptotic expansion of $\nu(s)$ in s: Lim is nonempty iff at least one eigenvalue $\nu(s)$ converges to 0 with order ≥ 1 in s, and

$$\xi = \lim_{s \to 0} \frac{\nu(s) - \nu(0)}{s} = \nu'(0).$$

The problem of finding $\lim_{|t|\to\infty}\lambda(t)$ can be treated as a special instance of Lidskii's theory applied to $E(s)=E+sA$ as $|s|\to 0$ when there exists at least one trivial Jordan block for $0\in\sigma(E)$.

More precisely, the order of convergence in s is 1 (resp. >1) if $\xi\neq 0$ (resp. $=0$).

3.8.1 Definitions

Let $\nu(s)$ is an eigenvalue of

$$E(s)=E+sA=J+sA=J(s), \tag{3.8.1}$$

where $E=J$ is under Jordan form, A is a given matrix, J, A in $\mathbb{C}^{n\times n}$, and s, the complex perturbation parameter, tends to 0. As $s\to 0$,

$$\nu(s)\to\nu\in\sigma(J)$$

with algebraic (resp. geometric) multiplicity m (resp. g). Without loss of generality, we assume that $\nu=0$. It is supposed to be *defective* ($1\leq g<m$). Let n_j, $j=1,\ldots,q$, $q\geq 1$ be the *different* sizes of the Jordan blocks for $0\in\sigma(J)$ ordered by *decreasing* value

$$n_1>n_2>\cdots>n_q.$$

We remark that if $n_q=1$, then $q\geq 2$ under (Δ). Each block of size n_j is repeated r_j times. The structure of $0\in\sigma(E)=\sigma(J)$ is therefore $(0^{n_1})^{r_1}(0^{n_2})^{r_2}\ldots(0^{n_q})^{r_q}$.

Let us define

$$f_i=\sum_{j=1}^{i}r_j$$

with $f_q=g$ and

$$m_i=\sum_{j=1}^{i}r_jn_j$$

with $m_q=m$.

We want to determine, when possible, for each of the m eigenvalues $\nu(s)$ converging to $\nu=0$, its order of convergence, p, and its nonzero leading coefficient $\xi\neq 0$:

$$\nu(s)=\xi s^p+o(s^p),\quad p\in\mathbb{Q},\quad 0\neq\xi\in\mathbb{C},$$

which is the first order term in the asymptotic expansion for $\nu(s)$. It is known since Puiseux in the 19th century [48, 42] that, when there is no interaction between the Jordan blocks, the possible exponents are the q rational numbers $1/n_i$, $i=1,\ldots,q$. These exponents are called *Puiseux exponents*, they are *generic* [41]. But no attention was given then to the coefficients ξ.

In a fundamental work published in 1965 [45], Lidskii looks at the coefficients ξ. He proposes an algorithm to compute them under the non interaction assumption of Puiseux.

3.8.2 The generic Lidskii process

The Lidskii algorithm constructs a sequence of imbedded matrices Γ_j of increasing order f_j, starting from Γ_1 of order $f_1 > 0$, and setting $f_0 = 0$. In a 2×2 block representation, this gives, for $j = 1, \ldots, q$:

$$\Gamma_1 = \Delta_1, \quad \Gamma_j = \left[\begin{array}{c|c} \Gamma_{j-1} & R \\ \hline L & \Delta_j \end{array}\right]. \tag{3.8.2}$$

Lidskii [45] introduces the following assumption:

Assumption (Li): Γ_{j-1} is nonsingular for all $j = 2, \ldots, q$. Equivalently, Γ_j is nonsingular for $j = 1, \ldots, q-1$.

Because Γ_{j-1} is nonsingular for $j \geq 2$, the Schur complement

$$\Omega_j = \Delta_j - L\Gamma_{j-1}^{-1}R$$

of Γ_{j-1} in Γ_j is well-defined. In addition, since Γ_j is nonsingular, Ω_j is itself nonsingular for $j < q$, because $\det \Gamma_j = \det \Gamma_{j-1} \det \Omega_j$, $j = 2, \ldots, q$. For $j = q$, Ω_q may be singular. By a slight abuse of language, we still say that $\Omega_1 = \Delta_1 = \Gamma_1$ is a Schur complement (it corresponds to Γ_0 inexistent: $f_0 = 0$).

Under the assumption (Li), the Lidskii process asserts that for each step j there are exactly $n_j r_j$ eigenvalues with Puiseux order $1/n_j$ and coefficients deduced from $\sigma(\Omega_j)$ by taking the n_j roots of order n_j of each of the r_j eigenvalues. This follows from the resolution of the equation

$$\det \Gamma_j(t) = 0, \tag{3.8.3}$$

with $t = z^{n_j}$ and

$$\Gamma_j(t) = \left(\begin{array}{c|c} \Gamma_{j-1} & R \\ \hline L & \Delta_j - tI_{r_j} \end{array}\right).$$

For $\det \Gamma_{j-1} \neq 0$ by the Schur complement formula:

$$\det\Gamma_j(t) = \det\Gamma_{j-1} \det(\Omega_j - tI_{r_j}).$$

For each j, the eigenvalues of Ω_j, yield the n_j roots equal to the n_j coefficients which are sought for.

The set of matrices A that do not satisfy (Li) is *nongeneric* in $\mathbb{C}^{n\times n}$ [41]. See section 3.8.3.

3.8.3 The nongeneric step where Γ_j is singular

The challenge to go beyond (Li) has been faced in [11] and more directly in [6, 7, 46, 19, 20]. In [11, 46], a *nongeneric* step in the Lidskii process is analyzed under the hypothesis:

(H): there exists $j\,,\ 1 \le j \le q$ such that $\det \Gamma_j = 0$ and $\det \Gamma_i \neq 0\,,\ i \neq j\,.$

What happens when the Lidskii process hits the first nongeneric step j ?

It is shown in [11] that the spectral projection $P(s)$ associated with $0 \in \sigma(J)$ can be represented as the sum of two analytic projections in $s\,.$ The first (resp. second) one concerns the m_{j-1} (resp. $m - m_{j-1}$) eigenvalues $\nu(s)$ converging with order $\le 1/n_{j-1}$ (resp. $\ge 1/n_j$). Analyticity cannot be guaranteed any more. However, it is clear that more algebraic computation can be performed [46, 19, 20].

3.8.4 What can we conclude when (Li) does not hold ?

What happens when one Γ_j is singular? Baumgärtel [11] does not face the computational issue, which is studied in [46]. Results presented in [6, 7] rely partly on [46]. However, the conclusions presented in [46] are not entirely correct. The necessary modifications are presented in [19, 20]. In section 3.9, we will depend on these modifications to establish that $L_1 = \sigma(\Omega) \subseteq \mathrm{Lim}$ under an assumption weaker than (Li).

We explicit what happens when either the first or last step is nongeneric. When $j = 1$ is nongeneric, $\det \Gamma_1 = 0\,.$ In this case, some eigenvalues converge with orders $> 1/n_1$ and others with orders $< 1/n_2\,.$

If we look at the case $j = q\,, \det \Gamma_q = 0\,, \det \Gamma_i \neq 0\,,\ i < q\,.$ The new feature is that there are no subsequent steps. At step $q-1\,,$ all m_{q-1} eigenvalues have been classified. When $\det \Gamma_q = 0\,,\ n_q(r_q - \omega_q)$ (resp. $n_q\omega_q$) eigenvalues converge with order $= 1/n_q$ (unknown orders $> 1/n_q$), where ω_q is the algebraic multiplicity of $0 \in \sigma(\Omega)\,.$

Let us illustrate the generic theory by the following example adapted from [46] which uses the explicit formulation of Lidskii to find the sequence of imbedded matrices Γ_j of increasing order $f_j\,,\ j = 1, 2\,.$

Example 3.8.1 *Let E be a 7×7 Jordan matrix with a unique zero eigenvalue and three Jordan blocks with respective dimensions $3\,,\ 3\,,$ and $1\,.$ Lidskii's results show that for a small $|s|\,,$ there are six eigenvalues for $E + sA$ of order $s^{1/3}\,,$ and one of order $s\,.$ Lidskii has shown [45] that the coefficients of the leading terms depend only*

on the elements of the matrix A *marked with a diamond in the following matrix:*

$$A = \left[\begin{array}{ccc|ccc|c} * & * & * & * & * & * & * \\ * & * & * & * & * & * & * \\ \diamond & * & * & \diamond & * & * & \diamond \\ \hline * & * & * & * & * & * & * \\ * & * & * & * & * & * & * \\ \diamond & * & * & \diamond & * & * & \diamond \\ \hline \diamond & * & * & \diamond & * & * & \diamond \end{array}\right] = [a_{ij}]. \tag{3.8.4}$$

More precisely, let Γ_1 *be the* 2×2 *matrix whose four entries are the four diamonds in the top left block of* A *in (3.8.4), that is,*

$$\Gamma_1 = \left[\begin{array}{cc} a_{31} & a_{34} \\ a_{61} & a_{64} \end{array}\right] \quad \textit{of order} \quad f_1 = 2.$$

Γ_1 *is assumed to be invertible under* (Li) *. Then the perturbed matrix* $E(s) = E + sA$ *has six eigenvalues with leading terms in* s *of the form*

$$(\xi_1^{(k)})^{1/3} s^{1/3}, \quad k = 1, 2,$$

for $\xi_1^{(1)}$, $\xi_1^{(2)} \in \sigma(\Gamma_1)$ *. Now, for the* 7^{th} *eigenvalue,* $\nu(s)$ *,*

$$\Gamma_2 = \left[\begin{array}{c|c} \Gamma_1 & R \\ \hline L & \Delta_2 \end{array}\right] = \left[\begin{array}{cc|c} a_{31} & a_{34} & a_{37} \\ a_{61} & a_{64} & a_{67} \\ \hline a_{71} & a_{74} & a_{77} \end{array}\right].$$

The Schur complement *of* Γ_1 *in* Γ_2 *is*

$$\xi_2 = \Delta_2 - L\Gamma_1^{-1}R = a_{77} - [a_{71}\ a_{74}]\ \Gamma_1^{-1} \left[\begin{array}{c} a_{37} \\ a_{67} \end{array}\right].$$

This shows that the perturbed matrix $E(s)$ *has one eigenvalue* $\nu(s)$ *with leading term* $(\xi_2)s$ *for* $\xi_2 \neq 0$ *: this corresponds to the trivial Jordan block associated with* (0^1) *. If* $\xi_2 = 0$ *, then the eigenvalue* $\nu(s)$ *converges to* 0 *with order* > 1 *.*

This concludes the presentation given in [46].

Because Δ_2 *is of order* 1 *, we can say more by using (1.6.8).*

$$\det\Gamma_2(z) = \det \left[\begin{array}{cc} \Gamma_1 & R \\ L & \Delta_2 - z \end{array}\right] = (\Delta_2 - z)\det\Gamma_1 - L^T \mathrm{adj}\Gamma_1 R,$$

and ξ_2 *is a root of* $\det\Gamma_2(z)$ *which is a polynomial of degree* 1 *(resp.* 0*) iff* $\det\Gamma_1 \neq 0$ *(resp.* $= 0$*).*

If $\det\Gamma_1 \neq 0$*,* ξ_2 *is the* 1×1 *Schur complement predicted by Lidskii. If* $\det\Gamma_1 = 0$*,* $\det\Gamma_2(z)$ *reduces to the constant* $-L^T \mathrm{adj}\Gamma_1\ R$*, which may be* 0 *(*$\xi_2 \in \mathbb{C}$*) or* $\neq 0$ *(*ξ_2 *does not exist).*

$\triangle$

3.9 Algebraic determination of L_1 under (Δ)

3.9.1 Notations associated with $\sigma(E)$ when Ker $E \cap$ Im $E = S \neq \{0\}$

The following notations are associated with $0 \in \sigma(E)$ when Ker $E \cap$ Im $E = S \neq \{0\}$:

$m =$ algebraic multiplicity of $0 \in \sigma(E)$,

$g =$ geometric multiplicity of 0, $1 \leq g =$ dim Ker E, $n - r = g < m$, $r =$ rank $E < n$,

$r' =$ rank $G =$ rank $(V^H U)$, $r' \leq r =$ rank E,

$g' =$ number of *trivial* Jordan blocks corresponding to 0, $0 \leq g' < g$,

$f =$ number of Jordan blocks of size ≥ 2 (*nontrivial* Jordan blocks corresponding to 0), $f = g - g' \leq g$,

$K' =$ eigenspace generated by the $g' = g - f$ eigenvectors belonging to a (trivial) Jordan chain of length 1, $g' =$ dim K',

$S =$ subspace generated by the f eigenvectors *starting* a non trivial Jordan chain,

$T =$ subspace generated by the f vectors *ending* a non trivial Jordan chain.

3.9.2 Application of Lidskii's theory when $g' \geq 1$: the matrix $\tilde{\Pi}(z)$ of order g

In this section, we restrict our attention to (Δ) with $g > g' \geq 1$ when $0 \in \sigma(E)$ is defective. This is equivalent to the assumption $n_q = 1$, hence $q \geq 2$, $r_q = g'$, $0 < f = g - g' < g$. We look for the possible convergence of $\nu(s)$ to 0 with order ≥ 1 in s, that is

$$\nu(s) = \xi s + o(s), \tag{3.9.1}$$

and $\xi \in$ Lim is possibly 0.

The g Jordan blocks for $\nu = 0 \in \sigma(E)$ are ordered by non increasing size. The same arrangement for non zero eigenvalues induces a complete Jordan basis X such that $E = XJX^{-1}$, therefore

$$E + sA = X(J + sX^{-1}AX)X^{-1} = X(J + sB)X^{-1}, \tag{3.9.2}$$

where J is a Jordan form for E, and $X^{-1}AX = B$. We assume that $\nu = 0$ is placed first and $\bar{X} = \left[\begin{array}{ccc} e_1 & \dots & e_m \end{array}\right]$ (resp. $\bar{Y}^T = \left[\begin{array}{c} e_1^T \\ \vdots \\ e_m^T \end{array}\right]$) represents the right (resp. left) Jordan basis for 0 of algebraic multiplicity m. $\bar{X}$ (resp. $\bar{Y}^T$) consists of the m first canonical vectors (resp. rows).

In these bases, we select the *eigenvectors*. This defines the $n \times g$ matrix $\tilde{X} = [Z, X']$ (resp. $\tilde{Y} = [W, Y']$) such that Z, X' (resp. W, Y') are the eigenvectors starting the non trivial and trivial Jordan blocks for J (resp. J^T) respectively. Therefore Z is a basis for S, W a basis for T and X' a basis for K' which satisfy

$$Y'^T X' = I_{g'}, \quad W^T Z = 0_f \quad \text{and } \tilde{Y}^T \tilde{X} = \left[\begin{array}{cc} 0_f & 0 \\ 0 & I_{g'} \end{array}\right] \tag{3.9.3}$$

[45, 46, 18].

$\tilde{P} = \tilde{X}\tilde{Y}^T$ is not a projection (on $\operatorname{Ker} E$ or $K' = \operatorname{Im} X'$). In fact it satisfies $\tilde{P}^2 = P' = X'Y'^T \neq \tilde{P}$, where $P' = X'Y'^T$ is the eigenprojection on K'.

Let us define $\tilde{\Pi} = \tilde{Y}^T B \tilde{X} = \left[\begin{array}{cc} \Gamma & R \\ L & \Pi' \end{array}\right]$, with

$$\Gamma = W^T BZ, \text{ of order } f, \quad L = Y'^T BZ, \quad R = W^T BX'$$

and where $\Pi' = Y'^T BX'$ of order g' represents the Galerkin approximation $P'BP'$ restricted to K', whereas $\tilde{\Pi}$ does *not* correspond to a Galerkin approximation ($\tilde{P}$ is not a projection).

We define $\tilde{\Pi}(z) = \left[\begin{array}{cc} \Gamma & R \\ L & \Pi' - zI_{g'} \end{array}\right]$, so that

$$\tilde{q}(z) = \det \tilde{\Pi}(z) \tag{3.9.4}$$

is an scalar polynomial in z of degree $\leq g'$.

Such a construction is possible for $g' \geq 1$, which we assume below. We shall look at the case $g' = 0$ later. When $g' \geq 1$, then $n_q = 1$ for the last step in Lidskii's algebraic theory. Lidskii's theory asserts that the coefficients ξ are the roots of $\tilde{q}(z) = 0$ with $g' \geq 1$ under the assumption (Li). Can this assumption be weakened?

3.9.3 (G): $\det\Gamma \neq 0 \Rightarrow L_1 = \sigma(\Omega) \subseteq \mathrm{Lim}$

When Γ is invertible, let Ω be the Schur complement of the block Γ in $\tilde{\Pi}$, that is

$$\Omega = \Pi' - L\Gamma^{-1}R. \tag{3.9.5}$$

Then, we can write

$$\Omega = \Pi^{'} - L\Gamma^{-1}R = {Y^{'}}^{T}(I_n - B(Z\Gamma^{-1}W^T))BX^{'}. \tag{3.9.6}$$

Assumption (G): $g^{'} \geq 1$ and $\Gamma = W^TBZ$ invertible ($\Leftrightarrow$ $\det\Gamma \neq 0$).

The following Proposition shows that the elements of Lim are deduced in part from $\sigma(\Omega)$.

Proposition 3.9.1 *[18] Under the assumption (G), $l_* \geq g^{'}$ and $L_1 = \sigma(\Omega) \subseteq$ Lim.*

Proof. By assumption Γ exists and is invertible. Then, we have

$$\tilde{q}(z) = \det\tilde{\Pi}(z) = \det\Gamma\ \det(\Omega - zI_{g^{'}}),$$

for Ω defined in 3.9.5 as the Schur complement of Γ in $\tilde{\Pi} \in \mathbb{C}^{g^{'}\times g^{'}}$. This shows that $\deg\det\tilde{\Pi}(z) = g^{'}$, and $Z(\tilde{q}(z)) = \{z \in \mathbb{C} : \tilde{q}(z) = 0\} = \sigma(\Omega)$. And of course, $l_* \geq g^{'} \geq 1$, since $g^{'} \geq 1$ by assumption.

The assumption (G) on Γ guarantees that at least $g^{'}$ eigenvalues stay at finite distance. But there is no guarantee that exactly $n - g^{'}$ eigenvalues diverge to ∞. Hence, $\sigma(\Omega) \subseteq$ Lim.

□

Example 3.9.1 *Let* $E = J = diag\left[\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, 0\right]$ *be in Jordan form and* $A = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} = [a_{ij}]$. $g = 3$, $g^{'} = 1$.

Then $\Gamma_1 = a_{31} = 0$, *and* $\Gamma_2 = \begin{bmatrix} a_{31} & a_{34} \\ a_{51} & a_{54} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ *which is nonsingular. For* $\tilde{X} = [e_1, e_4, e_6]$ *and* $\tilde{Y} = [e_3, e_5, e_6]$, *one has*

$$\tilde{\Pi} = \tilde{Y}^TA\tilde{X} = \Gamma_3 = \left[\begin{array}{cc|c} a_{31} & a_{34} & a_{36} \\ a_{51} & a_{54} & a_{56} \\ \hline a_{61} & a_{64} & a_{66} \end{array}\right] = \left[\begin{array}{cc|c} 0 & 1 & 1 \\ 1 & 0 & 1 \\ \hline 0 & 0 & 0 \end{array}\right], \text{ and } \Pi^{'} = (0).$$

Then

$$\Omega = (a_{66} - \begin{bmatrix} a_{61} & a_{64} \end{bmatrix}\Gamma_2^{-1}\begin{bmatrix} a_{36} \\ a_{56} \end{bmatrix}) = (0 - 0) = (0).$$

This example satisfies the condition (G) but not (Li), *where* $q = 3$, $n_3 = 1$ *and* $\Gamma_2 = \Gamma_{q-1}$ *is nonsingular. This shows that* $\sigma(\Omega) = \{0\} \subseteq$ Lim. *In fact, by computing the eigenvalues of* $A(t) = A + tE$ *for increasingly large* $|t|$, *one finds that* $\{0\} =$ Lim.

△

Under (G), let us define

$$F = Z\Gamma^{-1}W^T \in \mathbb{C}^{n\times n}. \tag{3.9.7}$$

Lemma 3.9.2 *[18] (i) The matrix F defined in (3.9.7) satisfies $F^2 = 0$ and* $\operatorname{rank} F = 1$.
(ii) $Q = FB$ is a rank $1-$ projection.

Proof. (i) Using the properties of W and Z in (3.9.3) and according to the definition of the matrix F in (3.9.7) one has $F^2 = Z\Gamma^{-1}(W^TZ)\Gamma^{-1}W^T = 0$ since $W^TZ = 0$ and $F \neq 0$. Therefore $\operatorname{rank} F = 1$, and F is nilpotent.

(ii) The matrix $Q = FB = Z\Gamma^{-1}W^TB$ satisfies

$$Q^2 = Z\Gamma^{-1}[W^TBZ\Gamma^{-1}]W^TB = Q, \quad \text{since } W^TBZ = \Gamma,$$

and $Q = FB$ shows that $1 \leq \operatorname{rank} Q \leq \min(\operatorname{rank}F, \operatorname{rank}B) = 1$. Therefore Q is a rank $1-$ projection.

□

Lemma 3.9.3 *[18] Ω represents the map $P'B(I-Q)P'$ restricted to $K' = \operatorname{Im} X'$.*

Proof. The assertion is derived from the following equalities.

$$P'B(I-Q)P' = X'({Y'}^TBX' - {Y'}^TBZ\Gamma^{-1}W^TBX'){Y'}^T$$

$$= X'(\Pi' - L\Gamma^{-1}R){Y'}^T = X'\Omega{Y'}^T.$$

□

Theorem 3.9.4 *[18] When (Σ) does not hold, but (G) is valid, the matrix Ω replaces Π. This amounts to replace PAP by $P'B(I-Q)P'$, where $I-Q$ is a projection with rank $n-1$.*

Proof. Clear by Lemma 3.9.2 and 3.9.3.

□

Under (Σ), $1 \leq g' = g$ and $Z = W = 0$. Therefore $Q = 0$. In fact $Q \neq 0 \Longleftrightarrow (\Delta)$. Theorem 3.9.4 indicates how PAP and Π are modified to get $P'B(I-Q)P'$ and Ω when $Q \neq 0$ and $1 \leq g' < g$ under (Δ).

3.9.4 (Li) $\Rightarrow$ $L_1 = \sigma(\Omega) = \mathrm{Lim}$

The stronger result $\mathrm{Lim} = \sigma(\Omega)$ requires the assumption (Li). As we saw, this stronger assumption (which implies (G)) makes it sure that there is no interaction between the Jordan blocks of different sizes. Therefore $\mathrm{Lim} = \sigma(\Omega)$ where Ω can be singular. And exactly $n - g^{'}$ eigenvalues diverge to ∞, by Lidskii's theory [45, 46].

Proposition 3.9.5 *[18] Under the assumption* (Li), $l_* = g^{'}$ *and* $\mathrm{Lim} = L_1 = \sigma(\Omega)$.

This Proposition is the generic case studied in [45, 46]. Under (Li) $l_* = g^{'}$ and $\mathrm{Lim} = L_1 = \sigma(\Omega)$ generalizes $\mathrm{Lim} = \sigma(\Pi)$ which is valid under (Σ). Note that (Li) reduces to (G) when $q = 2$.

Proposition 3.9.6 *[41] Let the matrix E be given in $\mathbb{C}^{n\times n}$. The set of matrices A such that (A, E) satisfies* (Li) *is a dense open subset in $\mathbb{C}^{n\times n}$.*

The Proposition 3.9.6 leads naturally to the following definition.

Definition 3.9.7 *When $0 \in \sigma(E)$ is defective with at least one trivial Jordan block, the deviation process (A, E) is generic iff* (Li) *is satisfied.*

Via the Proposition 3.9.5, when (A, E) is generic, there are exactly $g^{'} \geq 1$ eigenvalues converging to $\sigma(\Omega)$ and $n - g^{'}$ eigenvalues diverging to ∞.

Definition 3.9.8 *We call kernel points for (A, E) the values in $\sigma(\Omega)$ which are in the resolvent set $re(A)$. We denote the set of kernel points, as*

$$K(A, E) = \sigma(\Omega) \cap re(A).$$

We know that under (G), $\sigma(\Omega) \subseteq \mathrm{Lim}$ and under (Li), $\sigma(\Omega) = \mathrm{Lim}$. Therefore under (G), $K(A, E) \subseteq \Lambda(A, E)$ and under (Li), $K(A, E) = \Lambda(A, E)$. And under (Σ), $K(A, E) = \Lambda(A, E) = \sigma(\Pi) \cap re(A)$.

Example 3.9.2 *Let A be the companion matrix associated with $\pi(z) = z^{11} + 1$, in upper Hessenberg form, and first column e_2.*
Let $E = UV^T$ with $U = [e, e_2]$ and $V = [e_{11}, e_3]$ of rank $r = 2$ for $e = [1, \ldots, 1]^T$.
E is in Jordan form for $0 \in \sigma(E)$, with 1 block of size 2 for $0 \in \sigma(E)$, $g = n - 2 = 9$, $g^{'} = 8$, $m = 10 < n = 11$.
We use the complete Jordan decomposition of E with the similarity transformation matrix X into the Jordan form J. The Jordan blocks associated with $0 \in \sigma(E)$ satisfy $n_1 = 2 > n_2 = 1$, for $r_1 = 1$, $r_2 = 8$, $q = 2$. Then we get:

$$J = diag\left[\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, 0, \ldots, 0, 1\right],\ J = X^{-1}EX \text{ and } B = X^{-1}AX,$$

$$\tilde{\Pi} = \tilde{Y}^T B \tilde{X} = \Gamma_2 = \left[\begin{array}{c|c|cccccc} 1 & 0 & 1 & \cdots & \cdots & \cdots & 1 \\ \hline 0 & 0 & -1 & -1 & \cdots & -1 & 0 \\ & -- & -- & -- & -- & -- & -- \\ \vdots & 0 & -1 & -1 & \cdots & \cdots & -1 \\ \vdots & 0 & 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \ddots & \ddots & \\ 0 & 0 & 0 & & 0 & 1 & 0 \end{array}\right] \quad \textit{of order } 9.$$

for $\tilde{X} = [e_1, e_3, \ldots, e_{10}]$ *and* $\tilde{Y} = [e_2, e_3, \ldots, e_{10}]$. *Therefore,* $\Gamma_1 = e_2^T B e_1 = 1 \neq 0$, *and* Ω *is the down right* 8×8 *diagonal block of* Γ_2 *of order* $g^{'} = 8$. *Hence* $(G) = (\mathrm{Li})$ *holds.*
Ω *has a simple block diagonal structure:* 1 *scalar* 0, 1 *block of order* 7 *which is in companion form, associated with* $p(z) = z^7 + z^6 + \cdots + 1$. *We conclude that* $\sigma(\Omega) = Zer(p(z)) \cup \{0\} = F(A, E) \cup \{-1\}$ *: a direct computation shows that* $\det \hat{A}(z) = \hat{\pi}(z) = z(z^7 + z^6 + \cdots + 1) = z(z+1)(z^2+1)(z^4+1)$.

In summary, $(G) = (\mathrm{Li})$ *holds, the eigenvalue* 0 *in* $\sigma(\Omega)$ *is simple, and* $\sigma(\Omega) = \mathrm{Lim}$: 7 *eigenvalues* $\nu(s)$ *converge with order* 1 *and* 1 *eigenvalue* $\nu(s)$ *converges with order* > 1. *See numerical illustration in chapters 6 and 7.*

△

3.9.5 $g^{'} \geq 1$ and Γ is singular

When Γ is not invertible, Ω does not exist, and $0 \leq deg\ \tilde{q}(z) \leq g^{'} - 1$. The trivial and nontrivial Jordan blocks interact. For $f \geq 1$, all we can say is $\mathrm{Lim} \supseteq L_1 = Z(\tilde{q}(z))$ when $\tilde{q}(z) \not\equiv 0$.

When $f = 1$ (hence $q = 2$ for $g^{'} \geq 1$), Γ is restricted to the scalar $\gamma = 0$. In general, for $\gamma \in \mathbb{C}$,

$$\tilde{\Pi}(z) = \left[\begin{array}{cc} \gamma & \mathbf{r}^T \\ \mathbf{l} & \Pi^{'} - zI \end{array}\right].$$

The formula (1.6.8) yields

$$\tilde{q}(z) = \det\tilde{\Pi}(z) = \gamma \det(\Pi^{'} - zI) - \mathbf{r}^T \mathrm{adj}(\Pi^{'} - zI)\mathbf{l}.$$

The coefficient of $z^{g^{'}}$ is $\gamma(-1)^{g^{'}}$. When $z \notin \sigma(\Pi^{'})$, then $\det\tilde{\Pi}(z) = 0$ iff $\mathbf{r}^T(\Pi^{'} - zI)^{-1}\mathbf{l} = \gamma$. The case $q = 2$, $f = 1$ corresponds to a unique non trivial Jordan block, in addition to at least one trivial one. Because of its algorithmic significance, a special case for $q = 2$, $f = 1$ is treated in more detail in Chapter 8.

3.9.6 $q = f = 1$: a case where $g^{'} = 0$

When there is a unique Jordan block with $n_1 > 1$ then $g = 1$ and $g^{'} = 0$. Generically the Puiseux exponent is $1/n_1 < 1$, and all eigenvalues $\lambda(t)$ escape to

∞. However $\mathrm{Lim} \neq \emptyset$ is possible nongenerically. This is illustrated by the following Example.

Example 3.9.3 *[5]*

$$A = \begin{bmatrix} 0 & 0 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, E = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \pi(z) = z^3 + 1.$$

$$U = [e_1, e_2]\,,\ V = [e_2, e_3]\,,\ G = V^H U = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \text{ has rank } 1\,.$$

$$B_z = \mathrm{adj}\,(zI_3 - A) = \begin{bmatrix} z^2 & -1 & -z \\ z & z^2 & -1 \\ 1 & z & z^2 \end{bmatrix},$$

and

$$Q(z) = V^H B_z U = \begin{bmatrix} z & z^2 \\ 1 & z \end{bmatrix}$$

$\det\ Q(z) \equiv 0$ *and* $Zer = Z(\det Q) = \hat{Z} = \mathbb{C}$.
For $z \notin \sigma(A)$, $M_z = \frac{1}{\pi(z)} \begin{bmatrix} z & z^2 \\ 1 & z \end{bmatrix}$ *has the spectrum* $\sigma(M_z) = \{0, \frac{2z}{z^3+1}\}$.
Therefore, $F(A, E) = re(A) \subset \mathbb{C}$ *and* $C(A, E) = \{0\}$.
$\lambda(t)$ *is any of the three roots of* $\lambda^3 - 2\lambda t + 1 = 0$*: one tends to* 0, *two roots tend to* ∞ *as* $|t| \longrightarrow \infty$. *Hence* $\mathrm{Lim} = \{0\}$.

$\triangle$

Example 3.9.3 shows a case where E is nilpotent, A is companion, $F(A, E) = re(A)$ is continuous but $C(A, E)$ is discrete. As a result, the eigenvalues evolve. Note also that $\mathrm{Lim} = \{0\}$ is not empty even though $g' = 0$.

However, when $g^{'} = 0$ in general, we do not know theoretically how to predict L_1.

3.10 The collective behaviour

So far we have looked at the *individual* behaviour of each of the m eigenvalues $\lambda_j(t) = \frac{\nu_j(s)}{s}$, $j = 1, \ldots, m$ which may possibly have a finite limit as $|t| \to \infty$. It is classical that Proposition 2.5.3 remains valid for the arithmetic mean

$$\hat{\nu}(s) = \frac{1}{m} \sum_{j=1}^{m} \nu_j(s),$$

under the perturbation sA for E [32, 33, 18], provided that P denotes now the *spectral* projection for $0 \in \sigma(E)$ on the *invariant* subspace $\mathrm{Ker}\, E^m$ of dimension m. Therefore $\Pi = PAP_{\restriction \mathrm{Ker} E^m}$ is of order m.

Proposition 3.10.1 *[18] Without assumption on* $0 \in \sigma(E)$,

$$\hat{\lambda}(t) = \frac{\hat{\nu}(s)}{s} \to \frac{1}{m} \text{ tr } \Pi, \quad as \quad |t| \to \infty.$$

The arithmetic mean of m *eigenvalues stays at finite distance.*

Among the m eigenvalues $\lambda_j(t)$, $j = 1, \ldots, m$ appearing in $\hat{\lambda}(t)$, l_* of them, $0 \leq l_* \leq m$ converge to Lim, and $m - l_*$ escape to ∞. This escape is realized in such a way that $\hat{\lambda}(t)$ converges to $\frac{1}{m}$ trΠ. Even if some of the m eigenvalues escape to ∞, they remain connected: their sum converges to tr Π in $\mathbb{C}$.

This connection will explain some of the amazing numerical results obtained in the experimental Chapters of Part II.

3.11 A summary for (Δ)

In general, we have

$$Zer = Zer(\det Q(z)) = (\sigma(A))^{r-1} \cup \hat{Z}$$

for $\hat{Z} = Zer(\hat{\pi}(z))$, $\hat{\pi}(z) = \det \hat{A}(z)$ and

$$\hat{A}(z) = \begin{bmatrix} zI - A & -U \\ V^H & 0 \end{bmatrix}.$$

Under (Δ), one has

1. $0 \leq g' < g < m \leq n \Longleftrightarrow \deg(\det Q(z)) < (n-1)r$
   $\Longleftrightarrow$ deg $\hat{\pi} < n - r \Longleftrightarrow (\Delta)$,
2. $F(A, E) = \hat{Z} \cap re(A)$, and card $\hat{Z} < n - r = g$, or $\hat{Z} = \mathbb{C}$,
3. card Lim $= l_*$, $0 \leq l_* \leq m \leq n$,
4. When $g' \geq 1$ and (G) holds $\Longrightarrow$ $\Omega = \Pi' - L\Gamma^{-1}R$ of order g' exists.

I. When $\hat{\pi}(z) \not\equiv 0$, then $\{C(A, E),\ \Lambda(A, E)\} \subseteq F(A, E) = \hat{Z} \cap re(A) \subset re(A)$,

1. when $g' \geq 1$, for $\tilde{\Pi}(z) = \begin{bmatrix} \Gamma & R \\ L & \Pi' - zI_{g'} \end{bmatrix}$, one has $L_1 = Zer(\tilde{\Pi}(z))$.
2. under (G), $L_1 = \sigma(\Omega) \subseteq$ Lim,
   (a) $K(A, E) \subseteq \Lambda(A, E) \subseteq F(A, E)$, with $K(A, E) = \sigma(\Omega) \cap re(A)$,
   (b) Ω represents the map $P'B(I - Q)P'$ restricted to $K' = \text{Im } X'$, for a rank $1-$ projection $Q = FB$ with $F = Z\Gamma^{-1}W^T$, and $F^2 = 0$.

3. under (Li), $L_1 = \sigma(\Omega) = \text{Lim}$ and $\{C(A,E),\ K(A,E) = \Lambda(A,E)\} \subseteq F(A,E)$.
4. when $r = 1$, $\Lambda(A,E) \subseteq F(A,E)$ and $Zer = \hat{Z}$.

II. When $\hat{\pi}(z) \equiv 0$, then $\hat{Z} = Zer = \mathbb{C}$, and $F(A,E) = re(A)$, and

1. if $C(A,E)$ contains at most $n-1$ distinct points, then $C(A,E)$ is discrete and the eigenvalues evolve. In addition, under (G) $K(A,E) = \sigma(\Omega) \cap re(A) \subseteq \Lambda(A,E)$.
2. if $C(A,E)$ contains at least n distinct points, then $C(A,E) = re(A)$ is a continuous set and

$$K(A,E) = \Lambda(A,E) = \emptyset, \quad \text{Lim} = \sigma(A) = \sigma(A(t)), \quad \forall\ t \in \mathbb{C},$$

$l_* = n$, and there is no evolution for the eigenvalues.

When (G) does not hold ($g' = 0$ or $\det \Gamma = 0$), there is no theoretical answer to the algebraic determination of L_1. A computational approach is to plot the eigenvalue maps: $t \in \mathbb{C} \mapsto \lambda_i(t)$, as t varies in $\mathbb{C}$ [27, 51, 30, 31, 5]. See the numerical experiments for $t \mapsto \sigma(A(t))$ in Chapters 6,7 and 8 of Part II.

3.12 A summary for (Σ)

To better appreciate the role of 0 defective / 0 semi-simple we list below the results of Chapters 2 and 3 obtained under (Σ). $G = V^H U$ has rank r.

Under (Σ), we still use the notations Zer, $\hat{Z}$, $\hat{\pi}(z)$, and $\hat{A}(z)$ of the section 3.11. One has

1. $1 \le g' = g = n - r = m < n \iff \deg(\det Q(z)) = (n-1)r$
$\iff \deg \hat{\pi} = n - r \iff (\Sigma)$,
2. $F(A,E) = \hat{Z} \cap re(A)$,
3. $K(A,E) = \Lambda(A,E)$, with cardinal $\le g = n - r$,
4. $\text{card}\, \hat{Z} = n - r = g = \text{card Lim} = l_*$,
5. for $r = 1$, $g = n - 1$
   (a) $Zer = \hat{Z}$ and $\sigma(\Pi) = \text{Lim}$, where Π is the $g \times g$ matrix representing PAP restricted to $\text{Ker}\, E$. $P = I - UG^{-1}V^H$ is the eigenprojection associated with $0 \in \sigma(E)$, which projects onto the eigenspace $\text{Ker}\, E$ along $\text{Im}\, E$,

(b) $\Lambda(A,E) \subseteq F(A,E) = C(A,E)$.

6. for $2 \leq r < n$,

   (a) $\sigma(\Pi) = \mathrm{Lim}$, $\mathrm{card}\,\sigma(\Pi) = \mathrm{card}\,\hat{Z} = g = n - r$, but in general, $\sigma(\Pi) \neq \hat{Z}$ is possible,

   (b) $\{C(A,E),\ \Lambda(A,E)\} \subseteq F(A,E)$.

7. $Q = 0$; Ω (resp. $P'B(I-Q)P'$) is replaced by Π (resp. PAP).

8. If card $\Lambda(A,E) = n - r$ then

   $\sigma(A) \cap \sigma(\Pi) = \hat{Z} \cap \sigma(A) = \emptyset$ and

   $\hat{\pi}(z) = \det(zI_g - \Pi)\det G$, $\sigma(\Pi) = \hat{Z}$.

Chapter 4

Regular matrix pencils and HD theory

4.1 Introduction

Let $A, E \in \mathbb{C}^{n\times n}$ be two given matrices, where $\operatorname{rank} E = r < n$. The matrix E of rank r is written under the form $E = UV^H$ where $U, V \in \mathbb{C}^{n\times r}$ have rank r. 0 is an eigenvalue of E with algebraic (resp. geometric) multiplicity m (resp. $g = n - r \leq m$). Under the condition (Σ), one has $m = g = n - r$, and under (Δ), $g < m \leq n$. We consider the regular pencil $(A - zI) + tE$, defined for $t \in \mathbb{C}$ and for the parameter $z \in re(A) = \mathbb{C}\backslash\sigma(A)$, where $\sigma(A)$ denotes the spectrum of A.
In this Chapter, we analyze how the structure of the pencil $(A - zI) + tE$ varies with the parameter $z \in re(A)$. Our analysis follows Gantmacher [37]. It turns out that the appropriate notion is that of *frontier points*, which is essential in Homotopic Deviation theory. As it was shown in Chapters 2 and 3, at frontier points, $\operatorname{rank}(M_z) < r$, where $M_z = V^H(zI - A)^{-1}U$ is an $r \times r$ matrix defined for $z \in re(A)$.
We shall show in section 4.3 that, when z is not (resp. is) a frontier point, the structure of the pencil $(A - zI) + tE$ depends only on $r = \operatorname{rank} E$ but not on A (resp. depends on a_z, $1 \leq a_z \leq r$, the algebraic multiplicity of $0 \in \sigma(M_z)$). In section 4.4, as an example, the structure of the regular pencil $A + tE$ for A invertible ($z = 0 \in re(A)$) is determined.

4.2 Strictly equivalent forms for pencils of matrices

4.2.1 Pencils of rectangular matrices

The present section deals with the following problem:

Given four matrices A, E, A_1 and E_1 all of dimension $m \times n$ with elements from $\mathbb{C}$, it is required to find under what conditions there exist two square non-singular matrices P and Q of orders m and n, respectively, such that

$$PAQ = A_1, \quad PEQ = E_1. \tag{4.2.1}$$

By introduction of the pencils of matrices $A + tE$ and $A_1 + tE_1$ the two matrix equations (4.2.1) can be replaced by the single equation

$$P(A + tE)Q = A_1 + tE_1. \tag{4.2.2}$$

Definition 4.2.1 *Two pencils of rectangular matrices $A+tE$ and A_1+tE_1 of the same dimensions $m \times n$ connected by the equation (4.2.2) in which P and Q are constant square non-singular matrices (i.e., matrices independent of t) of orders m and n, respectively, will be called strictly equivalent.*

A criterion for equivalence of the pencils $A+tE$ and A_1+tE_1 follows from the general criterion for equivalence of matrix polynomials and consists in the equality of the invariant polynomials or, what is the same, of the elementary divisors of the pencils $A + tE$ and $A_1 + tE_1$ [37].

Now, we shall establish a criterion for strict equivalence of two pencils of matrices and we shall determine for each pencil a strictly equivalent canonical form. This strict equivalence has the following natural geometrical interpretation given in [37].

We consider a pencil of linear operators $A + tE$ mapping $\mathbb{C}^n$ into $\mathbb{C}^m$. For a definite choice of bases in these spaces the pencil of operators $A + tE$ corresponds to a pencil of rectangular matrices $A+tE$ (of dimension $m \times n$); under a change of bases in $\mathbb{C}^n$ and $\mathbb{C}^m$ the pencil $A + tE$ is replaced by a strictly equivalent pencil $P(A + tE)Q$, where P and Q are square nonsingular matrices of order m and n. Thus, a criterion for strict equivalence gives a characterization of that class of matrix pencils $A + tE$ mapping $\mathbb{C}^n$ into $\mathbb{C}^m$ for various choices of bases in these spaces.

In order to obtain a canonical form for a pencil it is necessary to find bases for $\mathbb{C}^n$ and $\mathbb{C}^m$ in which the pencil of operators $A + tE$ is described by matrices of the simplest possible form.

4.2.2 Regular pencils of matrices

All the pencils of matrices $A+tE$ of dimension $m \times n$ fall into two basic categories: *regular* and *singular* pencils.

Definition 4.2.2 *A pencil of matrices $A + tE$ is called regular if*

1. *A and E are square matrices of the same order n; and*

2. the determinant $\det(A+tE)$ *does not vanish identically.*

In all other cases ($m \neq n$*, or* $m = n$ *but* $\det(A+tE) \equiv 0$*), the pencil is called singular.*

A criterion for strict equivalence of regular pencils of matrices and also a canonical form for such pencils were established by Weierstrass in 1867 on the basis of his theory of elementary divisors. The analogous problems for singular pencils were solved later, in 1890, by the investigations of Kronecker [37].

Let us consider the special case where the pencils $A+tE$ and A_1+tE_1 consist of square matrices ($m=n$) such that $\det E \neq 0$, $\det E_1 \neq 0$. In this case, the two concepts of *equivalence* and *strict equivalence* of pencils coincide [37].

Theorem 4.2.3 *[37] Two pencils of square matrices of the same order* $A+tE$ *and* A_1+tE_1 *for which* $\det E \neq 0$ *and* $\det E_1 \neq 0$ *are strictly equivalent if and only if the pencils have the same elementary divisors in* $\mathbb{C}$.

Example 4.2.1 *[37] Let* $A+tE = \begin{bmatrix} 2 & 1 & 3 \\ 3 & 2 & 5 \\ 3 & 2 & 6 \end{bmatrix} + t \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 2 \\ 1 & 1 & 3 \end{bmatrix}$, *and* $A_1+tE_1 = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} + t \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$. *It is shown that each of the pencils* $A+tE$ *and* A_1+tE_1 *has one elementary divisor,* $t+1$, *and* $\det(A+tE) \equiv \det(A_1+tE_1) \equiv t+1$ *does not vanish identically. However, they are not strictly equivalent, since the matrices* E *and* E_1 *are of rank* 2 *and* 1, *respectively. If an equation (4.2.2) were to hold, it would follow from it that the ranks of* E *and* E_1 *are equal.*

△

According to the definition 4.2.2, it is quite possible in a regular pencil to have $\det E = 0$ (and even $\det A = \det E = 0$). By the way the example 4.2.1 shows that Theorem 4.2.3 does not hold for the regular pencils satisfying definition 4.2.2 where $\det E = 0$ or $\det E = \det E_1 = 0$.

In order to preserve Theorem 4.2.3, we need to introduce the concept of *infinite* elementary divisors of a pencils. We shall consider the pencil $A+tE$ as a *homogeneous* function of the two parameters t, $\lambda \in \mathbb{C}$: $\lambda A + tE$. Then the determinant $\det(\lambda A+tE)$ is a homogeneous function of t and λ. By determining the greatest common divisor $D_k(t,\lambda)$ of all the minors of order k of the matrix $\lambda A+tE$ ($k=1,2,\ldots,n$), we obtain the invariant polynomials by the well known formulas

$$i_1(t,\lambda) = \frac{D_n(t,\lambda)}{D_{n-1}(t,\lambda)}, \quad i_2(t,\lambda) = \frac{D_{n-1}(t,\lambda)}{D_{n-2}(t,\lambda)}, \quad \ldots$$

where all the $D_k(t,\lambda)$ and $i_j(t,\lambda)$ are homogeneous polynomials in t and λ.

Splitting the invariant polynomials into powers of homogeneous polynomials irreducible over $\mathbb{C}$, we obtain the elementary divisors $e_\alpha(t,\lambda)$ ($\alpha = 1,2,\dots$) of the pencil $\lambda A + tE$ in $\mathbb{C}$.

When we set $\lambda = 1$ in $e_\alpha(t,\lambda)$ we are back to the elementary divisors $e_\alpha(t)$ of the pencil $A + tE$. Conversely, from each elementary divisor $e_\alpha(t)$ of degree q we obtain the correspondingly elementary divisor $e_\alpha(t,\lambda)$ by the formula $e_\alpha(t,\lambda) = \lambda^q e_\alpha(\frac{t}{\lambda})$. We can obtain in this way all the elementary divisors of the pencil $\lambda A + tE$ apart from those of the form λ^q.

Elementary divisors of the form λ^q exist iff $\det E = 0$ and are called *infinite* elementary divisors of the pencil $A + tE$.

Since strict equivalence of the pencils $A + tE$ and $A_1 + tE_1$ implies strict equivalence of the pencils $\lambda A + tE$ and $\lambda A_1 + tE_1$, we see that for strictly equivalent pencils $A + tE$ and $A_1 + tE_1$ not only their *finite* but also their *infinite* elementary divisors must coincide.

Theorem 4.2.4 *[37] Two regular pencils $A + tE$ and $A_1 + tE_1$ are strictly equivalent iff they have the same finite and infinite elementary divisors.*

In the example 4.2.1, the pencils have the same *finite* elementary divisor $t + 1$, but different *infinite* elementary divisors (the first pencil has one infinite elementary divisor λ^2 and the second has two: $\lambda,\ \lambda$). Therefore these pencils turn out to be not strictly equivalent.

Theorem 4.2.5 *[37] Every regular pencil $A + tE$ can be reduced to a (strictly equivalent) canonical quasi-diagonal form*

$$\{N^{(u_1)}, N^{(u_2)}, \dots, N^{(u_s)},\ J + tI\} \quad (N^{(u)} = I_u + tH^{(u)}), \tag{4.2.3}$$

where the first s diagonal blocks correspond to infinite elementary divisors $\lambda^{u_1}, \lambda^{u_2}, \dots, \lambda^{u_s}$ of the pencil $A + tE$ and where the normal form of the last diagonal block $J + tI$ is uniquely determined by the finite elementary divisors of the given pencil.

Proof. Suppose that $A + tE$ is an arbitrary regular pencil. Then there exists a number c such that $\det(A + cE) \neq 0$. The given pencil can be represented in the form $A_1 + (t - c)E$, where $A_1 = A + cE$, so that $\det A_1 \neq 0$. We multiply the pencil on the left by A_1^{-1}: $I + (t-c)A_1^{-1}E$. By a similarity transformation we put the pencil in the form

$$I + (t - c)\{J_0, J_1\} = \{I_{n_0} - cJ_0 + tJ_0,\ I_{n_1} - cJ_1 + tJ_1\}, \tag{4.2.4}$$

where $\{J_0, J_1\}$ is the quasi-diagonal normal form of $A_1^{-1}E$, J_0 is a nilpotent Jordan matrix, $\det J_1 \neq 0$ and $n_i = \operatorname{order} J_i$ for $i = 0, 1$.

We multiply the first diagonal block on the right-hand side of (4.2.4) by $(I_{n_0} - cJ_0)^{-1}$ and obtain: $I_{n_0} + t(I_{n_0} - cJ_0)^{-1}J_0$. Here the coefficient of t is a nilpotent

matrix: from $J_0^l = 0$ for some integer $l > 0$, it follows that $[(I_{n_0} - cJ_0)^{-1} J_0]^l = 0$. Therefore by a similarity transformation we can put this pencil into the form

$$I + tJ_0 = \{N^{(u_1)}, N^{(u_2)}, \ldots, N^{(u_s)}\}, \quad (N^{(u)} = I_u + tH^{(u)}), \tag{4.2.5}$$

where $H^{(u)}$ is a matrix of order u whose elements in the first superdiagonal are 1, while the remaining elements are zero, that is

$$H^{(u)} = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix}.$$

We multiply the second diagonal block on the Right-hand side of (4.2.4) by J_1^{-1}; it can then be put into the form $J + tI$ by a similarity transformation, where J is a matrix of *normal* form (or of *Jordan* form).

□

We recall that two families of square matrix pencils have been encountered in HD theory. They play an essential role. They are for $z \in \mathbb{C}$:

a) $\hat{A}(z) = z \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} A & U \\ -V^H & 0 \end{bmatrix}$ of order $n + r$,

b) $\tilde{\Pi}(z) = \begin{bmatrix} \Gamma & R \\ L & \Pi' \end{bmatrix} - z \begin{bmatrix} 0 & 0 \\ 0 & I_{g'} \end{bmatrix}$ of order g.

In each case, we supposed that the matrix pencils were *regular*. Then, we looked for the roots of the polynomials $\hat{\pi}(z) \not\equiv 0$ and $\tilde{q}(z) \not\equiv 0$.

Therefore it is natural, in HD theory, to analyze the structure of regular pencils. In what follows, $t \in \hat{\mathbb{C}}$ defines the pencil $(A - zI) + tE$, and $z \in re(A) \subset \mathbb{C}$ is a parameter.

4.3 The structure of the regular pencil $(A - zI) + tE$ depending on the parameter z in $re(A)$

We consider the pencil $(A - zI) + tE$ for z given in $re(A)$: it is a *regular* pencil, since its determinant is nonzero for $t = 0$. How does the structure of the pencil evolve as z varies in $re(A)$?

Proposition 4.3.1 *For $z \in re(A)\backslash F(A,E)$, the pencil $(A - zI) + tE$ is strictly equivalent to the 2×2 block matrix*

$$\left[\begin{array}{ccc|ccc} 1 & t\varepsilon & 0 & & & \\ & \ddots & \ddots & & & \\ & & 1 & & & \\ \hline & & & \frac{1}{\mu_{1z}} & \varepsilon & 0 \\ & & & & \ddots & \ddots \\ & & & & & \frac{1}{\mu_{rz}} \end{array}\right] \tag{4.3.1}$$

corresponding to the partition $n = g + r$, where ε represents 0 or 1 and $\mu_{iz} \in \sigma(M_z)$, $i = 1, \ldots, r$, $t \in \mathbb{C}$.

Proof. From the Lemmas 3.2.1 and 3.3.1, we know that, for any $z \in re(A)\backslash F(A,E)$, M_z is invertible. In this case, the equality $\det((A - zI) + tE) = 0$ for any $t \in \mathbb{C}$ is equivalent to say that z is an eigenvalue of $A + tE$ in HD and

$$z \in \sigma(A + tE) \Longleftrightarrow t = \frac{1}{\mu_{iz}}, \quad \text{for } 0 \neq \mu_{iz} \in \sigma(M_z) \text{ and } i = 1, \ldots, r.$$

This quantifies the normal form of the last diagonal block of (4.2.3) by substituting $t = \frac{1}{\mu_{iz}}$ for $i = 1, \ldots, r$. The r finite eigenvalues of $(A - zI) + tE$ are the r numbers $\frac{1}{\mu_{iz}}$. There are $g = n - r$ infinite eigenvalues.

□

The structure of the pencil is given by the partition $n = g + r$ for any $z \notin F(A,E)$: it depends only on $r = \operatorname{rank} E$ but *not* on A. This is the generic situation when $F(A,E)$ is a discrete set. When z is close to being a frontier point, then some of the finite eigenvalues of the pencil are large.

When $z \in F(A,E)$, there is an abrupt change in structure of the pencil $(A - zI) + tE$. Let

$$a_z, \quad 1 \leq a_z \leq r \tag{4.3.2}$$

be the algebraic multiplicity of $0 \in \sigma(M_z)$, for $z \in F(A,E)$.

Proposition 4.3.2 *For $z \in F(A,E)$, the structure of the pencil $(A - zI) + tE$ is determined by the partition $n = (g + a_z) + (r - a_z)$ which depends on z in $F(A,E)$.*

Proof. For $z \in F(A,E)$, there is at least one $i \in \{1, \ldots, r\}$ such that $0 = \mu_{iz} \in \sigma(M_z)$ which is equivalent to $|t_i| = 1/|\mu_{iz}| = \infty$. Thus for a_z defined in (4.3.2), the strictly equivalent structure (4.3.1) of the pencil $(A - zI) + tE$ is determined by the partition $n = (g + a_z) + (r - a_z)$ which shows how the structure of the pencil $(A - zI) + tE$ depends on a_z (and of course, on z in $F(A,E)$).

□

According to the Proposition 4.3.2, if for example, the critical set $C(A, E)$ is not empty and $z \in C(A, E)$, then $a_z = r$ and the pencil has no finite eigenvalue. When the frontier set is discrete, there are at most $n-r$ frontier points z in $re(A)$ where the structure of the pencil depends on A. In particular, the number $r - a_z$ of finite eigenvalues for the regular pencil is always smaller than the generic value r, the rank of E. These frontier points signal a tight algebraic coupling between A and E such that $R(\infty, z)$ does not exist (there is no analyticity at ∞).

Example 4.3.1 *Let*

$$A = \begin{bmatrix} 1 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 2 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \text{ and } E = \begin{bmatrix} 0 & 4 & 1 & 2 & 0 & 0 \\ -1 & 3 & 4 & 2 & 0 & 1 \\ 0 & 4 & 1 & 2 & 0 & 0 \\ -1 & 4 & 3 & 3 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 & 0 & 0 \end{bmatrix}.$$

Now, one can use (1.6.1) to get $E = UV^H$ for

$$U = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \text{and } V = \begin{bmatrix} 0 & 0 & -1 \\ 3 & 1 & 0 \\ 2 & -1 & 2 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Here, $\operatorname{rank} E = 3 = r$, $\det(V^H U) = -15 \neq 0$ *and* $0 \in \sigma(E)$ *is semi-simple with the geometric multiplicity* $g = 3$. *Then*

$$\pi(z) = z(z^5 - 2z^4 - 4z^3 + 1),$$

and

$$\operatorname{detQ}(z) = -(15z^3 + 8z^2 - 8z + 1)(z^5 - 2z^4 - 4z^3 + 1)^2 z^2 = (\pi(z))^2 \hat{\pi}(z),$$

for $\hat{\pi}(z) = -(15z^3 + 8z^2 - 8z + 1)$.

This shows that $F(A, E) = \{-1.0828,\ 0.1568,\ 0.3926\}$. *These points are denoted by* f_k, $k = 1, 2, 3$. *The algebraic multiplicity of* $0 \in \sigma(M_{z_k})$, *for each frontier point* f_k, *is* $a_{f_k} = 1$, $k = 1, 2, 3$. *See Table 4.1 below. Therefore according to the Proposition 4.3.2, the structure of the pencil* $(A - zI) + tE$ *for* $z \in F(A, E)$ *is determined by the partition* $n = 6 = (g+1)+(g-1) = 4+2$. *This means that the pencil* $(A - zI) + tE$, *for* $z = f_k \in F(A, E)$, $k = 1, 2, 3$, *is strictly equivalent to the* 2×2 *block matrix corresponding to the partition* $6 = 4+2$

$$\left[\begin{array}{cccc|cc} 1 & t\varepsilon & 0 & 0 & & \\ & 1 & t\varepsilon & 0 & & \\ & & 1 & t\varepsilon & & \\ & & & 1 & & \\ \hline & & & & \frac{1}{\mu_{1f_k}} & \varepsilon \\ & & & & & \frac{1}{\mu_{2f_k}} \end{array}\right] \tag{4.3.3}$$

where ε *represents* 0 *or* 1 *and* $\mu_{jf_k} \in \sigma(M_{f_k})$, $j = 1, 2$, $t \in \mathbb{C}$.

$f_k \in F(A,E)$	μ_{1f_k}	μ_{2f_k}	μ_{3f_k}
$f_1 = -1.0828$	2.6680	-4.0644	-3.3422×10^{-15}
$f_2 = 0.1568$	$2.6864 + 3.9821i$	$2.6864 - 3.9821i$	1.0192×10^{-15}
$f_3 = 0.3926$	$0.3082 + 2.1113i$	$0.3082 - 2.1113i$	-4.9706×10^{-15}

Table 4.1: The computed values of μ_{jf_k}, $j = 1, 2, 3$ for each $f_k \in F(A, E)$

For $z \notin F(A, E)$, *the pencil* $(A - zI) + tE$ *is strictly equivalent to the* 2×2 *block matrix corresponding to the partition* $6 = 3 + 3$

$$\left[\begin{array}{ccc|ccc} 1 & t\varepsilon & 0 & & & \\ & 1 & t\varepsilon & & & \\ & & 1 & & & \\ \hline & & & \frac{1}{\mu_{1z}} & \varepsilon & 0 \\ & & & & \frac{1}{\mu_{2z}} & \varepsilon \\ & & & & & \frac{1}{\mu_{3z}} \end{array}\right] \tag{4.3.4}$$

where ε *represent* 0 *or* 1 *and* $\mu_{jz} \in \sigma(M_z)$, $j = 1, 2, 3$, $t \in \mathbb{C}$.

To visualize the change at frontier points, it is useful to consider the map $z \mapsto \max_{1 \le i \le 3} \frac{1}{|\mu_{iz}|}$ *for* $z \in \mathbb{C}$.

Since the 3 *frontier points are real, we plot in Figure 4.1* $\max_{1 \le i \le 3} \frac{1}{|\mu_{iz}|}$ *as* z *is maintained real and varies in* $[-1.5\ 1.5]$ *to indicate how the corresponding eigenvalue,* $\frac{1}{\mu_{iz}}$, *for the pencil* $(A - zI) + tE$ *escapes to infinity at* $z = f_k \in F(A, E)$, $k = 1, 2, 3$.

The 2D *and* 3D *versions of the frontier portrait* $\phi_2 : z \mapsto \rho(M_z^{-1})$ *related to this example will be displayed in Example 7.2.5 of Chapter* 7.

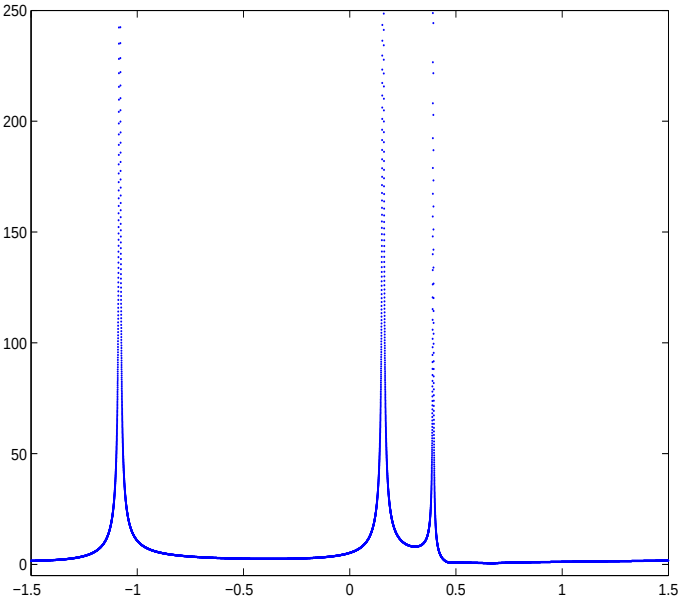


Figure 4.1: $\max_{1\leq i\leq 3} \frac{1}{|\mu_{iz}|}$ versus $z \in [-1.5\ 1.5]$

△

4.4 $z = 0 \in re(A)$: a reduction method for solving GEVP

In this section, we determine the structure of the regular pencil $A + tE$ when A is nonsingular, by applying the theory of Section 4.3 for $z = 0$ fixed in $\mathbb{C}$.

Let $A, E \in \mathbb{C}^{n\times n}$, $r = \operatorname{rank} E < n$, E is written in the form $E = UV^H$ for U and V defined in (1.6.1) and A is *nonsingular*. An application of the lemma 3.2.1 for this case and where $0 = z \notin \sigma(A)$ states

$$\det(A + tE) = \det(A)\ \det(I_r - tM), \tag{4.4.1}$$

where $M_0 = M = -V^H A^{-1} U$ is of order r.

When the matrix A is nonsingular, a direct application of the Propositions 4.3.1 and 4.3.2 for the matrix pencil $A + tE$ is as follows.

Proposition 4.4.1 *Let $A \in \mathbb{C}^{n\times n}$ be nonsingular, $E \in \mathbb{C}^{n\times n}$ has rank $r < n$, and $A + tE$ is a regular matrix pencil. The $r \times r$ matrix $M = -V^H A^{-1} U$ is defined in (4.4.1).*

(a) When $0 \notin \sigma(M)$*, the pencil* $A+tE$ *is strictly equivalent to* 2×2 *block matrix*

$$\left[\begin{array}{ccc|ccc} 1 & t\varepsilon & 0 & & & \\ & \ddots & \ddots & & & \\ & & 1 & & & \\ \hline & & & \frac{1}{\mu_1} & \varepsilon & 0 \\ & & & & \ddots & \ddots \\ & & & & & \frac{1}{\mu_r} \end{array}\right] \tag{4.4.2}$$

corresponding to the partition $n = g + r$*, where* ε *represents* 0 *or* 1 *and* $\mu_i \in \sigma(M)$*,* $i = 1, \ldots, r$*,* $t \in \mathbb{C}$*.*

(b) When $0 \in \sigma(M)$ *with algebraic multiplicity* $1 \leq a_0 \leq r$*, then the pencil* $A + tE$ *has* $r - a_0$ *finite eigenvalues.*

Proof. For $0 = z \in re(A)$, $(A - zI) + tE = A + tE$.

□

The Proposition 4.4.1 offers a reduction method for solving the following generalized eigenvalue problem, GEVP,

$$sp(A, E) = \{t \in \mathbb{C} : \det(A + tE) = 0\} \tag{4.4.3}$$

associated with the regular pencil $A+tE$, where the matrix A is nonsingular and $\operatorname{rank} E = r < n$: when the matrix M is nonsingular, GEVP (4.4.3) has $r < n$ finite eigenvalue which are $sp(A, E) = \sigma(M^{-1})$, and when $0 \in \sigma(M)$, GEVP (4.4.3) has less than r finite eigenvalues, which are the inverse of the *nonzero eigenvalues* of the $r \times r$ *ordinary eigenvalue problem*

$$Mx = \mu x \quad 0 \neq x \in \mathbb{C}^r. \tag{4.4.4}$$

Therefore, finding the generalized eigenvalues in (4.4.3) of order n, is equivalent to finding the non zero eigenvalues of the matrix M of order $r = \operatorname{rank} E$.

This may be computationally effective when n is large but $r \ll n$. To the best of my knowledge, the SVD decomposition for E has not been used in the currently available software for GEVP [3, 1, 10].

Chapter 5

Homotopic backward analysis, I Basic concepts

5.1 Introduction about backward analysis

Let $A \in \mathbb{C}^{n\times n}$ be a given matrix , z arbitrary in $\mathbb{C}$. We consider the problem (P): *find* $\triangle A \in \mathbb{C}^{n\times n}$ *such that* $A + \triangle A - zI$ *is singular.*

That is

(P): *find* $\triangle A$ *such that* z *is an eigenvalue of* $A + \triangle A$.

In other words, given z and A what are the modifications ΔA of A such that z is an exact eigenvalue of $A + \Delta A$? If $\Delta A = 0$, then $z \in \sigma(A)$ is an exact eigenvalue for A itself.

With no further assumption on $\triangle A \in \mathbb{C}^{n\times n}$, the problem (P) has an infinity of solutions, and one looks for modifications ΔA which are small in some sense.

In section 5.3, the problem (P) is considered in the framework of the classical normwise backward analysis which looks for $\triangle A$ with minimum norm [23].

On the other hand, in homotopic deviation theory [5, 16, 18, 30, 21], $\triangle A$ has a **prescribed structure** E such that $\triangle A = tE\,, t \in \mathbb{C}$,and $E \in \mathbb{C}^{n\times n}$ being fixed. In this case z is the eigenvalue of at most n matrices $A + t_k E$, $k = 1, .., r \leq n$, with $t_k \in \hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$. This is the subject of the section 5.4.

Modifications $\triangle A$ of A with a prescribed structure seem to play an important role in our current understanding of the evolution of living organisms [39].

The detailed comparison of the two backward analyses, normwise and homotopic, shows that the latter is computationally much richer than the first [5, 4, 2, 29, 18, 21]. In section 5.2, some general tools related to backward analysis are presented.

5.2 Theoretical tools for inexact computation

Finite precision computation requires the assessment of the computed result with respect to machine precision. Similarly, in exact computation, when the operator, for example, is replaced by a discrete approximation, one also wants, in Theoretical Numerical Analysis, to assess the validity of the approximate solution (computed exactly) with respect to the truncation error. Both problems can be treated within the unique framework of backward analysis which is based on the simple but powerful notion of backward error developed by Givens and Wilkinson for round-off errors in the late 1950s [23, 54].

5.2.1 Classes of modification ΔA

The validity of the conclusions of any backward stability analysis strongly depends on the adequacy of the class of modification to represent the phenomenon which is the source of modifications. In fact while software developers focus on a class of modification appropriate to represent faithfully the modifications generated by finite precision, physicists may be more interested in the one generated by measurement uncertainties on the data or by the variation of specific parameters in the model.

5.2.2 Norms

After deciding on the data to be modified and the class of modifications to be applied, we have to choose a norm to measure the modifications on the data and their effect on the solution.
In mathematics, the distance of the approximate solution $\tilde{x}$ to the exact solution x is measured by the absolute norm $\|\tilde{x} - x\| = \|\triangle x\|$, the ideal being that $\|\triangle x\| \to 0$ as the source of modifications vanishes.

On the other hand, numerical analysis and physicists most often consider a relative formulation of the norm $\frac{\|\triangle x\|}{\|x\|}$ or $\frac{\|\triangle x\|}{\|\tilde{x}\|}$ (with x or $\tilde{x}$ fixed) instead of $\|\triangle x\|$ such that, if possible, this fraction be small with respect to some threshold. The threshold can be chosen as machine precision or the level of uncertainty in the data. (The reason is that in numerical software as in physics, often, it is the *relative* assessment of an approximate solution $\tilde{x}$ which makes sense, not an absolute one). Such a strategy for measuring relative variations has been systematically developed in [23] under the generic name of *scaling*.

5.2.3 Scaling

Scaling consists in applying some linear transformation on the data and the solution. For example, one considers the transformed data $\hat{z} = S_1 z$ and the transformed solution $\hat{x} = S_2 x$, where S_1 and S_2 are linear transformations. So $x = g(z)$

becomes $\hat{x} = S_2 \; o \; g \; o \; S_1^{-1}(\hat{z})$ if we assume that S_1 is invertible: The nonlinear map g is transformed into $\hat{g} = S_2 \; o \; g \; o \; S_1^{-1}$.

Here, we describe three classes of modifications, mainly the componentwise, normwise and homotopic modification which are used in numerical analysis. Let us consider $A \in \mathbb{C}^{n \times n}$, and the modification $\triangle A$.

1. For the componentwise case (upperscript C), $\triangle A \in \mathbb{C}^{n \times n}$ with the scaled norm satisfies
$$\|\triangle A\|^C = \max_{ij} \frac{|\triangle a_{ij}|}{b_{ij}}$$
where $B = (b_{ij})$ is a matrix having a prescribed structure ($b_{ij} \geq 0$ and, if $b_{ij} = 0$, then $\triangle a_{ij} = 0$). The norm is absolute when the matrix B is such that $b_{ij} = 1$ if $a_{ij} \neq 0$ and $b_{ij} = 0$ otherwise. It is relative when $B = |A|$.

2. For the normwise case (upperscript N), $\triangle A \in \mathbb{C}^{n \times n}$ with the scaled norm satisfies
$$\|\triangle A\|^N = \frac{\|\triangle A\|}{\alpha}$$
where $\| \cdot \|$ is any subordinate norm. The norm is absolute if $\alpha = 1$ and relative if $\alpha = \|A\|$.

   For example, a normwise modification such that $\|\triangle A\|^N = \varepsilon$ can be obtained with $|a_{ij}| = \varepsilon\alpha$ where in this case all the components of A can be perturbed.

3. For the homotopic case (upperscript H), $\triangle A$ is such that $\triangle A = tE$ where E is the given deviation matrix, $t \in \mathbb{C}$. It is normed by choosing $\|\triangle A\|^H = \frac{|t|}{\alpha}$. The formulation is absolute if $\alpha = 1$ and relative if $\alpha = \frac{\|A\|}{\|E\|}$.

Once the data to be perturbed have been selected, we have to choose a metric to quantify the size of the perturbation on data and solution [23].

In mathematics, convergence to the exact solution is characterized by the condition $\|\triangle A\| \to 0$ for some norm. By comparison, in numerical analysis, in physics, and in all experimental sciences there is unavoidable uncertainty on the data. Therefore, the best one can do to assess the validity of the computed solution is to satisfy a relative criterion such as $\frac{\|\triangle A\|}{\|A\|}$ small compared with the level of uncertainty on the data.

5.3 Normwise backward analysis for the eigenvalue problem

5.3.1 Normwise backward error

We define the backward error associated with a normwise analysis for the eigenvalue problems. For any given z in $re(A)$, $\triangle A \in \mathbb{C}^{n \times n}$ is arbitrary and (P) has an

infinity of solutions. In the normwise analysis we use scaled norms.

Definition 5.3.1 *Let z be a given point in $re(A) = \mathbb{C}\backslash\sigma(A)$ for the eigenproblem $Ax = \lambda x$. The normwise backward error β_z^N corresponding to z is defined by*

$$\begin{aligned}\beta_z^N &= \min\{\varepsilon;\ \ \exists\ u\ \neq 0,\ (A+\triangle A)u = zu \text{ such that } \|\triangle A\|^N \leq\ \varepsilon\} \\ &= \min\{\|\triangle A\|^N; A+\triangle A - zI \text{ is singular}\}. \end{aligned} \tag{5.3.1}$$

There are some equivalent formulations for β_z^N [23]: β_z^N is the distance of $A - zI$ to singularity. Equivalently, it measures, in terms of $\|\triangle A\|^N \in \mathbb{R}^+$, the smallest amount by which z fails to be an eigenvalue of A.

Figure 5.1 shows how β_z^N indicates the frontier for the two distinct groups of ΔA, for which $(A + \Delta A - zI)^{-1}$ exists or not. $(A + \Delta A - zI)^{-1}$ exists for $\|\Delta A\| < \beta_z^N$, and it does not exist for $\|\Delta A\| \geq \beta_z^N$.

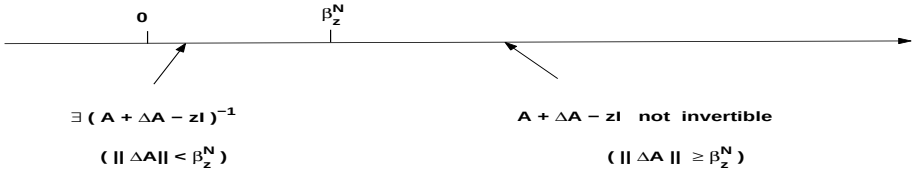


Figure 5.1: Existence and non existence of $(A + \Delta A - zI)^{-1}$ depending on $\|\Delta A\|$ compared with β_z^N

Remark: Using the notation of subsection 5.2.3, the formulation for β_z^N is absolute if $\alpha = 1$, and is relative if $\alpha = \|A\|$.

Lemma 5.3.2 *[23] The normwise backward error associated with $z \in re(A)$, for the eigenproblem $Ax = \lambda x$ can be computed as*

$$\beta_z^N = \frac{1}{\alpha\|(A - zI)^{-1}\|}\ . \tag{5.3.2}$$

Proof. Let us consider a non zero $u \in \mathbb{C}^n$ such that $(A + \triangle A)u = zu$. It means that $Au - zu = -\triangle Au$ and so,

$$\|(A - zI)u\| \leq \|A - zI\|\|u\| = \|\triangle A\|\|u\| \leq \alpha\varepsilon\|u\|. \tag{5.3.3}$$

Now, considering this fact that for $z \notin \sigma(A)$, $\frac{1}{\|(A-zI)^{-1}\|} \leq \|(A - zI)\|$, evidently (5.3.3) implies the desired result.

□

5.3.2 Normwise spectral portrait of A

During the decade of the 90 s, special attention has been given to the map $\psi : z \longmapsto \|(A - zI)^{-1}\|$, defined for $z \in re(A)$. The reason is that the normwise backward error defined in (5.3.2) for $z \in re(A)$, measures by how much z fails to be an eigenvalue of A. The map ψ is called the *normwise spectral portrait* of A. See the many plots in [23].

5.4 Homotopic backward analysis for the eigenvalue problem

Let $r = \operatorname{rank} E$ and μ_{kz}, $k = 1, \ldots, \hat{r}_z \leq r$ denote the nonzero eigenvalues of the matrix M_z defined in (2.4.2) for $\hat{r}_z = r - a_z$, where $a_z \geq 0$ is algebraic multiplicity of $0 \in \sigma(M_z)$ (> 0 for $z \in F(A, E)$). The same z is an eigenvalue of the $\hat{r}_z$ (not necessarily distinct) matrices $A + t_k E$ with $t_k = \frac{1}{\mu_{kz}} \in \mathbb{C}$, $k = 1, \ldots, \hat{r}_z \leq r$. $\hat{r}_z$ depends on z when $z \in F(A, E)$. Otherwise $\hat{r}_z = r$ for $z \in re(A) \backslash F(A, E)$.

z is an exact eigenvalue for $\hat{r}_z \leq r$ matrices $A(t_k)$, where $|t_k| < \infty$. If one is willing to consider $t \in \hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$, then z is an exact eigenvalue for *exactly* r matrices $A(t)$ with $|t| \leq \infty$ (see section 5.4.1).

Amongst these $\hat{r}_z$ matrices, one at least is closest to A, it is defined by t_* of minimum modulus:

$$|t_*| = \min_k |t_k| = \frac{1}{\max_k |\mu_{kz}|}. \tag{5.4.1}$$

$A(t_*)$ is a matrix of the family $A(t)$ closest to A, if the distance is measured by the modulus of t.

Similarly, amongst the $\hat{r}_z$ matrices, one at least is furthest from A, it is defined by t^* of maximum modulus:

$$|t^*| = \max_k |t_k| = \frac{1}{\min_k |\mu_{kz}|}. \tag{5.4.2}$$

$A(t^*)$ is a matrix of the family $A(t)$ furthest from A, if the distance is measured by the modulus of t.

There is not a unique way to define the homotopic distance. For example, one can choose the following.

Definition 5.4.1 *The homotopic distance to spectral singularity of* z *is*

$$|t_*| = \min\{|t|; \quad z \ \textit{is an eigenvalue of} \quad A + tE\}.$$

It is also called **homotopic backward error** *which we denote it by* β_z^H, *where the upperscript* H *refers to homotopic.*

Using the definition of $|t_*|$ in (5.4.1), one has

$$\beta_z^H = |t_*| = \frac{1}{\rho(E(A - zI)^{-1})} = \frac{1}{\rho(M_z)}, \tag{5.4.3}$$

where $|t_*|$ measures by how much z, which is an eigenvalue of $A(t_*) = A + t_*E$, fails to be an eigenvalue of A.

The bound

$$\rho(E(A - zI)^{-1}) \leq \|E\|\|(A - zI)^{-1}\| \leq \|(A - zI)^{-1}\|$$

valid for all E such that $\|E\| \leq 1$ is useful to compare homotopic and normwise backward errors.

5.4.1 Homotopic deviations from singularity at $z \in re(A)$

For $z \in re(A)$, there are at most r ways (or matrices) by which $A - zI$ fails to be singular: these are the deviations $t_{kz}E$, with $t_{kz} = \frac{1}{\mu_{kz}}$, for $\mu_{kz} \neq 0$, $k = 1, \ldots, \hat{r}_z \leq r$.

We assume that $\mu_{kz} = 0 \Leftrightarrow |t_{kz}| = \infty$. When $z \notin F(A, E)$, there are exactly r finite deviations from singularity, namely $t_{kz}E$, for $t_{kz} \in \mathbb{C}$. Below, we consider that $t \in \hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$. Observe that the situation in homotopic deviation is *very different* from that in normwise analysis. There is a *finite* ($\leq r$) number of ways to achieve singularity. Whereas in normwise analysis, the set of modifications is not countable. This difference makes the homotopic backward analysis computationally *much richer* than its normwise counterpart.

5.5 Two kinds of homotopic backward analyses at $z \in re(A)$ for $t \in \mathbb{C}$

Let z be fixed in $re(A)$. When we consider the deviation tE with the complex parameter t, it is important to introduce a qualitative difference on deviations, deriving from $t = |t|e^{i\theta}$ with $|t| = h$, and $\theta \in [0, 2\pi[$. The existence of the

resolvent $R(t,z)$, defined in (2.2.3), can be analyzed from 2 points of view: the modulus or argument of t.

Let be given the $\hat{r}_z$ points t_{kz} in $\mathbb{C}$ where $R(t,z)$ does not exist. Then, $R(t,z)$ exists for t with $|t| \neq |t_{kz}|$, or, for θ which satisfies $\theta \neq \theta_{kz}$ *mod* 2π. This leads to 2 types of homotopic backward analyses: *metric* or *angular*, and two ways to partition the complex plane in which t varies.

5.5.1 Homotopic metric rings related to $z \in re(A)$, $r > 1$

For $t = |t|e^{i\theta}$ $(|t| = h)$, the *metric homotopic error* is equal to the distance from the set of deviations $\{e^{i\theta}E,\ 0 \leq \theta < 2\pi\}$ [4, 2]. For z fixed in $re(A)$, there are at most r metric homotopic errors: $|t_{kz}| = \frac{1}{|\mu_{kz}|}$, $k = 1, \cdots, r$, bounded or not.

Definition 5.5.1 *Let z be given in $re(A)$. The circles*

$$O(h_k) = \{t = h_k e^{i\theta};\ \textit{for } h_k = |t_{kz}| \textit{ fixed, and } \theta \in [0, 2\pi[\ \},\ k = 1, \cdots, \hat{r}_z \leq r, \tag{5.5.1}$$

represent the sets of t with equal metric homotopic error h_k.

These circles define (at most) $\hat{r}_z - 1$ metric rings, in which $R(t,z)$ exists for every t which does not belong to any of the (at most) r circles with radii $|t_{kz}|$, $k = 1, \ldots, \hat{r}_z$. When $z \notin F(A,E)$, $\hat{r}_z = r$ and the r circles define $\hat{r}_z + 1$ regions in which $R(t,z)$ exist. In this case, $R(t,z)$ is analytic in t inside of the smallest circle and outside of the largest one (around 0, and ∞). $R(t,z)$ exists in the $r-1$ circular rings for $r > 1$.

See the 5 metric rings and the 2 regions of analyticity on Figure 5.2 for $r = 6$. Inside of the smallest circle and outside of the largest one, lie the regions where $R(t,z)$ is analytic in t, they are hatched.

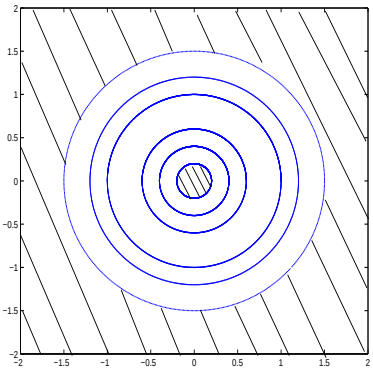


Figure 5.2: The 5 metric rings and the 2 regions of analyticity for $R(t,z)$, $r = 6$, $t \in \mathbb{C}$, z given in $re(A)$

A numerical illustration is given in the following example.

Example 5.5.1 *We are given the following matrices,*

$$A = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 2 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 9 & 1 & 2 \\ 1 & 2 & 5 & 2 & 0 & 2 \end{bmatrix}, U = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 2 & 2 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 8 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, V = \begin{bmatrix} 0 & 2 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix}$$

where $E = UV^H$ *and* $r = \operatorname{rank} E = 4 < n = 6$. *Given* z *in* $re(A) \backslash F(A, E)$, *there are exactly* 4 *points* $t_{kz} = \frac{1}{\mu_{kz}}$ *in* $\mathbb{C}$ *where* $R(\frac{1}{\mu_{kz}}, z)$ *does not exist,* $k = 1$ *to* 4.

We choose z *in* $re(A) \backslash F(A, E)$ *as* $z_\star = 17.9763 + 2.3561i$, *the values and the magnitudes of* $t_{kz} = 1/\mu_{kz}$, *for* $k = 1, \ldots, 4$ *are shown in table 5.1. We drop the index* z. *They satisfy* $|t_1| < |t_2| < |t_3| < |t_4|$ *and define* 3 *rings. These* 4 *points are plotted by* $*$ *with appropriate color in Figure 5.3. The* 3 *metric rings are shown in Figure 5.3.*

	t_k	$\lvert t_k \rvert$	color
k=1	1.0781+0.2185i	1.1	green
k=2	14.9463+2.0656i	15.0883	magenta
k=3	-26.2868-1.6567i	26.3390	blue
k=4	-46.19485-9.7705i	47.2204	red

Table 5.1: The values and the magnitudes of $t_k = 1/\mu_{kz}$, $k = 1, \ldots, 4$ for $z = z_\star$

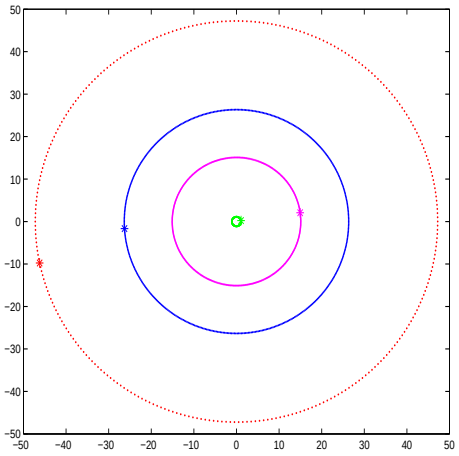


Figure 5.3: The 4 circles for $z_\star = 17.9763 + 2.3561i$, the 3 metric rings and the 2 regions of analyticity for $t \in \mathbb{C}$

In this example $R(t, z_\star) = R(t, 17.9763 + 2.3561i)$ exists for every t which does not belong to the 4 circles with radii $|t_{kz}|$, $k = 1, \ldots, 4$. The 4 circles define 5 regions in which $R(t, z)$ exist. It is analytic in t inside of the smallest circle and outside of the largest one (around 0, and ∞). See Figure 5.3.

$\triangle$

5.5.2 Homotopic angular sectors

The angular homotopic error for θ or ($e^{i\theta}$) is the distance from the set of deviations $\{hE,\ h \in \mathbb{R}^+\}$ [4, 2]. There are at most r *angular* homotopic errors : $e^{i\theta_{kz}}$ (or θ_{kz} if one considers the logarithm), $k = 1, \ldots, \hat{r}_z \leq r$ where θ_{kz} are the arguments of $t_{kz} = \frac{1}{\mu_{kz}} \pmod{2\pi}$ for $\mu_{kz} \neq 0$.

Observe that the argument of $\mu_{kz} = 0$ is undetermined.

Definition 5.5.2 *Let z be given in $re(A) \backslash F(A, E)$. The oriented half lines*

$$S(\theta_k) = \{t = he^{i\theta_{kz}},\ \textit{for } \theta_{kz} \textit{ fixed in } [0, 2\pi[, \textit{and } |t| = h \in \mathbb{R}^+\},\ k = 1, \cdots, \hat{r}_z \leq r, \tag{5.5.2}$$

represent the sets of t with equal angular homotopic error θ_{kz}.

The $\hat{r}_z$ points t_{kz} define (at most) $\hat{r}_z$ angular sectors.

Example 5.5.2 *We consider the family $A + tE - zI$ defined in Example 5.5.1, where z is fixed at the same value $z_\star = 17.9763 + 2.3561i$. The angles $\theta_k = \theta_{kz}$ in $t = |t|e^{i\theta_k}$ for $k = 1, \ldots, 4$ together with the corresponding colors used in the figure 5.4 are shown in the table 5.2.*

In this example $R(t, z_\star)$ exists for all $\theta \neq \{\theta_{1z}, \cdots, \theta_{4z}\}$ (mod 2π) which are the 4 different arguments of $t_k = 1/\mu_{kz}$, $k = 1, \ldots, 4$. They define 4 different angular sectors in which $R(t, z_\star)$ exists. See figure 5.4.

	θ_k	color
k=1	0.2000	green
k=2	0.1373	magenta
k=3	-3.0787	blue
k=4	-2.9332	red

Table 5.2: The values and the corresponding colors of θ_k, $k = 1, \ldots, 4$ for $z = z_\star$

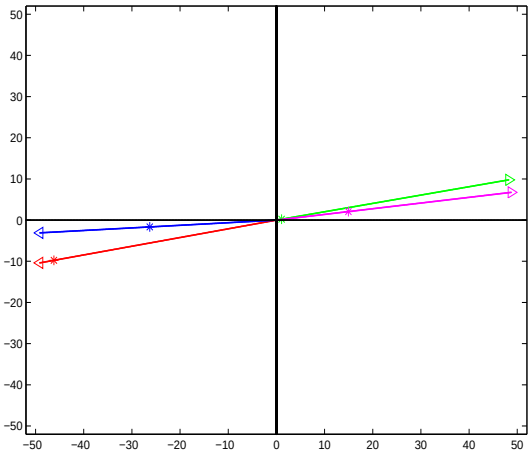


Figure 5.4: The $r = 4$ angular sectors for $z = z_\star$

△

5.6 Homotopic normwise and spectral portraits

The map $\phi_0 \ : \ z \mapsto \|M_z\|$ is the homotopic analogue of the popular normwise portrait map $\psi \ : \ z \mapsto \|(A - zI)^{-1}\|$, [23]. In ϕ_0, the matrix $(A - zI)^{-1}$ of order n is replaced by M_z of order $r < n$.
HD theory suggests to complement the qualitative study of $z \mapsto M_z$ by the alternative spectral portrait

$$\phi_1 \ : \ z \mapsto \rho(M_z),$$

where $\|\cdot\|$ is replaced by $\rho(\cdot)$.

An important consequence when $r > 1$ is that ϕ_1 can localize the critical points ($\rho = 0$) when they are isolated, whereas the normwise spectral portrait cannot [21]. To distinguish between the two portraits ϕ_0 and ϕ_1, we call ϕ_0 (resp. ϕ_1) the normwise (resp. spectral) portrait, see Chapter 7.

Definition 5.6.1 *A map $z \in \mathbb{C} \mapsto f(z) \in \mathbb{R}^+$ is said to have a peak (resp. a well) at $z \in \mathbb{C}$ iff $f(z)$ is not defined (resp. $f(z) = 0$).*

To increase the visibility, in practice, we scale the plot of the normwise (resp. spectral) portrait by $\phi_0 \ : \ z \mapsto log_{10}(\|M_z\|)$ (resp. $\phi_1 \ : \ z \mapsto log_{10}(\rho(M_z))$). Since $-\infty \leq log_{10}(a) < 0$ for $0 \leq a < 1$, hence a part of the plot $\phi_0 \ : \ z \mapsto log_{10}(\|M_z\|)$ (resp. $\phi_1 \ : \ z \mapsto log_{10}(\rho(M_z))$) which corresponds to $\|M_z\| < 1$ (resp. $\rho(M_z) < 1$) appears below the plane $a = 0$ in the 3D cartesian coordinates (x, y, a).
In general, the plot of ϕ_0 (resp. ϕ_1) has peaks at the eigenvalues of A (resp. has peaks at the eigenvalues of A and wells at certain frontier points). See in Chapter 7

a detailed discussion about the notions of peaks and wells and their role in detecting the spectrum of A, the frontier points, critical points and the set Lim.

Part II

HD in finite precision: computer experiments for a qualitative analysis

Chapter 6

A qualitative study of HD based on the spectral field $t \mapsto \sigma(A(t))$

6.1 Introduction

This Chapter is devoted to a general presentation of computer experiments which can be performed to realize a qualitative study of HD in finite precision. The reason to do this is that the theory is not complete: in practice, the sufficient conditions (Li), or (G) encountered in sections 3.8.2 and 3.9.3 respectively may not be satisfied.

To get information about Lim, we use plots of the vector map $t \mapsto \sigma(A(t))$ consisting of the n eigenvalues $\lambda_i(t)\,,\ i = 1, \ldots, n$ in $\sigma(A(t))\,$. We present, in section 6.2, the two families of curves which have been introduced in [27, 29, 31, 51]. In section 6.2.1, the coloring will be used to show the variation of spectral rays and spectral orbits. In section 6.2.2, we present two meshes of rays and orbits. The different rates of change for the spectral rays are examined in section 6.2.3.

In section 6.3, we introduce a practical method for finding the set Lim, and we discuss the robustness of the convergence $\lim_{|t|\to\infty} \sigma(A(t))$ to finite precision. In the spectral rays, except for the Figure 6.1, we display the eigenvalues of $\sigma(A)$ by $\circledast$ and the points in Lim by $\odot$. The dependence of the convergence on the parameters $\theta = \mathrm{Arg}\, t$ and $h = |t|$ in $t = he^{i\theta} = |t|e^{i\theta}$ will be discussed in section 6.3.4.

6.2 Spectral rays and spectral orbits

In order to visualize the n spectral maps $t \to \lambda_i(t) \in \sigma(A(t))\,,\ i = 1, \ldots, n\,,$ we write the homotopic parameter $t = he^{i\theta}\,,$ with $h = |t| \in \mathbb{R}^+$ and $\theta = \mathrm{Arg}\, t \in [0,\ 2\pi[\,$. By fixing either h or $\theta\,$, and letting the other be a free variable, we form two families of curves in the complex plane that are parameterized by a real parameter. Specifically:

1. the map $h \in \mathbb{R}^+ \mapsto \lambda_j(t),\ j = 1, \ldots, n$ defines the set $\Lambda(\theta)$, of n *spectral rays*, corresponding to $t = he^{i\theta}$ for a fixed θ in $[0,\ 2\pi[$,

2. the map $\theta \in [0,\ 2\pi[\mapsto \lambda_j(t),\ j = 1, \ldots, n$ defines the set $\Sigma(h)$, of (at most) n *spectral orbits*, corresponding to $t = he^{i\theta}$ for a fixed h in $\mathbb{R}^+$.

The rays show, for a fixed θ and increasing h, the trajectory in the complex plane for an eigenvalue of $A(t)$. The rays start at an eigenvalue of A and either diverge to infinity or converge to a limit point of (A, E) (possibly an eigenvalue of A). The orbits are closed curves in the complex plane, since $t = he^{i\theta}$ is 2π-periodic in θ. For a small enough h, each orbit encloses a distinct eigenvalue. Several orbits may enclose the same eigenvalue, depending on its (algebraic) multiplicity. For h large enough, each orbit encloses either a limit point, or the orbit escapes to infinity. Again, several orbits may enclose the same limit point, depending on its multiplicity. For a medium sized h, the same orbit may enclose several eigenvalues. See section 6.2.1. For an early discussion and many examples of spectral rays and orbits, the reader is referred to [27, 29, 31, 51].

6.2.1 A standard color chart to parameterize the variation of h or θ

In this section, we use the same coloring chart to represent the variation of the n *spectral rays* $\Lambda(\theta)$ (resp. *spectral orbits* $\Sigma(h)$) for a fixed value of θ (resp. h) respectively.

In the examples to follow, we use the M-file, *colormap*, in Matlab: it defines the colors (ranging from pure blue to pure red) assigned to the rays and orbits. We use *colorbar* in Matlab, to append a vertical color scale (on the right hand side of each figure).

Let $h = |t| \in [0,\ h_{max}]$ and $\theta \in [0,\ \theta_{max}] = [0,\ 2\pi - \varepsilon]$ for a small $\varepsilon > 0$. We define the one-to-one correspondence between the interval $[0,\ h_{max}]$ (resp. $[0,\ \theta_{max}]$) for a spectral ray (resp. for a spectral orbit) and the vertical color scale which displays the range $[0,\ 64]$ for the colors. When we use the color chart for the spectral rays, we use the following linear equation for finding the values of $h \in [0,\ h_{max}]$ using the values of $c \in [0,\ 64]$,

$$h = (\frac{h_{max}}{64})c.$$

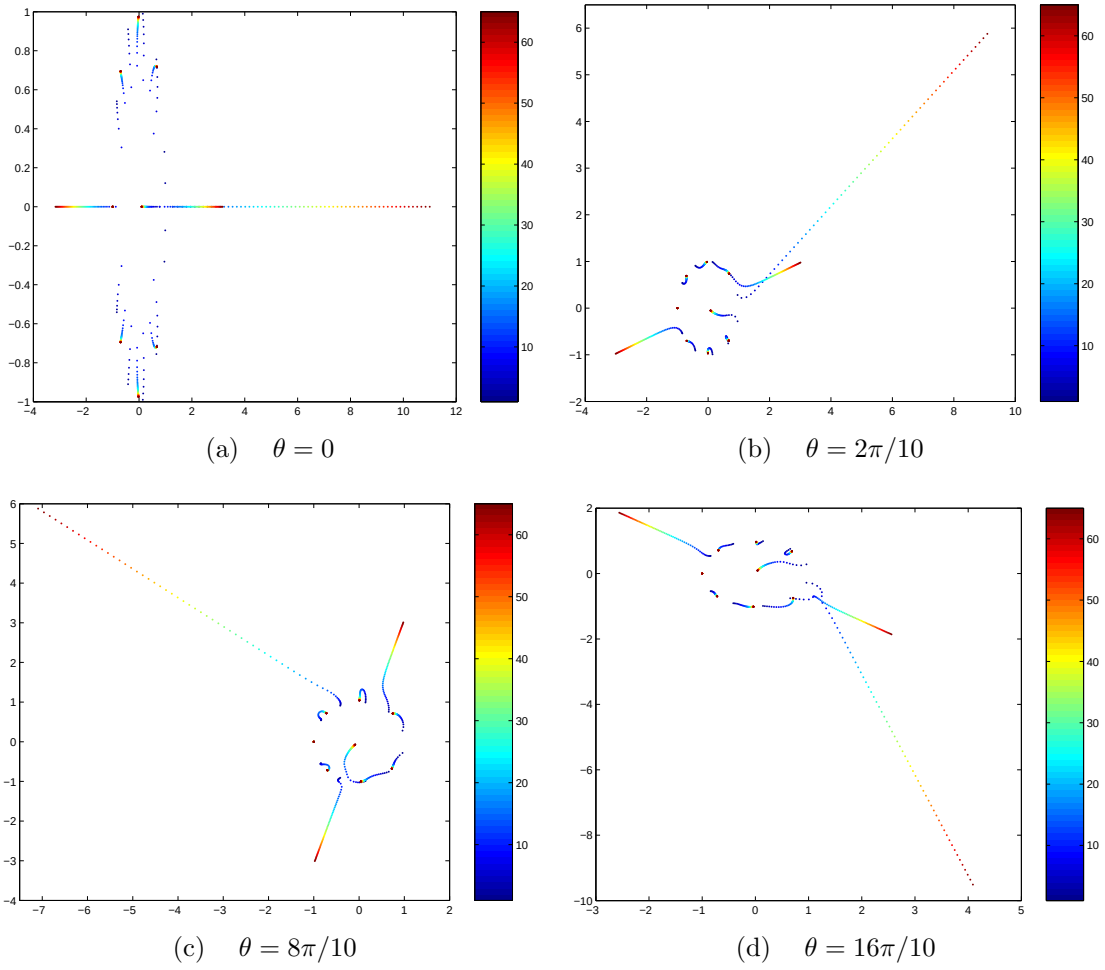
Similarly, when we use the color chart for the spectral orbits, we use the following linear equation for finding the values of $\theta \in [0,\ \theta_{max}]$ using the values of $c \in [0,\ 64]$,

$$\theta = (\frac{\theta_{max}}{64})c,$$

where $0 < \theta_{max} < 2\pi$.

Example 6.2.1 *Let us go back to the Example 3.9.2 where* A *of order* 11 *is the companion matrix associated with* $\pi(z) = z^{11} + 1$, *in upper Hessenberg form, and first column* e_2, *and* $E = UV^T$ *with* $U = [e, e_2]$ *and* $V = [e_{11}, e_3]$ *of rank* 2, $e = [1, \ldots, 1]^T$. *In this Example, we have* $(G) = (\mathrm{Li})$, *and* Lim *consists of the* 8 *points given by* $\mathrm{Lim} = \sigma(\Omega) = Zer(p(z)) \cup \{0\} = F(A, E) \cup \{-1\}$ *for* $p(z) = z^7 + z^6 + \cdots + 1$. *Three eigenvalues escape to* ∞.

On Figures 6.1, we display 4 *different sets of rays corresponding to the* 4 *values* $\theta = 0,\ 2\pi/10,\ 8\pi/10,\ 16\pi/10$ *respectively. Here,* $|t| \in [0,\ 10] = [0,\ h_{max}]$. *Figures 6.1,* (b), (c) *and* (d) *confirm that* 3 *rays escape to* ∞ *with different speeds.*



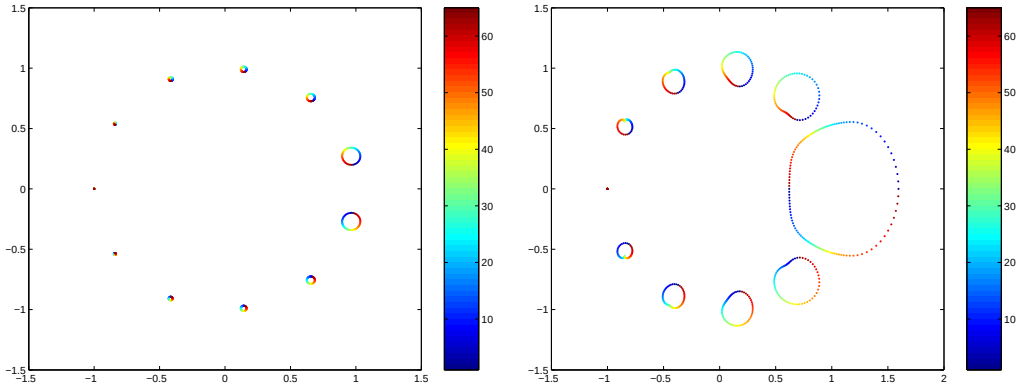
(a) $\theta = 0$ (b) $\theta = 2\pi/10$

(c) $\theta = 8\pi/10$ (d) $\theta = 16\pi/10$

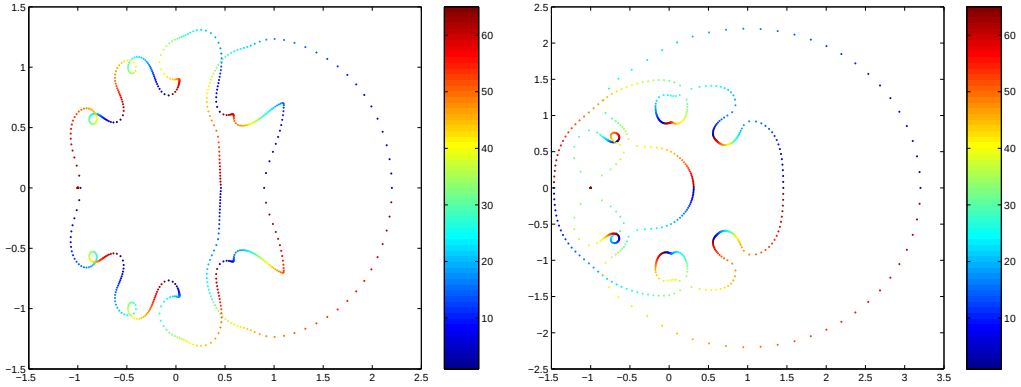
Figure 6.1: Four different sets of 11 rays for $t \mapsto \sigma(A(t))$ where $h_{max} = 10$

On Figures 6.2, we display 4 *different orbits corresponding to the* 4 *values* $h = 0.1,\ 0.6,\ 1.2,\ 2.2$ *respectively. Since* θ *varies in* $[0,\ 2\pi[$, *therefore we let* $\theta \in [0,\ \theta_{max}]$ *with* $\theta_{max} = 2\pi - \varepsilon$ *for* $\varepsilon = 2\pi/63$. *One can see* 11 *orbits for*

$h = 0.1$ but there are less than 11 orbits for the cases $h = 0.6,\ 1.2,\ 2.2$ because they enclose several eigenvalues.



(a) $h = 0.1$ (b) $h = 0.6$



(c) $h = 1.2$ (d) $h = 2.2$

Figure 6.2: Four different sets of orbits for $t \mapsto \sigma(A(t))$

△

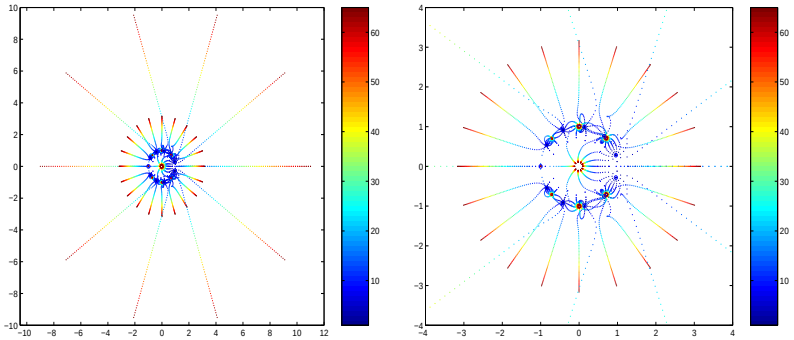
6.2.2 Meshes of rays and orbits

In this section, we use the standard color chart introduced in section 6.2.1 for some different rays $\Lambda(\theta_k)$, $k = 1, \ldots, d_\theta$ (resp. orbits $\Sigma(h_k)$, $k = 1, \ldots, d_h$) which are associated with d_θ (resp. d_h) different values of θ (resp. h). We plot the corresponding sets of rays (resp. orbits) on the same Figure to get what we call a *mesh* of rays (resp. orbits).

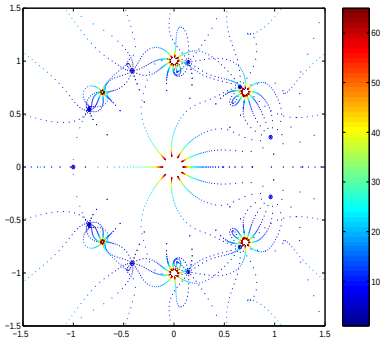
1. **A mesh of rays**

Example 6.2.2 *Let us go back to the Example 6.2.1. We partition the interval* $[0,\ \theta_{max}]$, *with* $\theta_{max} = 2\pi - \frac{\pi}{5}$.

The segment $[0,\ \theta_{max}] = [0,\ 2\pi - \frac{\pi}{5}] = [0,\ \frac{9\pi}{5}]$ *is evenly divided by* 10 *points which realize a discretization of size* $\frac{\frac{9\pi}{5}}{10-1} = \frac{\pi}{5}$. *We plot on Figures 6.3 the* 10 *sets of rays* $\Lambda(\theta_k)$ *corresponding to the* 10 *discrete values* $\theta_k = k\frac{\pi}{5}$, $k = 0, 1, \ldots, 9$. *And we let* $|t| = h$ *vary in* $[0,\ 10]$ *with a discretization step equal to* $10/63 = 0.1587$.



(a) Global mesh (b) Zoom on the mesh



(c) Close-up on the mesh

Figure 6.3: Mesh of rays for 10 values of θ and $|t| = h \in [0,\ 10]$

The 3 *Figures 6.3,* (a), (b) *and* (c) *show the global view and two successive zooms for the mesh according to the description below. Since there are* 3 *escaping rays for each set of spectral rays, therefore there are* 30 *escaping rays for* 10 *sets of spectral rays* $\Lambda(\theta_k)$, $k = 0, \ldots, 9$.

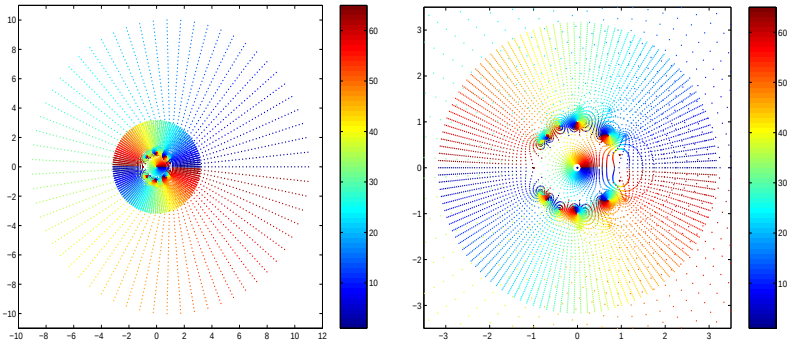
On Figure 6.3, (a), *we can see the global shape of the mesh. Each one of* 10 *rays (which is an escaping ray) corresponds to one of the* 10 *values of* θ *considered in* $[0,\ \theta_{max}]$. *Figure 6.3,* (b), *is a zoom on the middle core of*

the mesh: it is the part of the global mesh which lies in $[-4,\ 4]^2$ *. It shows the* 20 *rays which are again associated with the* 10 *values of* θ *considered in* $[0,\ \theta_{max}]$ *. On Figure 6.3,* (c) *, we display a close-up of the mesh: it is the part of the global mesh which lies in* $[-1.5,\ 1.5]^2$ *and shows the trajectories of the rays starting from pure blue (*$\sigma(A)$*) and ending in pure red at the* l_* *points of* Lim *for large enough* h *.*

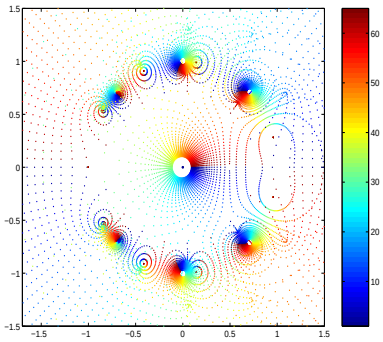
△

2. **A mesh of orbits**

Example 6.2.3 *Let us go back to the example 6.2.1. We evenly divide the segment* $[0, 10]$ *by* 41 *points. They realize a discretization of size* 0.25 *of the segment* $[0, 10]$ *. We plot on Figures 6.4 the* 40 *orbits* $\Sigma(h_k)$ *corresponding to the discrete values* $h_k = (0.25)k$ *,* $k = 1, 2, \ldots, 39, 40$ *.*



(a) Global mesh of orbits (b) Zoom on mesh of orbits



(c) Close-up on mesh of orbits

Figure 6.4: Mesh of orbits

Some parts of Figures 6.4, (a) *and* (b) *, which look like rays, are in fact orbits. This is a consequence of the discretization used for* θ *, which in theory varies*

continuously. On Figures 6.4, (c), we can see how the spectrum of A and the points of Lim *are enclosed by orbits for h small and large respectively.*

Zoom on Figures 6.4, (b) means a close view of the part of main (global) mesh which lies in $[-3.5,\ 3.5]^2$. Close-up on Figures 6.4, (c) means a closer view of the part of main (global) mesh which lies in $[-1.7,\ 1.5] \times [-1.5,\ 1.5]$.

△

6.2.3 Rate of convergence / divergence for $\lambda(t)$: an analysis by colors

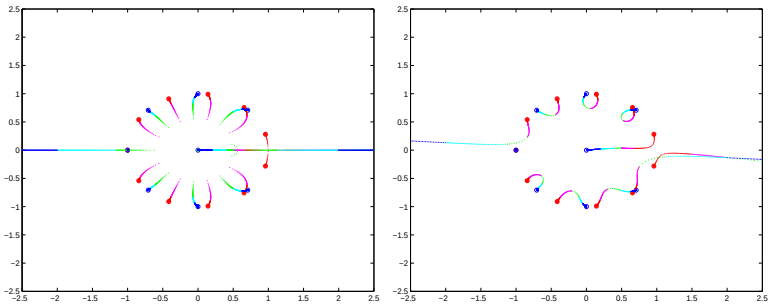
As should be expected, usually the rate of change for $\lambda_i(t)$, $i = 1, \ldots, n$ as $|t|$ varies is not uniform. One way to illustrate this fact is to display all $\lambda_i(t)$, $i = 1, \ldots, n$ with the same color for all i, but varying with the range of $|t|$. For instance, red for $0 < |t| < b_1$, magenta for $b_1 \leq |t| < b_2$, and so on.

Example 6.2.4 *For the Example 3.9.2, the different ranges of $h = |t|$ and the corresponding colors are listed in the table 6.1. The computation of $t = he^{i\theta} \mapsto \sigma(A(t))$, for 3 different values of θ are shown in Figures 6.5, (a), (b) and (c) where the range of $|t|$ and the corresponding colors for all eigenvalues in $\sigma(A(t))$ are the same.*

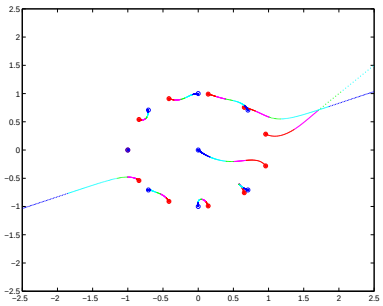
$\|t\|$ varies in	color for $\lambda(t)$
$[0,\ 0.5[$	red
$[0.5,\ 1[$	magenta
$[1,\ 1.5[$	green
$[1.5,\ 4[$	cyan
$[4,\ 300[$	blue

Table 6.1: The 5 intervals for $|t|$ and their associated colors

The three Figures 6.5, (a), (b) and (c) show the different rates of change for $\lambda_i(t)$, $i = 1, \ldots, n$ as $|t|$ varies. The length of each sub-interval associated with red, magenta, green and cyan is uniform but the lengths of the corresponding sub-rays inside $\lambda_i(t)$, $i = 1, \ldots, n$ are different.



(a) $\theta = 0$ (b) $\theta = -\pi/24$



(c) $\theta = \pi/4$

Figure 6.5: $t = he^{i\theta} \mapsto \sigma(A(t))$ for $h \in [0,\ 300]$

△

6.3 Robustness of convergence to finite precision

In finite precision, one expects that the computation of $\lim_{|t|\to\infty} \lambda(t)$ for $\lambda(t) \in \sigma(A(t))$ imposes some limitations on the maximum value of $|t|$ that should be considered, say h_M. In this section, we show that for finding the points of Lim, the determination of $h_M^{(i)}$ for each $\lambda_i(t)$ is very important. If we use $|t| > h_M^{(i)}$, some $\lambda_i(t)$ which are theoretically converging to a finite limit point, diverge from their theoretical limit: they may go to ∞ or to a wrong limit.

How can we determine $h_M^{(i)}$, $i = 1, \ldots, l_* = \text{card Lim} \leq n - r$ for each example? Does it depend on the data of each problem?

In section 6.3.1, we use an example to show that exceeding the specific value $h_M^{(i)}$ may cause wrong results.

In section 6.3.2, we present a heuristic method for finding $h_M^{(i)}$. We will see that the value $h_M^{(i)}$ varies from an example to another. In section 6.3.3, we shall present some cases that require, for some $\lambda_i(t)$, huge values for $h_M^{(i)}$ before convergence manifests itself.

6.3.1 An example where h is too large

Let us look at the following Examples.

Example 6.3.1 *Let A of order 6 be the companion matrix of $\pi(z) = z^6 + 1$, and $E = diag\left[\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, 0, 0, 0, 1\right]$. $\sigma(E) = \{0,\ 1\}$, and $0 \in \sigma(E)$ is defective. $r = \operatorname{rank} E = 2$, with $q = 2$, $n_q = 1$, $r_q = 3$. Using Maple, we find the factorization $\det Q(z) = z^3(z^6 + 1)$ and $\hat{\pi}(z) = z^3$. This shows that $0 \in \sigma(M_0)$ and $F(A, E) = \{0^3\}$. Now we have $\Gamma_1 = 1$, $\Gamma_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$, and $\Omega = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} [1] \begin{bmatrix} 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$.*

Therefore $(G) = (\mathrm{Li})$ and $\{0^3\} = \sigma(\Omega) = \mathrm{Lim}$. Observe that, in this case, $\mathrm{Lim} = F(A, E)$. The computation of $t \mapsto \sigma(A(t))$ for $\theta = \pi/3$ and $0 \le |t| \le 10^2$ shown in Figure 6.6, illustrates the existence of 3 rays converging to 0 and 3 rays escaping to ∞, as is expected from theory.

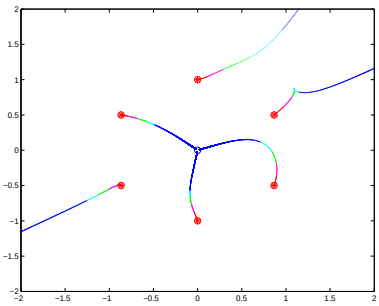


Figure 6.6: $t \mapsto \sigma(A(t))$

△

Example 6.3.2 *Let us consider the example 6.3.1 where we let $|t| > 10^2$. Figures 6.7, and 6.8 display the maps $log_{10}|t| \mapsto log_{10}|\lambda(t)|$ for each of the 3 eigenvalues converging to $0 \in \mathrm{Lim}$, for $|t| \in [\varepsilon,\ 10^{22}]$ (Figure 6.7) and $|t| \in [\varepsilon,\ 10^{33}]$ (Figure 6.8) with $\varepsilon = 5 \times 10^{-2}$. Here $\theta = \pi/3$.*

As we can see on Figures 6.7, and 6.8, the computation for $|t| > 10^{17}$ causes that all 3 $|\lambda(t)|$ increase from their local minimum obtained for $|t| \approx 10^{16}$. The increasing behaviour will continue until $|t| = 10^{32}$. From this value on, one of the $|\lambda(t)|$ still increases and the two others magnitudes stay near the value 1. This latter behaviour, which shows 2 eigenvalues staying at finite distance, and

4 escaping eigenvalues instead of 3, is preserved until the value $|t| = 10^{140}$. *This is the starting point for an oscillating behaviour of the diverging eigenvalue illustrated by Figure 6.9.*

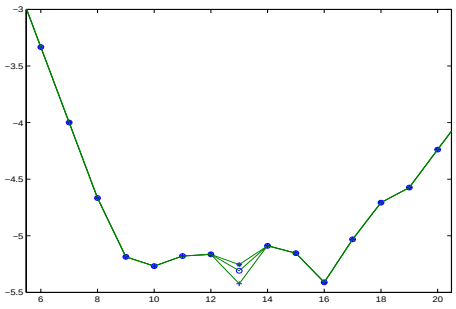


Figure 6.7: $log_{10}|t| \mapsto log_{10}|\lambda(t)|$ where $|t| \in [5 \times 10^{-2},\ 10^{22}]$, $\theta = \pi/3$

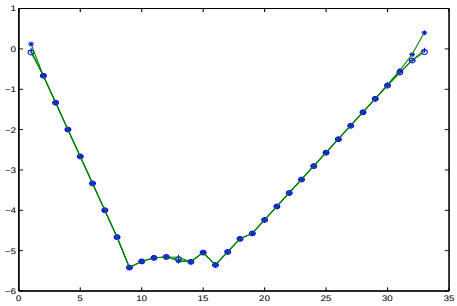


Figure 6.8: $log_{10}|t| \mapsto log_{10}|\lambda(t)|$ where $|t| \in [5 \times 10^{-2},\ 10^{33}]$, $\theta = \pi/3$

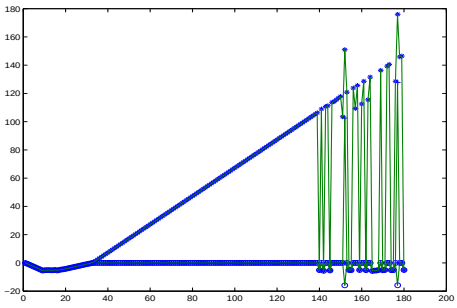


Figure 6.9: $log_{10}|t| \mapsto log_{10}|\lambda(t)|$ where $|t| \in [5 \times 10^{-2},\ 10^{180}]$, $\theta = \pi/3$

△

The Examples 6.3.1 and 6.3.2 show that the determination of $h_M^{(i)}$, for each $i = 1, \ldots, l_*$ is important to guarantee the validity of the computation of the points of Lim in finite precision.

6.3.2 A heuristic method for finding h_M

The value $h_M^{(i)}$ is problem-dependent and we need a way to determine this value. In addition, for each problem, $h_M^{(i)}$ may depend on i. This means that for a problem with $l_* = \text{card Lim}$, one may find up to l_* distinct values $h_M^{(i)}$, $i = 1, \ldots, l_*$.

We introduce a convergence criterion which provides us with values for $h_M^{(i)}$, $i = 1, \ldots, l_*$. To this end, we use the characteristics of the two families of eigenvalues, invariant and evolving, presented in sections 6.3.2.1 and 6.3.2.2 respectively, to introduce a fast and reliable numerical alternative for finding $h_M^{(i)}$, $i = 1, \ldots, l_*$.

6.3.2.1 Determination of invariant eigenvalues

We know that $\lambda \in \sigma^i \subset \sigma(A)$ is an *invariant* eigenvalue under the Homotopic Deviation $A + tE$ iff $\det(A + tE - \lambda I) \equiv 0$ for all $t \in \mathbb{C}$ [18]. An alternative way for finding invariant eigenvalues is to check whether $\det(A + tE - \lambda I)$ is *zero* or not for two or three different values of t. This should be done for all $\lambda_i \in \sigma(A)$, $i = 1, \ldots, n$.

In practice, calculation of $\det(A + tE - \lambda I)$ is too costly and instead one can check the equality $\lambda = \lambda(t)$ for a few different values of $t \neq 0$: usually, checking with 3 or 4 different values of t is enough for distinguishing invariant eigenvalues, if there are any.

Example 6.3.3 *Let*

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}, \quad E = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Then $\pi(z) = \det(zI - A) = z(z-1)(z-2)$, $\det Q(z) = q(z) = \hat{\pi}(z) = z - 1$, *and* $\det(A + tE - zI) = -(z-1)(z^2 + 2z + t)$. *This means that there is just one limit point which is equal to* 1. *Since* $1 \in \sigma(A)$, *the question about* 1 *is the following: is* 1 *an invariant eigenvalue, or does* $\lim_{|t| \to \infty} \lambda_i(t) = 1$ *for* $i = 1$ *or* $i = 2$ *or* $i = 3$?

The eigenvalues of $A + tE$ *listed in the Table 6.2 for 4 different values of* t *show that* $1 \in \sigma(A)$ *is invariant.*

	$\sigma(A+tE)$ for $t=0$	$\sigma(A+tE)$ for $t=10$	$\sigma(A+tE)$ for $t=10^5$	$\sigma(A+tE)$ for $t=10^{15}$
$\lambda_1(t)$	1	1	1	1
$\lambda_2(t)$	2	4.3166	317.2293	3.1623×10^7
$\lambda_3(t)$	0	-2.3166	-315.2293	-3.1623×10^7

Table 6.2: $\sigma(A+tE)$ for 4 values of t

△

6.3.2.2 A convergence criterion for finding h_M

An *evolving* $\lambda \in \sigma^e \subset \sigma(A)$ is an eigenvalue such that $\lambda(t) \neq \lambda$ for almost all $t \in \mathbb{C}$. To discriminate between eigenvalues converging to Lim and eigenvalues escaping to ∞, we should let $|t|$ increase. The convergence criterion is chosen as follows

$$r^e_{i,k} = \frac{|\lambda_i(t_k) - \lambda_i(t_{k-1})|}{|\lambda_i(t_k)|} \leq 10^{-n_{max}}, \tag{6.3.1}$$

for a given integer $0 \leq n_{max} \leq 15$. It uses the *relative error* associated with $\lambda_i(t)$ for the two successive values t_{k-1} and t_k of t. When the process is stopped for each $\lambda_i(t)$, we set $h^{(i)}_M = |t_k|$ and $\lambda_i(t_k)$ is considered as an approximation of $\lim_{|t|\to\infty} \lambda_i(t)$ with $n_{\max}$ significant digits. The first two successive values t_{k-1} and t_k which obtain the inequality (6.3.1) may be different for each eigenvalue $\lambda_i(t) \in \sigma(A(t))$.

When there is a risk of a small $|\lambda_i(t_k)|$, one may use the alternative *absolute criterion*

$$a^e_{i,k} = |\lambda_i(t_k) - \lambda_i(t_{k-1})| \leq 10^{-n_{max}}. \tag{6.3.2}$$

Example 6.3.4 *Let*

$$A = \begin{bmatrix} 0 & -1 & 1 & 1 & -1 \\ 0 & 2 & 0 & -1 & 0 \\ 2 & -1 & 0 & 2 & 0 \\ 0 & 2 & 0 & -1 & 1 \\ 2 & -1 & -1 & 2 & 2 \end{bmatrix}, \quad E = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

Then we have the following results.

a) $\det(zI - A) = (z-1)(z-2)(z+1)(z^2 - z - 1)$, *and* $\det Q(z) \equiv 0$, *so that* $F(A,E) = re(A)$.

b) Using the notations of section 3.9, $q = 2$, $n_1 = 3$, $r_1 = 1$, $n_q = 2$ *and* $r_q = 1$. *Also,* $\Gamma_1 = 2 \neq 0$ *and* $\Gamma_2 = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$ *which is singular. This yields*

$\Omega = 2 - 2(1/2)2 = 0$, *and,* $\sigma(\Omega) = \{0\}$. *We remark that here* $g' = 0$, *hence this example does not satisfy the condition* (G). *Therefore* $\mathrm{Lim} \supset \{0\}$ *does not hold necessarily. We have to resort to computation to determine* Lim.

c) *The computation of* $t \mapsto \sigma(A(t))$ *in Figure 6.10 shows that three eigenvalues* $\lambda_i(t)$, $i = 1, 2, 3$, *go to* ∞ *as* $|t| \to \infty$ *but two eigenvalues reach finite limits (i=4,5). The blue* $\odot$ *are the limit points, and the red* $\circledast$ *are the spectrum of* A.

To compute Lim, *we use the stopping criterion (6.3.1) when* $|t|$ *increases. Here,* $|t_{k+1}| = 10^2\ |t_k|$, *and* $\theta = \pi/3$ *is fixed.*

We can see on Figures 6.11 and 6.12 that $r^e_{4,18} \leq 10^{-15}$ *(red circles) and* $r^e_{5,9} \leq 10^{-15}$ *(green +). The points in* Lim *are found as* $\mathrm{Lim} = \{2.4142,\ -0.4142\}$ *(rounded to 5 significant digits).*

See Figure 6.11 for $n_{max} = 15$. *For the localization of points in* Lim, *this is the best that can be achieved in finite precision. This means that the condition* $r^e_{i,k} \leq 10^{-u}$ *has meaning for* $u \lesssim 15$ *and* $i = 4, 5$. *See Figure 6.12 for* $n_{max} = 16$.

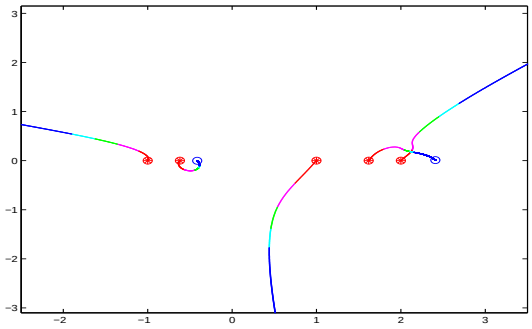


Figure 6.10: The plot of $\sigma(A(t))$ for $0 < |t| < 10^2$, $\theta = \pi/3$

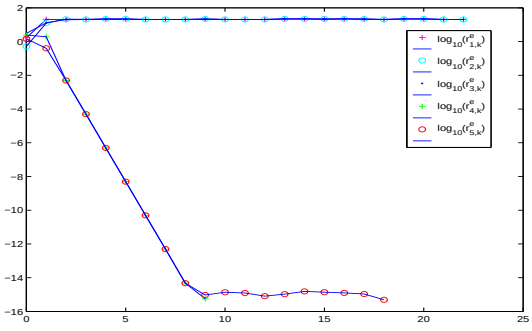


Figure 6.11: $log_{10}(|t_i|) \mapsto log_{10}(r^e_{i,k})$ where $n_{max} = 15$, $\theta = \pi/3$

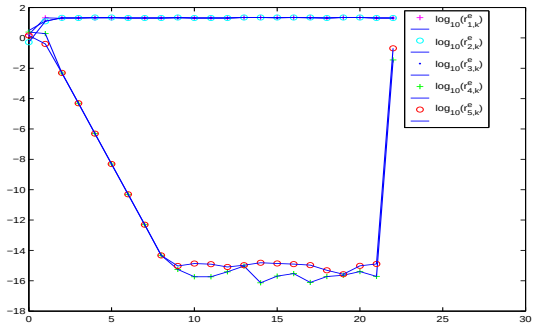


Figure 6.12: $log_{10}(|t_i|) \mapsto log_{10}(r^e_{i,k})$ where $n_{max} = 16$, $\theta = \pi/3$

Figure 6.11 indicates that for getting an accuracy of 15 *significant digits, we need* $h_M^{(5)} = 10^9$ *and* $h_M^{(4)} = 10^{18}$.

Moreover, one can see on Figure 6.12 that, if we insist on increasing $|t|$ *above* 10^{21}, *then we suddenly loose the accuracy (of* 15 *significant digits) previously achieved.*

△

6.3.3 Large values of the magnitude of some points in Lim

Depending on the entries of matrices A and E, it may happen that the magnitudes of some points in Lim are very large. In such cases, one has to use even larger values of $|t|$ to find such limit points with some accuracy. Example 6.3.5 illustrates one such case.

Example 6.3.5 *Let* a *be a given real parameter*

$$A = \left[\begin{array}{ccc|cc} 0 & -1 & 0 & 1 & 2 \\ 1 & 2 & 1 & 0 & 1 \\ 0 & -1 & 0 & 0 & 1 \\ \hline 0 & 0 & 1 & 1 & a \\ 0 & 0 & 0 & 2 & 2 \end{array}\right], \text{ and } E = \left[\begin{array}{ccc|cc} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{array}\right] = e_5 e_4^T \text{ of rank } 1.$$

The order of A *and* E *is* 5, E *is defective nilpotent with the eigenvalues* $(0^1)^3(0^2)$, $r = \operatorname{rank} E = 1$, *with* $q = 2$, $n_q = 1$, $r_q = 3$. *In* $A + tE$, *the* 2×2 *bottom right matrix* $\left[\begin{array}{cc} 1 & a \\ 2+t & 2 \end{array}\right]$ *is related to top left* 3×3 *block matrix by the scalar* 1: *the computation of* $t \mapsto \sigma(A(t))$ *for different values of* a, *suggests that for* $a \neq 0$, 2 *eigenvalues escape to* ∞ *and* Lim *consists of* 3 *points. The fact that for* $a \neq 0$, *exactly* 2 *eigenvalues escape to* ∞, *will be shown in Chapter* 8, *Theorem 8.6.2.*

We use $n_{max} = 7$ *for all the cases which are considered below. We consider* $t = he^{i\theta}$ *with a fixed* $\theta = \pi/6$.

1) *When* $a = 1$ *, it is shown on Figure 6.13 that for approximating the three points in* Lim *to* 7 *digits, we need to take* $h_M^{(3)} = 10^9$ *,* $h_M^{(4)} = 10^9$ *and* $h_M^{(5)} = 10^8$ *.*

2) *When* $a = 10^{-2}$ *, Figure 6.14 shows the necessity of taking* $h_M^{(3)} = 10^{14}$ *,* $h_M^{(4)} = 10^{11}$ *and* $h_M^{(5)} = 10^9$ *respectively to get to the same accuracy of* 7 *significant digits.*

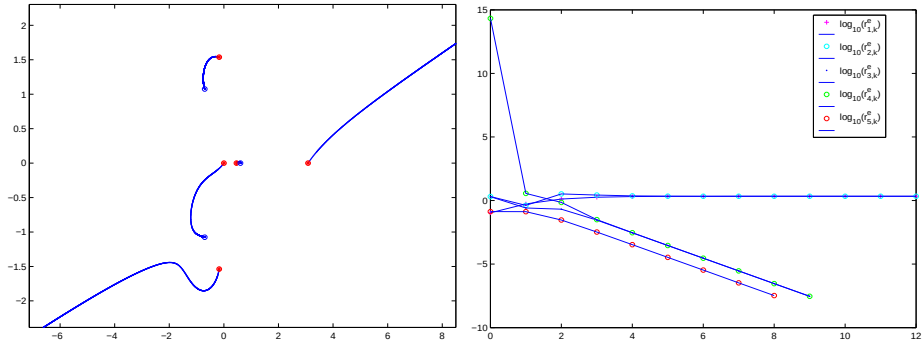
This phenomenon of smaller a *- larger* $h_M^{(3)}$ *continues until* $a = 10^{-102}$ *. For the value* $a = 10^{-102}$ *, the limit* $\lim_{|t|\to\infty} \lambda_3(t)$ *cannot be computed even with* $h = 10^{306}$ *(the maximum usable number in Matlab). The values of* a *with the corresponding* $h_M^{(3)}$ *are listed in the Table 6.3.*

The two other limit points, l_4 *and* l_5 *, have a bounded magnitude for every value of* a *.*

a	$h_M^{(3)}$	computed $l_3 = \lambda_3(t)$ with $\lvert t\rvert = h_M^{(3)}$
1	10^9	$-0.7028 + 1.0758i$
10^{-2}	10^{14}	$-1.0096 \times 10^2 + 5.5209 \times 10^{-7}i$
10^{-8}	10^{32}	$-1.000 \times 10^8 + 5.000 \times 10^{-1}i$
10^{-16}	10^{56}	$-1.0000 \times 10^{16} + 5.0000 \times 10^7 i$
10^{-75}	10^{233}	$-1.0000 \times 10^{75} + 5.0000 \times 10^{67}i$
10^{-95}	10^{293}	$-1.0000 \times 10^{95} + 5.0000 \times 10^{87}i$
10^{-102}	$\leq 10^{306}$	impossible

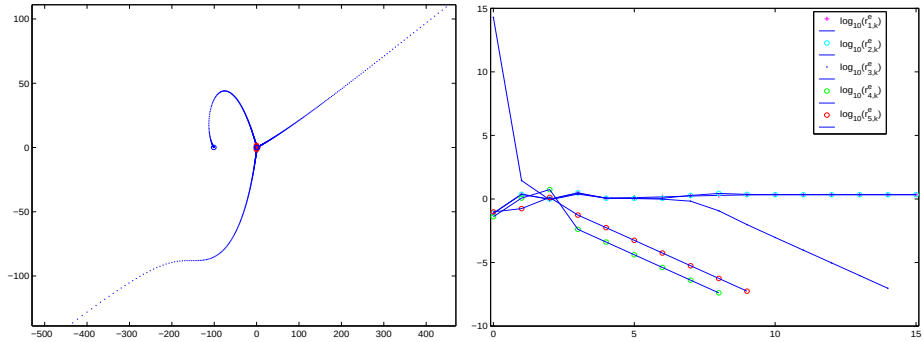
Table 6.3: Correspondence $a \mapsto h_M^{(3)}$, $\theta = \pi/6$

In each Figure 6.13, 6.14, . . . , 6.19, we respectively plot both $t \mapsto \sigma(A(t))$ *and* $log_{10}(|t|) \mapsto log_{10}(r_{i,k}^e)$ *for* $i = 1, \ldots, 5$ *corresponding to the cases mentioned in the Table 6.3. The* **scale factor** *is indicated when larger than* 1 *. The left plots,* (a) *, of the Figures 6.14, 6.15, . . . , 6.19 show that the scale factor increases as* a *decreases. More precisely, we observe that the automatic scale factor given by Matlab is of the order of* $1/a$ *.*



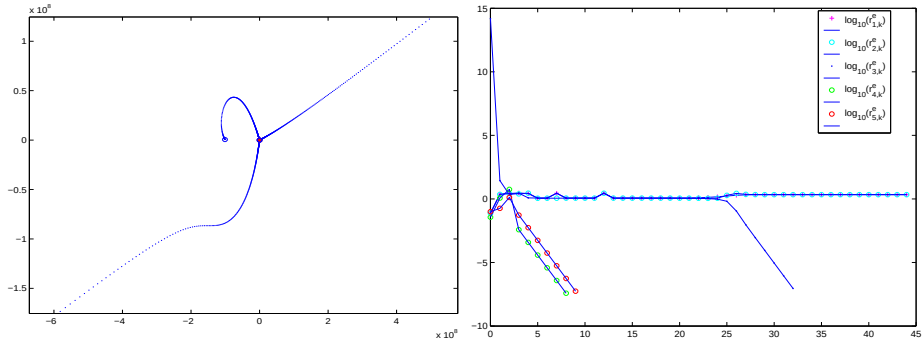
(a) $t \mapsto \sigma(A(t))$ (b) $log_{10}(|t|) \mapsto log_{10}(r^e_{i,k})$

Figure 6.13: $a = 1$



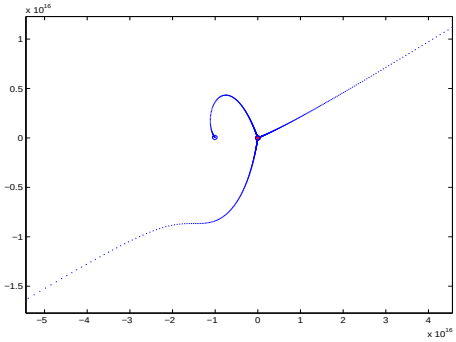
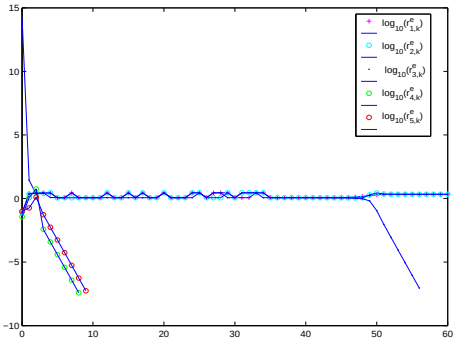
(a) $t \mapsto \sigma(A(t))$ (b) $log_{10}(|t|) \mapsto log_{10}(r^e_{i,k})$

Figure 6.14: $a = 10^{-2}$



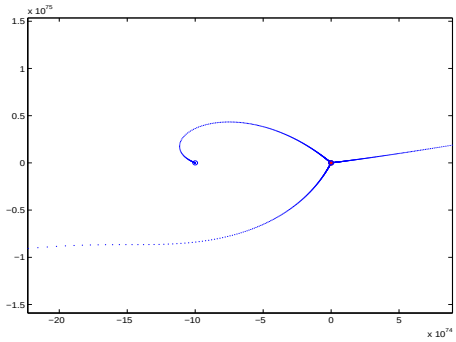
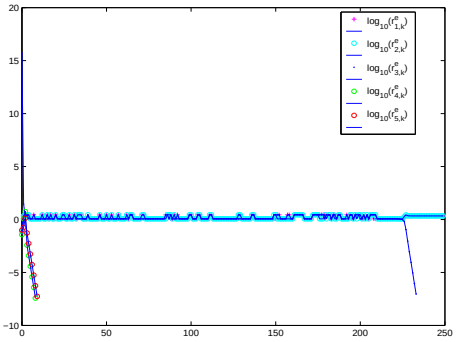
(a) $t \mapsto \sigma(A(t))$ with scale factor 10^8 (b) $log_{10}(|t|) \mapsto log_{10}(r^e_{i,k})$

Figure 6.15: $a = 10^{-8}$

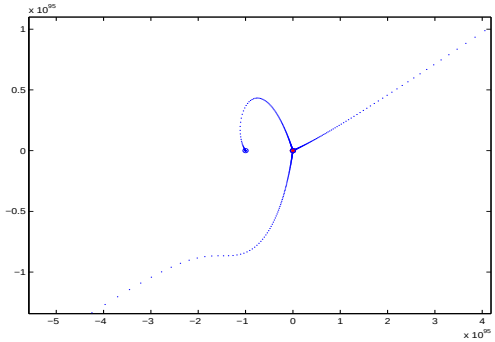
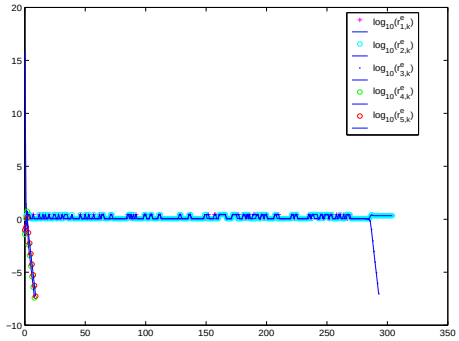
(a) $t \mapsto \sigma(A(t))$ with scale factor 10^{16} (b) $log_{10}(|t|) \mapsto log_{10}(r^e_{i,k})$

Figure 6.16: $a = 10^{-16}$

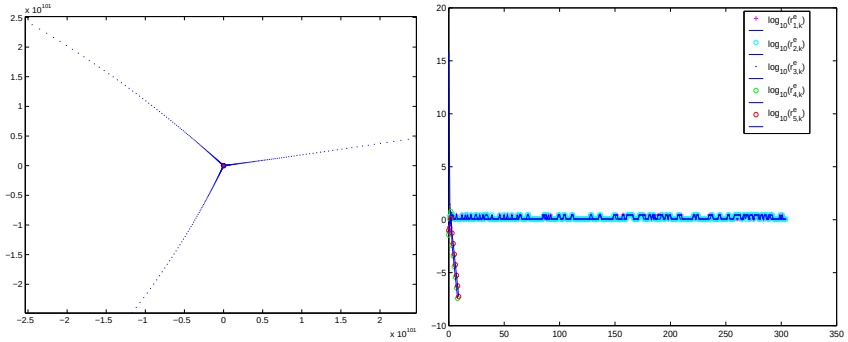
(a) $t \mapsto \sigma(A(t))$ with scale factor 10^{75} (b) $log_{10}(|t|) \mapsto log_{10}(r^e_{i,k})$

Figure 6.17: $a = 10^{-75}$

(a) $t \mapsto \sigma(A(t))$ with scale factor 10^{95} (b) $log_{10}(|t|) \mapsto log_{10}(r^e_{i,k})$

Figure 6.18: $a = 10^{-95}$



(a) $t \mapsto \sigma(A(t))$ with scale factor 10^{101} (b) $log_{10}(|t|) \mapsto log_{10}(r^e_{i,k})$

Figure 6.19: $a = 10^{-102}$

△

6.3.4 Dependence of the convergence on the parameter θ

In this Chapter, we have principally analysed the role of $h = |t|$ in finding the limit points. Here we shall make some remarks about the role of the parameter θ in the determination of Lim.

As we have seen in section 6.2, the value of the parameter θ in $[0,\ 2\pi[$ does not affect the number and the values of the points in $\text{Lim} = \{l_1, \ldots, l_{l_*}\}$. The role of the parameter θ is to govern the one-to-one correspondence

$$\{\lambda_1, \ldots, \lambda_{l_*}\} \subset \sigma(A) \mapsto \{l_1, \ldots, l_{l_*}\}. \tag{6.3.3}$$

To illustrate this fact, we ask the reader to look at the Figures 6.5, (a), (b) and (c) in Example 6.2.4: they show how the correspondence (6.3.3) is affected by the choice of different values for θ. We will present more Examples of this kind in Chapter 7.

6.3.5 A methodological remark

In the qualitative analysis of HD that we have presented, we have let $h = |t|$ take extremely large or small values, much in the spirit of *pure mathematics*, where a real variable h can, in theory, tend to 0 or ∞ without any difficulty.

However, this is not usually the case in finite precision computation, because of the intrinsically finite character of the computer arithmetic. Beginners in the art of computer simulations quickly learn the limiting role played by machine precision in the assessment of computer results [23].

In view of this well-established body of experience, it usually does not make sense to let a parameter be too small or too large in magnitude when compared with

machine precision. However, HD computer simulations reproduce extraordinarily well the mathematical predictions for $|t|$ extremely large (up to the largest usable number 10^{306}). The phenomenon is completely *unexpected*; this is the reason why we have chosen to present at length the "unreasonable" robustness of HD to finite precision.

Chapter 7

Homotopic backward analysis, II Numerical illustrations

This Chapter revisits the CERFACS technical report [5] to take into account the important new result that $\det Q(z) = (\pi(z))^{r-1}\hat{\pi}(z)$ for $z \in \mathbb{C}$. We mainly follow [22]. In section 7.1, we study the matrix M_z and its spectrum $\sigma(M_z)$ when $z \to \lambda \in \sigma(A)$ in exact arithmetic and in finite precision. Then in section 7.2, the three portraits associated with the map $z \in re(A) \mapsto M_z$ as visualisation tools in 2D and 3D are presented. Using the 2 homotopic portraits, $\phi_1 : z \mapsto \rho(M_z)$ and $\phi_2 : z \mapsto \rho(M_z^{-1})$, we propose a qualitative representation of the map $z \mapsto M_z$. Finally in section 7.3, we treat in detail the case $r = 2$ and we present numerical examples which illustrate the use of the visualization tools described in Chapters 6 and 7.

7.1 Study of M_z and $\sigma(M_z)$ as the parameter z in $re(A)$ tends to $\lambda \in \sigma(A)$

M_z and $\sigma(M_z)$ are well-defined for $z \in re(A)$. For $z \in re(A)$, we have

$$\det M_z = \vartheta(z) = \frac{\hat{\pi}(z)}{\pi(z)} = \Pi_{i=1}^{r}\mu_{iz},$$

where $\pi(z)$ and $\hat{\pi}(z)$ are polynomials in $z \in \mathbb{C}$ of respective degree n and $\leq n - r$ and where $\mu_{iz} \in \sigma(M_z)$. In what follows, we look at the situation when $z \in re(A) \to \lambda \in \sigma(A)$.

Remark: Certain limits $\lim_{z\to\lambda}\mu_{iz} = \mu_{i\lambda}$ may exist when $\lim_{z\to\lambda} M_z$ does *not* exist. Below the notation $\lim_{z\to\lambda}\max|\mu_{iz}| = \rho(M_\lambda) < \infty$ is just a notation. It does not imply that M_λ exists.

When $\lambda \in \sigma(A)$, $\pi(\lambda) = 0$. We assume that $\hat{\pi}(z) \not\equiv 0$. Let $m_\lambda > 0$ (resp.

λ	$\hat{Z} \cap \sigma(A)$	$\sigma(A) \backslash \hat{Z}$
τ_λ	$\mathbb{Q}^+$	∞
$\hat{m}_\lambda$	≥ 1	0

Table 7.1: The possible values for τ_λ and $\hat{m}_\lambda$ when $\hat{\pi}(z) \not\equiv 0$

$\hat{m}_\lambda \geq 0$) be the algebraic multiplicity of λ as a root of $\pi(z)$ (resp. $\hat{\pi}(z)$): $\hat{m}_\lambda = 0$ (resp. ≥ 1) if $\lambda \notin \hat{Z}$ (resp. $\lambda \in \hat{Z}$).

Let us define $\tau_\lambda = \frac{m_\lambda}{\hat{m}_\lambda}$. Then

$$\hat{m}_\lambda < m_\lambda \Longleftrightarrow \tau_\lambda > 1,$$

$$\hat{m}_\lambda \geq m_\lambda \Longleftrightarrow 0 < \tau_\lambda \leq 1.$$

The Table 7.1 displays the possible values of τ_λ and $\hat{m}_\lambda$ under the 2 conditions $\lambda \in \hat{Z} \cap \sigma(A)$ or $\lambda \in \sigma(A) \backslash \hat{Z}$.

7.1.1 Partition of $\sigma(A)$

The eigenvalues in $\sigma(A)$ are counted with their algebraic multiplicity. The behaviour of $\lambda(t)$ as $|t| \to \infty$ induces the following distinction between eigenvalues λ in $\sigma(A)$: $\sigma(A) = \sigma^e \cup \sigma^i$, with $\sigma^e \cap \sigma^i = \emptyset$, where

- $\lambda \in \sigma^e \Leftrightarrow \lambda$ is an *evolving* eigenvalue, that is $\lambda(t) \neq \lambda$ for almost all $t \in \mathbb{C}$.
- $\lambda \in \sigma^i \Leftrightarrow \lambda$ is an *invariant* eigenvalue, that is $\lambda(t) = \lambda$ for all $t \in \mathbb{C}$.

Looking at the limit of $\lambda(t)$ as $|t| \to \infty$ leads to the definition of $\sigma^f \subset \sigma(A)$:

$\lambda \in \sigma^f \Leftrightarrow \lambda$ is a *final* eigenvalue, that is $\lambda = \lim_{|t| \to \infty} \lambda(t)$, where $\lambda(t)$ originates in an evolving eigenvalue $\lambda(0) = \lambda' \in \sigma^e$. It is rare but possible that $\lambda' = \lambda$: this creates a loop, see Example 7.3.2 in Section 7.3.

Therefore Lim $= \sigma^i \cup \sigma^f \cup \Lambda(A, E)$, for $\sigma^i \cup \sigma^f \subset \sigma(A)$, and $\sigma^i \cap \sigma^f = \emptyset$.

7.1.2 Observability of $\lambda \in \sigma(A)$ by HD for $r \geq 2$

When $z \to \lambda$ then it is possible that for some $i \in \{1, \ldots, r\}$, $|\mu_{iz}| \to \infty$. This means that, when $z \to \lambda$ then the possibilities are one of the following:

$$\forall i = 1, \ldots, r, \quad |\mu_{iz}| \to \infty, \quad \text{or} \tag{7.1.1}$$

$$\exists i, j \in \{1, \ldots, r\} \text{ such that } |\mu_{iz}| \to \infty, \quad \text{and}, \quad \mu_{jz} \to \mu_{j\lambda} \in \mathbb{C} \quad \text{or} \tag{7.1.2}$$

$$\forall i = 1, \ldots, r, \quad \mu_{iz} \to \mu_{i\lambda} \in \mathbb{C}. \tag{7.1.3}$$

Each one of (7.1.1), (7.1.2), and (7.1.3) results in a specific kind of observability for $\lambda \in \sigma(A)$ (by HD for $r \geq 2$) as follows.

7.1.2.1 Spectral observability

Under the condition (7.1.1), $\forall i = 1, \ldots, r$, $\lim_{z\to\lambda} |\mu_{iz}| = \infty$ and therefore $\rho(M_\lambda) = \infty$. In this case, we say that λ is spectrally observable or σ-observable. In addition,

$$\min_{1\le i\le r} |\mu_{i\lambda}| = \infty \Longleftrightarrow \rho(M_\lambda^{-1}) = 0.$$

7.1.2.2 Partial observability

As it is written in (7.1.2), it is possible that only *some* eigenvalues μ_z, $|\mu_z| < \rho(M_z)$, have a limit as $z \to \lambda$, rather than the whole spectrum. In this case we say that $\lambda \in \sigma(A)$ is partially observable. A necessary condition is that $\mu_z = \frac{d(z)}{\pi(z)}$ with $d(\lambda) = 0$.

Let λ have algebraic multiplicity m_λ so that $\pi(z) = (z-\lambda)^{m_\lambda}\pi^{'}(z)$ with $\pi^{'}(\lambda) \neq 0$. The condition $d(z) = (z-\lambda)^{m_\lambda} d^{'}(z)$, where $d^{'}(z)$ is continuous around $z = \lambda$, implies that $\mu_z = \frac{d^{'}(z)}{\pi^{'}(z)}$ and $\mu_z \to \mu_\lambda = \frac{d^{'}(\lambda)}{\pi^{'}(\lambda)}$ as $z \to \lambda$. If $d^{'}(\lambda) = 0$ then $\mu_\lambda = 0$. An example is provided in section 7.3, Example 7.3.2.

For a partially observable $\lambda \in \sigma(A)$, we have $\min_{1\le i\le r} |\mu_{i\lambda}| < \infty$ which may be equal to 0. Also $\rho(M_\lambda) = \infty$.

7.1.2.3 Spectral nonobservability

Under the condition (7.1.3), $\forall i = 1, \ldots, r$, $\lim_{z\to\lambda} |\mu_{iz}| = |\mu_{i\lambda}| < \infty$. In this case, λ is said to be spectrally nonobservable, or σ-nonobservable. This is a case where $\min_{1\le i\le r} |\mu_{i\lambda}| < \infty$ with the possibility $\min_{1\le i\le r} |\mu_{i\lambda}| = 0$. This happens when the r eigenvalues $d(z)$ of $Q(z)$ contain at least the common factor $(z-\lambda)^{m_\lambda}$ which again cancels with that in $\pi(z)$.

7.1.2.4 Normwise observability of λ: $\lim_{z\to\lambda} M_z$ does not exist

This happens generically because the r^2 elements of the matrix $Q(z)$ do not have $z = \lambda$ as a common zero with multiplicity m_λ at least (see section 7.2). The presence of an eigenvalue λ for A is revealed by the fact that $\lim_{z\to\lambda} M_z$ does not exist : the map $z \mapsto \|M_z\|$ has a peak at any $\lambda \in \sigma(A)$ which is $\|\cdot\|-$ observable.

Since we have $\rho(M_z) \le \|M_z\|$, either $\sigma-$observability or partial observability results in normwise observability.

7.1.2.5 Normwise nonobservability: $\lim_{z\to\lambda} M_z = M_\lambda$

The r^2 element in $Q(z)$ contain at least a factor $(z-\lambda)^{m_\lambda}$, which cancels with the factor $(z-\lambda)^{m_\lambda}$ in $\pi(z)$. A characterisation is given in [18, 30] under the form $V^H D_i U = 0$, $D_i = (A-\lambda I)^i P_\lambda$, $i = 0$ to $l_\lambda - 1$ where P_λ is the spectral projection and l_λ is the ascent of λ [34].

It is clear that normwise nonobservability implies spectral nonobservability. However, the reciprocal does not hold for $r > 1$. Normwise observability does not forbid spectral non observability, since $\rho(M_z) \leq \|M_z\|$. This is illustrated by the

Example 7.1.1 *Let us consider the Example 3.9.2 where A is the companion matrix associated with $\pi(z) = z^{11} + 1$, in upper Hessenberg form, and first column e_2.*
$E = UV^T$ with $U = [e, e_2]$ and $V = [e_{11}, e_3]$ of rank $r = 2$, $e = [1, \ldots, 1]^T$.
Direct calculation leads to $M_z = \frac{1}{\pi(z)} Q(z)$ with

$$Q(z) = \begin{pmatrix} 1 + z + \cdots + z^{10} & z \\ (-1 - z - \cdots - z^7) + z^8 + z^9 + z^{10} & z^9 \end{pmatrix}$$

which is a matrix polynomial of order 2 and degree $\leq 10 = n - 1$, with

$$d^{\pm}(z) = \tfrac{1}{2} \left[\mathrm{tr}Q(z) \pm (\mathrm{tr}^2 Q(z) - 4\det Q(z))^{1/2} \right],$$

$$\mathrm{tr}Q(z) = (z+1)(1 + z^2 + z^4 + z^6 + z^8 + z^9).$$

$\det Q(z) = \pi(z)\hat{\pi}(z)$ *has degree $19 < 20$. This yields the factorizations*

$$\det Q(z) = (z+1)^2 z(1 - z + z^2 - \cdots - z^9 + z^{10})(z^2 + 1)(z^4 + 1),$$

$$\pi(z) = (z+1)(1 - z + z^2 - \cdots - z^9 + z^{10}), \text{ and}$$

$$\hat{\pi}(z) = (z+1)z(z^2+1)(z^4+1) = z(1 + \cdots + z^7) = \tfrac{\det Q(z)}{\pi(z)}.$$

We have also $\vartheta(z) = \frac{\hat{\pi}(z)}{\pi(z)} = \frac{z(z^2+1)(z^4+1)}{1 - z + z^2 - \cdots - z^9 + z^{10}}$, after simplification by $(z+1)$.

$\lambda = -1$ is a root of $\pi(z)$ with multiplicity $m_\lambda = 1$. It is also a root of $\hat{\pi}(z)$ with multiplicity $\hat{m}_\lambda = 1$. Clearly $\vartheta(-1)$ exists and

$$\vartheta(-1) = \frac{d_1(-1)}{\pi(-1)} \frac{d_2(-1)}{\pi(-1)} = \det M_{-1} \neq 0.$$

Here $z+1$ is a factor for both eigenvalues $d^{\pm}(z)$ and for $\pi(z)$, and $\lim_{z \to -1} \sigma(M_z)$ exists, equal to $\{2 \pm 4\sqrt{3}\} = \sigma_{-1}$: the eigenvalue $\lambda = -1$ for A is spectrally unobservable. However, $\lim_{z \to -1} M_z$ does not exist: $z + 1$ is not a common factor for the 4 elements of $Q(z)$: $\lambda = -1$ is normwise observable.

$\mu_z^+ = d^+(z)/\pi(z)$ (resp. $\mu_z^- = d^-(z)/\pi(z)$) does not exist (resp. exists) at the 10 eigenvalues $\sigma(A) \backslash \{-1\}$. Therefore all of the 10 eigenvalues, $\sigma(A) \backslash \{-1\}$, are partially observable. This also show that the 10 eigenvalues $\sigma(A) \backslash \{-1\}$ are normwise observable.

$\triangle$

In summary, the following conceptual inclusions hold for $2 \leq r < n$:

$$\left.\begin{array}{r} \sigma - \text{nonobs.} \subset \text{ partial obs.} \subset \sigma - \text{obs.} \subset \\ \|\cdot\| - \text{nonobs.} \subset \end{array}\right\} \ \|\cdot\| - \text{obs.}$$

where obs. stands for observability.

For $r = 1$, the two notions coalesce: $\sigma-$ obs. $\Leftrightarrow$ $\|\cdot\|-$ obs.

For $r = n$, the matrices M_z and $(zI-A)^{-1}$ are unitarily equivalent: they share the same singular values. They are unitarily similar if $V = U$, that is $E = UU^H = I$. Hence $A(t) = A + tI$, $\lambda(t) = \lambda + t$, and $\mu_{iz} = \frac{1}{z-\lambda_i}$ for $z \neq \lambda_i \in \sigma(A)$, $i = 1, \ldots, n$.

7.1.2.6 Dim observability in finite precision

The above discussion assumes exact arithmetic.
Normwise observable eigenvalues (M_λ does not exist) may nevertheless be difficult to detect in finite precision when the matrices $V^H D_i U$ have a small norm: no peak appears at λ for the global spectral portrait, unless the mesh size is sufficiently refined around λ. We say that λ is *dimly* normwise observable (or $\|.\|-$ observable for short).

Example 7.1.2 *Let A be the companion matrix associated with $\pi(z) = (z-1)^3(z-3)^4(z-7)$, in upper Hessenberg form and first column vector e_2. $E = UV^T$ with $U = [e_1, e]$ and $V = [e_6, e_8]$ of rank 2 for $e = [1, \ldots, 1]^T$. $V^T U = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$ has rank 1. Direct calculation leads to $M_z = \frac{1}{\pi(z)} Q(z)$ with*

$$Q(z) = \begin{pmatrix} z^2 & (z-7)(z-1)^2(z-3)^4 \\ 1 & (z-7)(z-1)^2(z-3)^4 \end{pmatrix}$$

which is a matrix polynomial of order 2 and degree $\leq 7 = n-1$, with

$$d(z) = \tfrac{1}{2}\left[\mathrm{tr}Q(z) \pm \left(\mathrm{tr}^2 Q(z) - 4\mathrm{det}Q(z)\right)^{1/2}\right],$$

$$\mathrm{tr}Q(z) = -567 + 1971z - 2726z^2 + 1947z^3 - 781z^4 + 177z^5 - 21z^6 + z^7 .$$

$\det Q(z) = \pi(z)\hat{\pi}(z)$ *has degree 9. This yields the factorization*

$$\mathrm{det}Q(z) = (z-1)^3(z-3)^4(z-7)(z+1) \quad \textit{for}$$

$$\pi(z) = (z-1)^3(z-3)^4(z-7), \ \textit{and}$$

$$\hat{\pi}(z) = (z+1) = \tfrac{\mathrm{det}Q(z)}{\pi(z)} .$$

We have also $\vartheta(z) = \frac{\hat{\pi}(z)}{\pi(z)} = \frac{(z+1)}{(z-7)(z-1)^3(z-3)^4}$.

The eigenvalue $\lambda = 7$ *with eigenprojection* P_λ *is difficult to capture by the map* $z \mapsto log_{10}(\rho(M_z))$ *because*

a) $\lambda = 7$ *competes with the two defective eigenvalues of ascent* 3 *and* 4 , *and*

b) $\|V^T P_\lambda U\| = O(10^{-4})$. *See on Figure 7.1 the global spectral portrait (mesh size* $\sim 10^{-2}$ *), and on Figure 7.2 a zoom around* 7 *(mesh size* $\sim 5\ 10^{-4}$ *).*

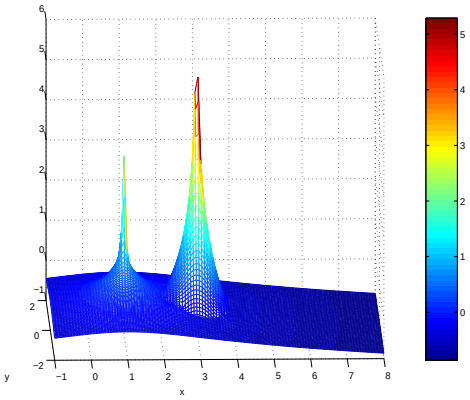


Figure 7.1: The global spectral portrait $z \mapsto log_{10}(\rho(M_z))$

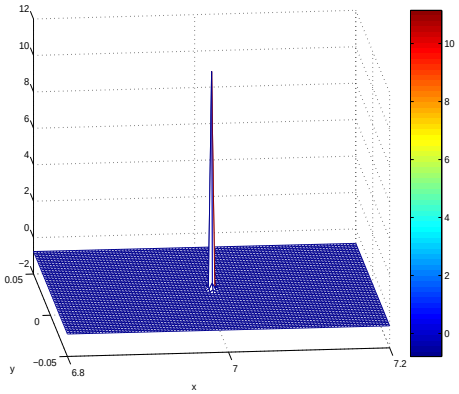


Figure 7.2: Zoom around $\lambda = 7$

△

It is well known [34] that, if M_z is highly nonnormal, the difference between $\rho(M_z)$ and $\|M_z\|$ may be blurred by finite precision. Another more subtle computation difficulty concerns the peaks and wells of the frontier portrait, and is addressed in the next section.

7.1.3 The product $\vartheta(z) = \Pi_{i=1}^r \mu_{iz}$ as $z \to \lambda$

We recall that $\det M_z = \vartheta(z) = \frac{\hat{\pi}(z)}{\pi(z)}$ for $z \in re(A)$, where $\pi(z)$ and $\hat{\pi}(z)$ are polynomials in $z \in \mathbb{C}$ of respective degree n and $\leq n - r$. We address this question: what is the relationship between observability and the product $\vartheta(z) = \Pi_{i=1}^r \mu_{iz} = \frac{\hat{\pi}(z)}{\pi(z)}$.

Proposition 7.1.1 *[22] Suppose that $\lambda \in \sigma(A)$.*
i) When λ is σ-observable, then the product $\vartheta(z) = \frac{\hat{\pi}(z)}{\pi(z)}$ does not exist at λ. Hence $\hat{m}_\lambda < m_\lambda$ ($\tau_\lambda > 1$).
ii) When λ is partially observable, then the product $\vartheta(\lambda) = \Pi_{i=1}^r \mu_{i\lambda}$ exists iff $\hat{m}_\lambda \geq m_\lambda \geq 1$ ($\tau_\lambda \leq 1$).
iii) When λ is σ-nonobservable, then the product $\vartheta(\lambda) = \Pi_{i=1}^r \mu_{i\lambda}$ exists. Hence $\hat{m}_\lambda \geq m_\lambda \geq 1$, and $\vartheta(\lambda) \neq 0$ (resp. $= 0$) iff $\hat{m}_\lambda = m_\lambda$ (resp. $\hat{m}_\lambda > m_\lambda \geq 1$).

Proposition 7.1.1 shows that when $\lambda \in \sigma(A)$ is simple and $\sigma-$observable, then $\lambda \notin \hat{Z}$. Another result is that under *ii)* and *iii)*, $\lambda \in \hat{Z}$ if $\vartheta(\lambda) = 0$.

Example 7.1.3 *Let us go back to Example 7.1.1. In this example $-1 \in \sigma(A)$ is $\sigma-$nonobservable and $m_{-1} = 1 = \hat{m}_{-1}$, therefore $0 \neq \vartheta(-1) < \infty$:* $0 \notin \sigma_{-1} = \{2 \pm 4\sqrt{3}\}$.

△

7.1.4 Possible extension of $F(A, E)$ or $C(A, E)$ into $\sigma(A)$

7.1.4.1 Closure of $F(A, E)$ into $\mathbb{C}$

Let us define $\bar{\sigma} = \{\lambda \in \sigma(A)\colon \min_{1 \leq i \leq r} |\mu_{i\lambda}| = 0\}$. When $\bar{\sigma} \neq \emptyset$, then $\lambda \in \bar{\sigma}$ is at most partially observable: $\lambda \in \bar{\sigma}$ is either $\sigma-$nonobservable or partially observable. In this case, λ can be added to $F(A, E)$ by continuity and we get what is called the *closure* of $F(A, E)$ into $\mathbb{C}$ (when $\hat{\pi}(z) \not\equiv 0$). It is denoted by $\bar{F}(A, E) = F(A, E) \cup \bar{\sigma} \subset \mathbb{C}$.

Observe that $\bar{\sigma} \not\subset \hat{Z}$ is possible when λ is partially observable. Also $\sigma^f \subset \bar{\sigma}$, a fact that is not true for σ^i: a necessary condition for $\lambda \in \sigma^f$ is $\min_{1 \leq i \leq r} |\mu_\lambda| = 0$.

See the Example 7.3.2 for a case where $F(A, E)$ can be extended.

7.1.4.2 Closure of $C(A, E)$ into $\mathbb{C}$

Let us define $\sigma_c = \{\lambda \in \sigma(A)\colon |\mu_{i\lambda}| = 0,\ \forall i = 1, \ldots, r\}$. When $\sigma_c \neq \emptyset$, then $\lambda \in \sigma_c$ is necessarily $\sigma-$nonobservable and $\lambda \in \hat{Z}$. In this case, λ can be added to $C(A, E)$ by continuity. This yields the closure of $C(A, E)$ into $\mathbb{C}$ which is denoted by $\bar{C}(A, E) = C(A, E) \cup \sigma_c \subset \mathbb{C}$.

When $r = 1$, we have $\bar{F}(A, E) = \bar{C}(A, E)$, and $\sigma_c = \bar{\sigma}$.

Example 7.1.4 *Let us look at Example 7.1.1 again.*
$\lambda = -1$ *is* $\sigma -$ *unobservable , with* $\mu_{1\lambda} = 2 - 4\sqrt{3}$ *and* $\mu_{2\lambda} = 2 + 4\sqrt{3}$ *. Therefore, neither* $F(A, E)$ *nor* $C(A, E)$ *can be extended.*

△

7.2 The three homotopic portraits associated with $z \mapsto M_z$

It is clear that the map $z \mapsto M_z,\ z \in re(A)$ contains an essential information ruling the properties of the resolvent map $z \mapsto R(t, z)$. The complete information is contained in the r^2 elements of the matrix map $z \mapsto M_z$.
A more compact form for this information corresponds to the r numbers of the vector map $z \mapsto \sigma(M_z)$.
For $r > 1$, such maps are not easy to visualise graphically. We addressed this question in Chapter 6 with the spectral rays and orbits. Here we take a complementary approach. We introduce three maps $\mathbb{C} \mapsto \mathbb{R}^+$ associated with $z \mapsto M_z$, that we call "portraits".

7.2.1 Three portraits for $z \mapsto M_z$

These are visualisation tools in 2D and 3D , that is maps $\mathbb{C} \mapsto \mathbb{R}^+$ represented in $\mathbb{R}^2$ or $\mathbb{R}^3$, which attempt to give a qualitative representation of the map $z \mapsto M_z$, indicating in particular the location of its singularities (M_z and M_z^{-1} not defined). We shall consider specifically:

1. the *normwise portrait* $\phi_0 : z \mapsto \|M_z\|$, $z \in re(A)$,
2. the *spectral portrait* $\phi_1 : z \mapsto \rho(M_z)$, $z \in re(A)$,
3. the *frontier portrait* $\phi_2 : z \mapsto \rho(M_z^{-1})$, $z \in re(A) \backslash F(A, E)$.

We recall that $\rho(M_z) \leq \|M_z\|$. The portrait ϕ_1 (resp. ϕ_2) is related to the analyticity domain of $R(t, z)$ around 0 (resp. ∞).

We have introduced the portraits ϕ_0 and ϕ_1 in section 5.6 where we compared ϕ_0 with the well-known normwise spectral map $\psi : z \longmapsto \|(A - zI)^{-1}\|$. See more in Section 7.2.2. In section 7.2.3, we discuss the properties of the *frontier portrait*. In sections 7.2.4 and 7.2.5, we use some numerical examples to illustrate the abilities of the homotopic spectral and frontier portraits in detecting the eigenvalues and the frontier points.

The notions of peaks and wells have been defined in Chapter 5 , Definition 5.6.1. In what follows, we shall use these notions for the frontier portraits as well.

Remark. As it was mentioned in section 5.6, in practice and in what follows, we plot the log_{10} of the normwise portraits, spectral portraits, and frontier portraits. But, for simplicity, we do not write log_{10} for the plots of $log_{10}(\|M_z\|)$, $log_{10}(\rho(M_z))$ and $log_{10}(\rho(M_z^{-1}))$.

7.2.2 The normwise and spectral portraits ϕ_0 and ϕ_1

These two portraits introduced in Chapter 5 both give global informations about the map $z \mapsto M_z$, which are qualitatively different. We study these differences.

Proposition 7.2.1 *[5] ϕ_0 has a peak at $\lambda \in \sigma(A)$ if λ is $\|\cdot\|-$ observable. It cannot have wells.*

Proposition 7.2.2 *[5, 22] ϕ_1 has a peak at $\lambda \in \sigma(A)$ if λ is partially observable. It has a well at $z \in \bar{C}(A,E)$ when the set is nonempty. It has sinks at some $z \in \mathbb{C} \setminus \{\lambda \in \sigma(A),\ \lambda$ partially observable$\}$.*

It follows that if ϕ_1 has a sink at an eigenvalue λ, then λ is necessarily $\sigma-$ nonobservable. If $\lambda \in \sigma_c$, then the sink becomes a well.

Example 7.2.1 *Let us return to the Example 3.9.3 where for*

$$A = \begin{bmatrix} 0 & 0 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, E = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix},$$

$\pi(z) = z^3 + 1$, $\sigma(A) = \{-1,\ 0.5000 \pm 0.8660i\}$, *and*
$U = [e_1, e_2]$, $V = [e_2, e_3]$, *we get*

$$Q(z) = \begin{bmatrix} z & z^2 \\ 1 & z \end{bmatrix}$$

with $\det Q(z) \equiv 0$, *hence* $\hat{\pi}(z) \equiv 0$ *and* $Z(\det Q(z)) = \hat{Z} = \mathbb{C}$. *For* $z \notin \sigma(A)$, $M_z = \frac{1}{\pi(z)} \begin{bmatrix} z & z^2 \\ 1 & z \end{bmatrix}$ *has the spectrum* $\sigma(M_z) = \{0, \frac{2z}{z^3+1}\}$.
Therefore, $F(A,E) = re(A) \subset Z(\det Q(z))$ *and* $C(A,E) = \{0\}$.

The computation $t \mapsto \sigma(A(t))$ displayed on Figure 7.3 supports that one of the eigenvalues $\lambda(t)$ tends to 0 and the other two tend to ∞ as $|t| \to \infty$.

The matrix M_z does not defined at $\lambda \in \sigma(A)$. The 2D *and* 3D *normwise portraits $\phi_0 : z \mapsto \|M_z\|$ on Figures 7.4, (a) and (b) display* 3 *peaks corresponding to the* 3 *eigenvalues of A, confirming that the* 3 *eigenvalues are $\|\cdot\|-$ observable.*

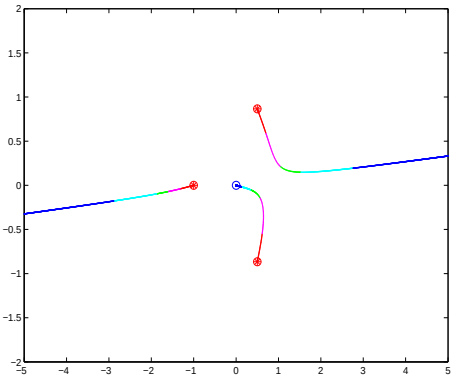
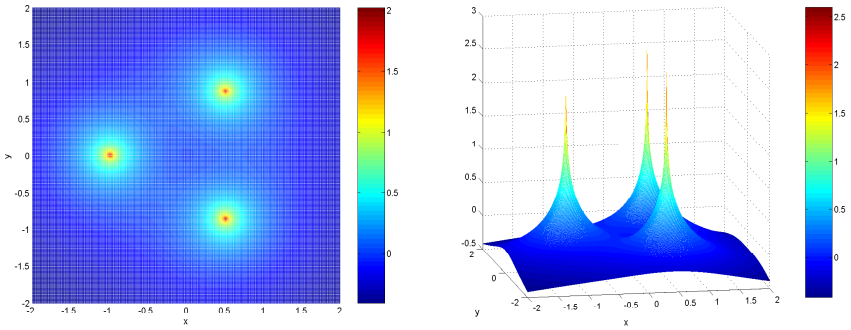
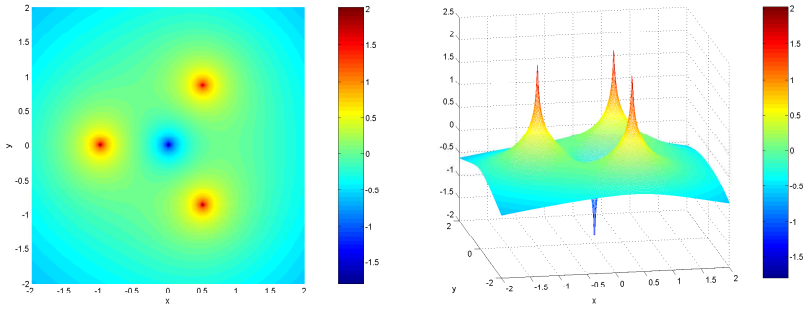


Figure 7.3: $t \mapsto \sigma(A(t))$ for $|t| = h \in [0,\ 300]$ and $\theta = \pi/24$



(a) 2D -normwise portrait (b) 3D -normwise portrait

Figure 7.4: Normwise portrait $\phi_0 \ : \ z \mapsto \|M_z\|$



(a) 2D -spectral portrait (b) 3D -spectral portrait

Figure 7.5: Spectral portrait $\phi_1 : z \mapsto \rho(M_z)$

Also the 3 *eigenvalues of* A *are partially observable:* $\lim_{z\to\lambda} |\mu_{iz}| = \{0,\ \infty\}$. *The* 2D *and* 3D *spectral portraits* $\phi_1 : z \mapsto \rho(M_z)$ *on Figures 7.5,* (*a*) *and* (*b*) *display*

3 peaks corresponding to the 3 eigenvalues of A and a well corresponding to the critical point 0, as expected.

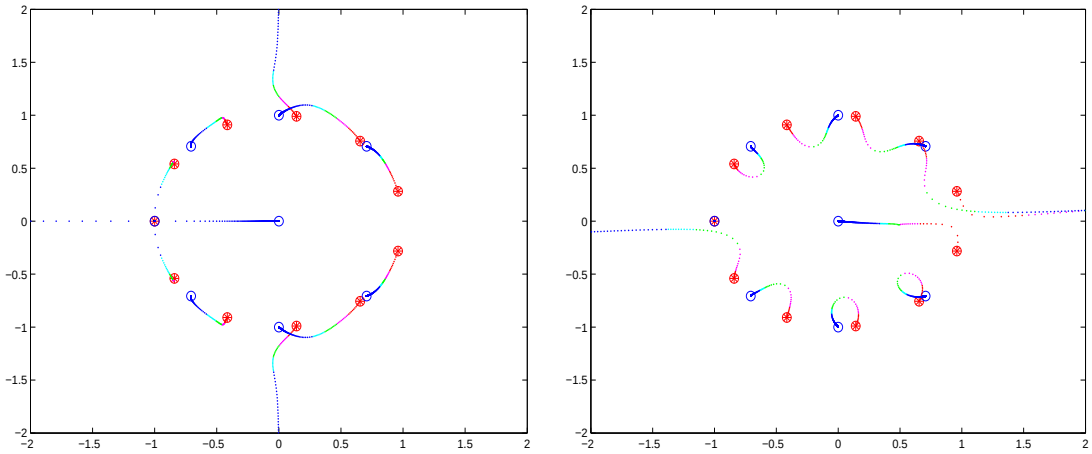
$\triangle$

Example 7.2.2 *Let us go back to the Example 7.1.1.*
The computation $t \mapsto \sigma(A(t))$ displayed on Figure 7.6 supports that 8 eigenvalues $\lambda(t)$ tend to the 8 points in Lim *and the other three tend to ∞ as $|t| \to \infty$.*

All the 11 eigenvalues of A are $\|\cdot\|-$observable. The 10 eigenvalues $\sigma(A)\backslash\{-1\}$ are partially observable but the eigenvalue $-1 \in \sigma(A)$ is $\sigma-$ nonobservable. No well is present: $\bar{C}(A, E) = \emptyset$.

The Figures 7.7 and 7.8 display the normwise and spectral portraits respectively. On Figure 7.7, (a) and (b), one can see 11 peaks corresponding to the eigenvalues of A. There is no well appearing on Figure 7.7, (c), nor on Figure 7.8. The portrait ϕ_1 displays 2 sinks on Figures 7.8, (c).

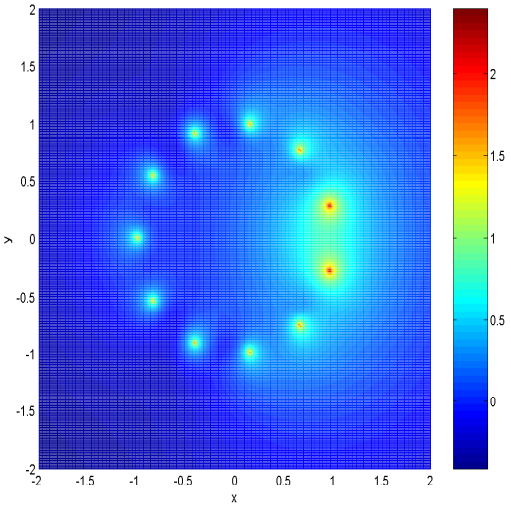
Figures 7.8, (a) and (b), indicate that there is no peak corresponding to the common point $\lambda = -1$ but there are 10 peaks corresponding to $\sigma(A)\backslash\{-1\}$. Figures 7.8, (c), and (d) show that there is no well corresponding to $\lambda \in \sigma(A)$. There are sinks at some $z \in re(A)$, but not at the $\sigma-$ nonobservable eigenvalue $\lambda = -1$.



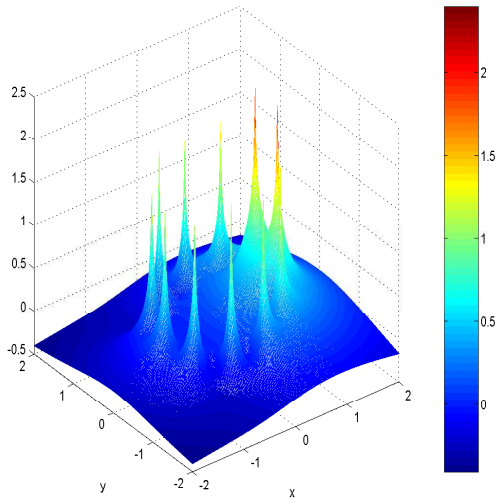
(a) $\theta = \pi$ (b) $\theta = 0.1$

Figure 7.6: $t \mapsto \sigma(A)$ for $h = |t| \in [0,\ 1000]$

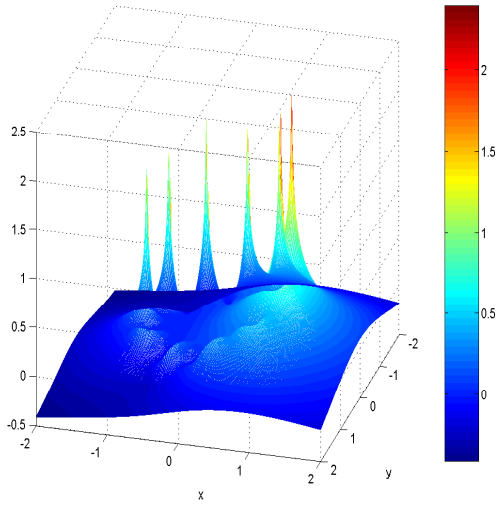
Observe that the absence of peak at $\lambda = -1$ modifies the form of the portrait ϕ_1 compared to the portrait ϕ_0 in the neighbourhood of -1.



(a) 2D -normwise portrait

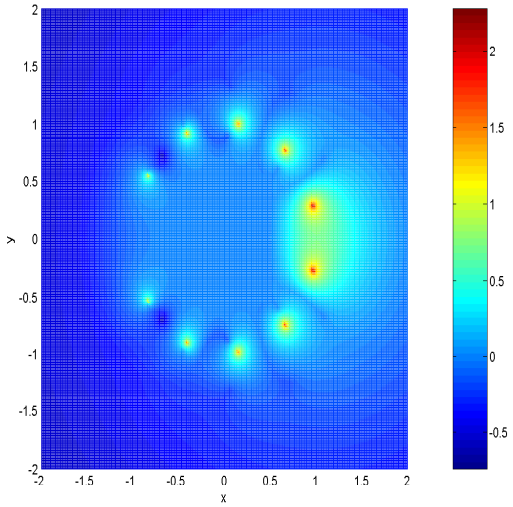


(b) 3D -normwise portrait

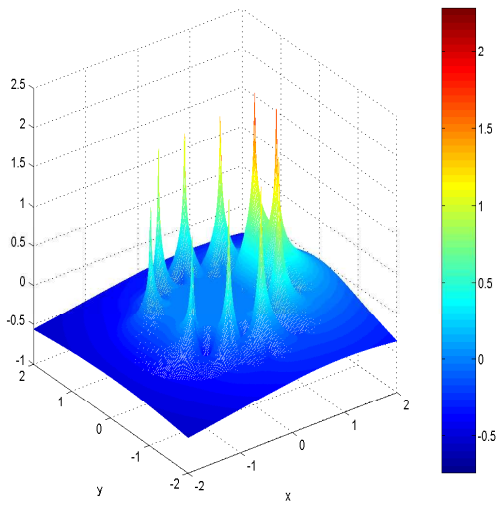


(c) 3D -normwise portrait from below

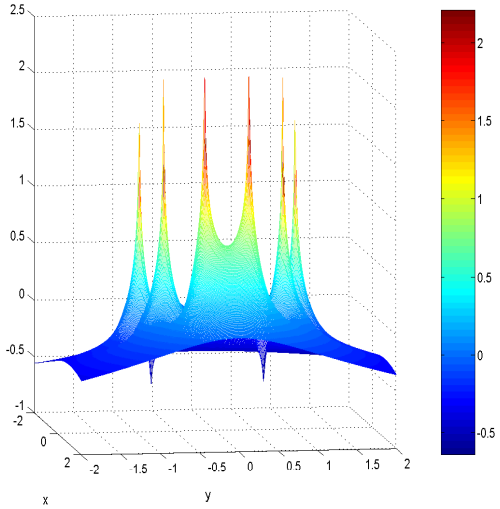
Figure 7.7: Normwise portrait $\phi_0 : z \mapsto \|M_z\|$



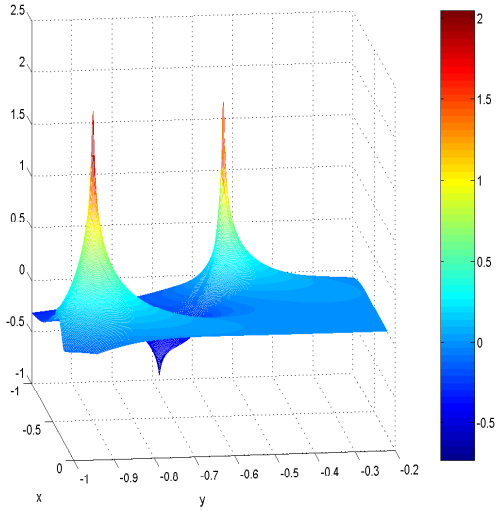
(a) 2D -spectral portrait



(b) 3D -spectral portrait



(c) 3D -spectral portrait from below



(d) Zoom around the left sink, near a $z \in F(A, E)$

Figure 7.8: Spectral portrait $\phi_1 : z \mapsto \rho(M_z)$

△

7.2.3 The frontier portrait ϕ_2

The *frontier portrait* is the map $z \mapsto \phi_2(z) = \rho(M_z^{-1})$ which is defined for $z \in re(A)$ at $z \notin F(A,E)$ [18, 21]. For $r = 1$, we have trivially $\rho(M_z) = 1/\rho(M_z^{-1})$. It is non trivial for $r \geq 2$. The study below follows [22].

Proposition 7.2.3 *[22] For $r \geq 2$, ϕ_2 has a peak at $z \in \bar{F}(A,E)$ and a well at $\lambda \in \sigma(A)$ if it is $\sigma-$ observable. It has sinks between peaks for some $z \in \mathbb{C}\backslash\bar{F}(A,E)$.*

It follows that if ϕ_2 has a sink at an eigenvalue λ, then λ is partially observable. If λ is $\sigma-$observable then the sink becomes a well.

For $r = 1$, the situation simplifies because $\vartheta(z) = \mu_z$, $\phi_0 = \phi_1$ and $\phi_2 = \phi_1^{-1} : z \mapsto \frac{1}{|\mu_z|}$. The notions of $\|\cdot\|-$observability and $\sigma-$observability coalesce.

Proposition 7.2.4 *[22] For $r = 1$, $\phi_2 = \phi_1^{-1}$ has no sink. It has peaks at $z \in \bar{F}(A,E) = \bar{C}(A,E)$ and wells at observable eigenvalues.*

It is remarkable that neither $\phi_0 = \phi_1$ nor $\phi_2 = \phi_1^{-1}$ presents any sink for $r = 1$. The general situation $r \geq 2$ is more complex with peaks, sinks and possibly wells.

For $r = 2$, M_z has 2 eigenvalues and it is easy to predict algebraically when λ is a $\sigma-$observable eigenvalue. See Section 7.3. For $r \geq 3$, the algebraic prediction is more difficult. It can be useful to zoom around an eigenvalue to appreciate whether the computed sink in the neighbourhood of an eigenvalue can actually be a well at the eigenvalue which is therefore $\sigma-$observable.

Lemma 7.2.5 *[22] For $z \in re(A)\backslash F(A,E)$ the inverse M_z^{-1} exists and satisfies the equivalent identities*

$$M_z^{-1} = \alpha(z)\mathrm{adj}Q(z) = \beta(z)\mathrm{adj}M_z,$$

where the coefficients $\alpha(z) = \frac{1}{(\pi(z))^{r-2}\hat{\pi}(z)}$ and $\beta(z) = \frac{1}{\vartheta(z)} = \frac{\pi(z)}{\hat{\pi}(z)} = \frac{1}{\det M_z}$ are meromorphic functions of $z \in \mathbb{C}$ for $1 \leq r < n$.

Proof. We recall that for the matrix M of order $r \geq 1$, $\mathrm{adj}\,(\beta M) = \beta^{r-1}\,\mathrm{adj}\,M$. Therefore $M_z = \frac{1}{\pi(z)}Q(z)$ implies, for $z \in re(A)\backslash F(A,E)$, that

$$M_z^{-1} = \pi(z)(Q(z))^{-1} = \frac{\pi(z)}{\det Q(z)}\mathrm{adj}Q(z) = \frac{1}{(\pi(z))^{r-2}\hat{\pi}(z)}\mathrm{adj}Q(z),$$

where $\alpha(z) = \frac{1}{(\pi(z))^{r-2}\hat{\pi}(z)}$ depends on r.
Equally $\mathrm{adj}\,M_z = \frac{1}{(\pi(z))^{r-1}}\,\mathrm{adj}\,Q(z)$ for $z \in re(A)$. Here $\beta(z) = \frac{1}{\det M_z} = \frac{\pi(z)}{\hat{\pi}(z)}$ is independent of r.

□

r	1	2	≥ 3
$\alpha(z)$	$\frac{\pi(z)}{\hat{\pi}(z)} = \beta(z)$	$\frac{1}{\hat{\pi}(z)}$	$\frac{1}{\pi(z)^{r-2}\hat{\pi}(z)}$

Table 7.2: Correspondence $r \mapsto \alpha(z)$

The functions $\alpha(z)$ and $\beta(z)$ are meromorphic in $\mathbb{C}$. And we observe that $\beta(z)$ is independent of r, whereas $\alpha(z)$ depends on r in the way that illustrated in Table 7.2.
The coefficient $\beta(z)$ relates M_z^{-1} to the adjoint matrix $\operatorname{adj} M_z$. Similarly $\alpha(z)$ relates M_z^{-1} to the *iterated* adjoint matrix $\operatorname{adj} Q(z) = \operatorname{adj}(V^H \operatorname{adj}(zI - A)U)$. Observe that for $r = 1$, $\operatorname{adj} Q(z) = \operatorname{adj} M_z = 1$.

Because $\alpha(z) \neq \beta(z)$ for $r \geq 2$, the question arises: are the peaks of the frontier portrait ϕ_2 related to the poles of $\alpha(z)$ or of $\beta(z)$? As a conclusion of the above theoretical study, one should remark that ϕ_2 follows $\beta(z)$ and not $\alpha(z)$. This fact is supported by the examples given below.

We begin with the particular case $r = 1$, $\phi_2 = \phi_1^{-1} : z \mapsto \frac{1}{|\mu_z|}$. ϕ_2 has peaks at $\bar{F}(A, E)$ and wells at any λ which is observable. $\alpha(z) = \beta(z) = \frac{\pi(z)}{\hat{\pi}(z)} = \frac{1}{\mu_z}$.

Example 7.2.3 *Let*

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 2 & 0 & 0 & 5 & 1 \\ 0 & 10 & 1 & 0 & 0 & 0 \\ 7 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \text{ and } E = uv^T \text{ with } u = e_1, \text{ and } v = e_6.$$

Here, $r = \operatorname{rank} E = 1$, $v^T u = 0$ *and* $0 \in \sigma(E)$ *is defective with*
$n_1 = 2 > n_2 = 1$, *for* $r_1 = 1$, $r_2 = 4$, *and* $q = 2$.
Then

$$\pi(z) = z^6 - 6z^5 + 13z^4 - 73z^3 + 61z^2 + 56z - 302,$$

$$\hat{\pi}(z) = z^4 - 4z^3 + 12z^2 - 73z + 314 = \det Q(z) = v^H \operatorname{adj}(zI - A)u,$$

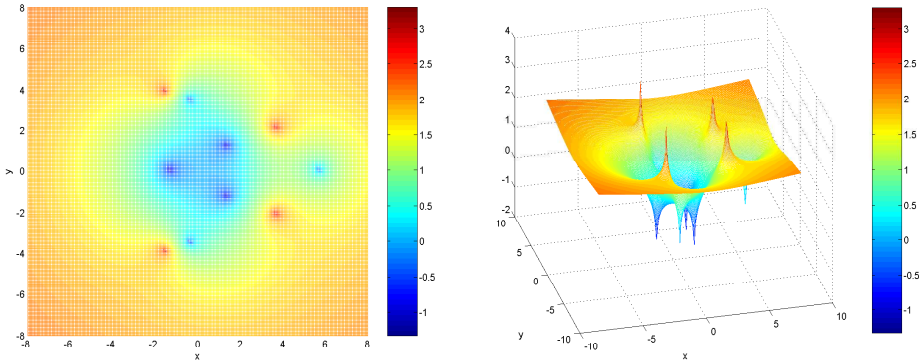
We calculate that $\hat{Z} \cap \sigma(A) = \emptyset$ *and* $F(A, E) = \hat{Z} = \{-1.6193 \pm 3.9117i,\ 3.6193 \pm 2.1024i\}$ *has an empty intersection with* $\sigma(A) = \{5.6461,\ -0.4265 \pm 3.4905i,\ 1.2853 \pm 1.2327i,\ -1.3638\}$.

All the 6 *eigenvalues of* A *are observable:* $\lim_{z \to \lambda} |\mu_z| = \infty$, *for* $\lambda \in \sigma(A)$.

In Figures 7.9, (a) and (b), we use the 2D*-frontier and the* 3D*-frontier portraits respectively to illustrate that the frontier portrait has* 4 *peaks at the frontier points in* $F(A,E)$ *and* 6 *wells at* $\lambda \in \sigma(A)$. *Here no simplification takes place:* $\hat{m}_\lambda = 0 < m_\lambda = 1$. *The rational functions*

$$\alpha(z) = \frac{z^6 - 6z^5 + 13z^4 - 73z^3 + 61z^2 + 56z - 302}{z^4 - 4z^3 + 12z^2 - 73z + 314} = \beta(z).$$

are defined for $z \notin F(A,E)$.



(a) 2D -frontier portrait (b) 3D -frontier portrait

Figure 7.9: Frontier portrait $\phi_2 : z \mapsto \rho(M_z^{-1})$

△

Example 7.2.4 *Let*

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & -2 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix} \text{ and } E = uv^T \text{ with } u = e_6 = v\,.$$

Here, $r = \operatorname{rank} E = 1$, $v^T u = 1$ *and* $0 \in \sigma(E)$ *is semi-simple. Then*

$$\pi(z) = z(z^5 - z^4 - 5z^3 + 2z^2 + 2z - 1)\,,$$

$$\hat{\pi}(z) = z(z+2)(z-1)^3 = \det Q(z) = d(z)\,,$$

We calculate that $\hat{Z} \cap \sigma(A) = \{0\}$, *for* $\sigma(A) = \{-1.8650,\ -0.7199,\ 0,\ 0.5152 \pm 0.1617i,\ 2.5544\}$ *and* $F(A,E) = \{-2,\ (1)^3\}$.

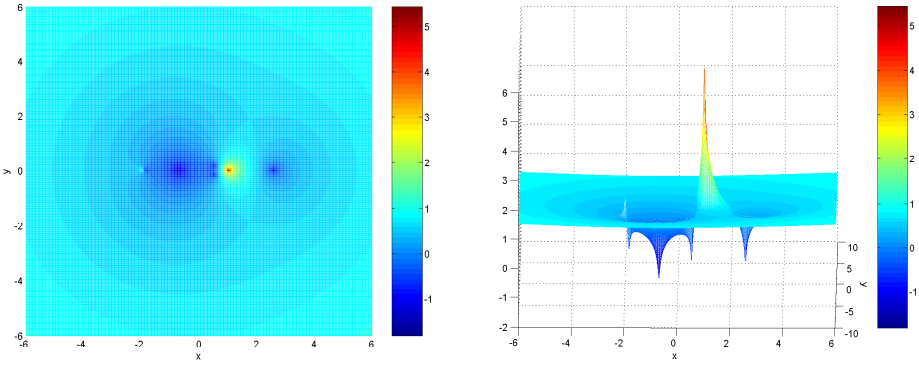
The 5 *eigenvalues* $\lambda \in \sigma(A)\backslash\{0\}$ *are observable:* $\lim_{z\to\lambda} |\mu_z| = \infty$, *for* $\lambda \in \sigma(A)\backslash\{0\}$. *Therefore they are normwise observable too. The eigenvalue* $\lambda = 0$

is nonobservable: $\lim_{z\to 0} |\mu_z| = 2 < \infty$. *In addition,* $\det Q(z) = Q(z)$, *therefore* $M_z = \mu_z = \frac{Q(z)}{\pi(z)}$ *is defined at* $\lambda = 0$ *which means that* $\lambda = 0$ *is nonobservable. Here we have*

$$\alpha(z) = \frac{z(z^5 - z^4 - 5z^3 + 2z^2 + 2z - 1)}{z(z+2)(z-1)^3} = \frac{(z^5 - z^4 - 5z^3 + 2z^2 + 2z - 1)}{(z+2)(z-1)^3} = \beta(z).$$

The simplification by z *entails that* $\alpha(\lambda) = \beta(\lambda) = 0$ *for* $0 \neq \lambda \in \sigma(A)$.

In Figures 7.10, (a) *and* (b), *we use the* 2D*-frontier and the* 3D*-frontier portraits respectively to illustrate that the frontier portrait has* 2 *peaks at the* 2 *distinct frontier points* $F(A, E) = \{-2,\ (1)^3\}$ *and* 5 *wells at* $\lambda \in \sigma(A)\backslash\{0\}$.



(a) 2D -frontier portrait (b) 3D -frontier portrait

Figure 7.10: Frontier portrait $\phi_2 : z \mapsto \rho(M_z^{-1})$

Since 1 *is triple and* -2 *simple in* $F(A, E)$, *it is difficult to capture the peak corresponding to* -2 *in the* 3D *frontier portrait* ϕ_2 *in finite precision.*

△

Remark. When the rank of matrix E is greater than 2, it becomes clear that the spectral and frontier portraits are valuable tools to detect the partially observable eigenvalues, the frontier points and the set Lim.

Example 7.2.5 *Let us return to the Example 4.3.1 where for*

$$A = \begin{bmatrix} 1 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 2 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \text{ and } E = \begin{bmatrix} 0 & 4 & 1 & 2 & 0 & 0 \\ -1 & 3 & 4 & 2 & 0 & 1 \\ 0 & 4 & 1 & 2 & 0 & 0 \\ -1 & 4 & 3 & 3 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 & 0 & 0 \end{bmatrix}, \text{ we use (1.6.1)}$$

to get $E = UV^T$ *with*

$$U = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \text{and } V = \begin{bmatrix} 0 & 0 & -1 \\ 3 & 1 & 0 \\ 2 & -1 & 2 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Here, $r = \text{rank}\, E = 3$, $\det(V^T U) = -15 \neq 0$ *and* $0 \in \sigma(E)$ *is semi-simple with the geometric multiplicity* $g = 3$. *Then*

$$\pi(z) = z(z^5 - 2z^4 - 4z^3 + 1), \quad \hat{\pi}(z) = -(15z^3 + 8z^2 - 8z + 1),$$

and

$$\det Q(z) = -(15z^3 + 8z^2 - 8z + 1)(z^5 - 2z^4 - 4z^3 + 1)^2 z^2 = (\pi(z))^2 \hat{\pi}(z).$$

Then $F(A, E) = \{-1.0828,\ 0.1568,\ 0.3926\}$ *has an empty intersection with* $\sigma(A) = \{-1.3293,\ 0,\ 0.5915,\ 3.2294,\ -0.2458 \pm 0.5774i\}$.

In Figures 7.11, (a), (b), *we use the* 2D*-frontier and the* 3D*-frontier portraits respectively to illustrate that the frontier portrait has* 3 *peaks at the frontier points. There is no well on* ϕ_2, *but we can see sinks at* $\{0.5915,\ -0.2458 \pm 0.5774i\}$. *There are also* 2 *sinks around* -1.3293. *See Figure 7.12 . Here*

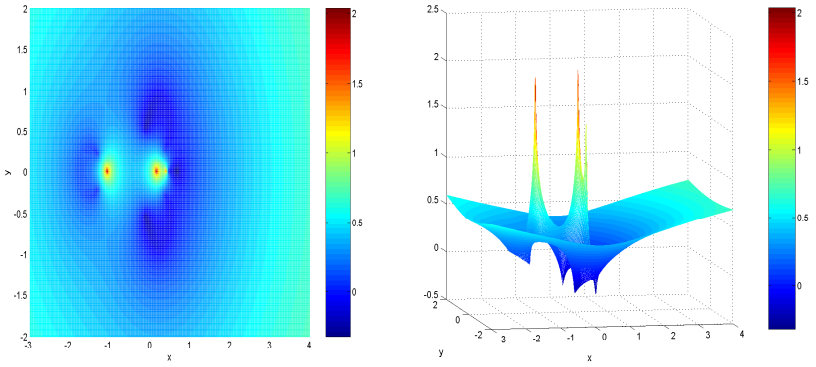
$$\alpha(z) = \frac{1}{-z(z^5 - 2z^4 - 4z^3 + 1)(15z^3 + 8z^2 - 8z + 1)},$$

is not defined for $\lambda \in \sigma(A) \cup \hat{Z}$, *and*

$$\beta(z) = \frac{z(z^5 - 2z^4 - 4z^3 + 1)}{-(15z^3 + 8z^2 - 8z + 1)},$$

is not defined for $z \in \hat{Z}$.

We have $\beta(\lambda) = 0, \quad \forall\ \lambda \in \sigma(A)$, *that is* $\vartheta(\lambda)$ *does not exist. Therefore all* 6 *eigenvalues are partially observable. See the spectral portrait* ϕ_1 *on Figure 7.13 with* 6 *peaks corresponding to* $\lambda \in \sigma(A)$. *There is also a sink at a real value between* $0.3926 \in F(A, E)$ *and* $\lambda = 0.5915 \in \sigma(A)$ *for* ϕ_1. *See Figure 7.14,* (a) *and a zoom around the real sink in Figure 7.14,* (b).



(a) 2D-frontier portrait (b) 3D-frontier portrait from above

Figure 7.11: Frontier portrait $\phi_2 : z \mapsto \rho(M_z^{-1})$

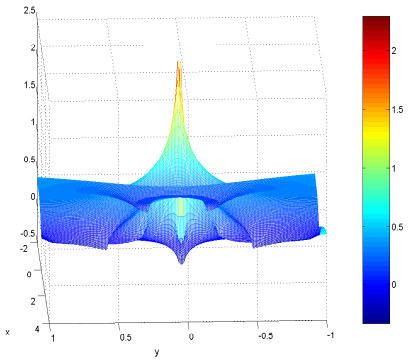
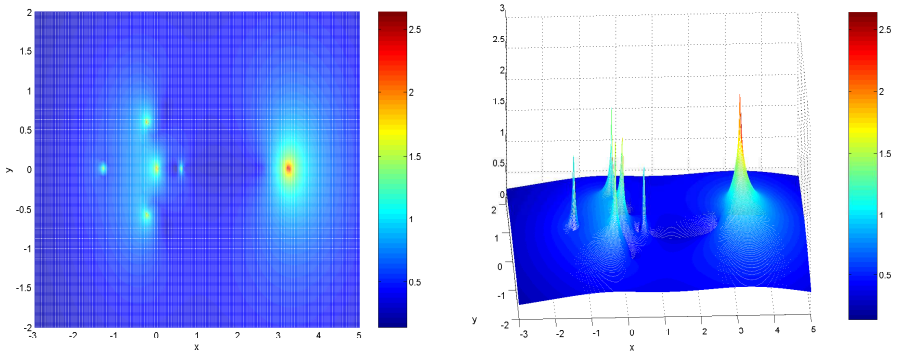
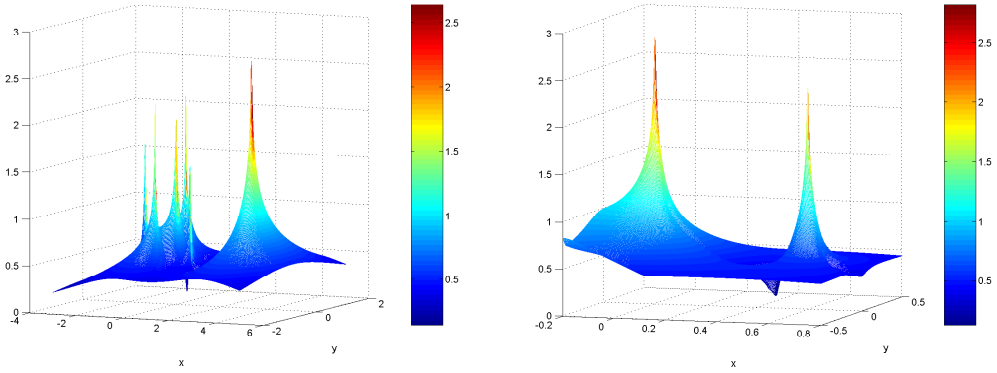


Figure 7.12: Frontier portrait $\phi_2 : z \mapsto \rho(M_z^{-1})$ from below



(a) 2D-spectral portrait (b) 3D-spectral portrait from above

Figure 7.13: Spectral portrait $\phi_1 : z \mapsto \rho(M_z)$



(a) From below (b) Zoom around $z = 0.3926$ and $\lambda = 0.5915$

Figure 7.14: Spectral portrait $\phi_1 : z \mapsto \rho(M_z)$

△

Example 7.2.6 *Let*

$$A = \begin{bmatrix} 1 & 0 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \end{bmatrix} \text{ and } E = \begin{bmatrix} 0 & 2 & 1 & 2 & 0 & 0 \\ -1 & 1 & 4 & 2 & 0 & 1 \\ 0 & 2 & 1 & 2 & 0 & 0 \\ -1 & 2 & 3 & 3 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 & 0 & 0 \end{bmatrix}, \text{ we use}$$

(1.6.1) to get $E = UV^T$ *with*

$$U = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \text{and } V = \begin{bmatrix} 0 & 0 & -1 \\ 1 & 1 & 0 \\ 2 & -1 & 2 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

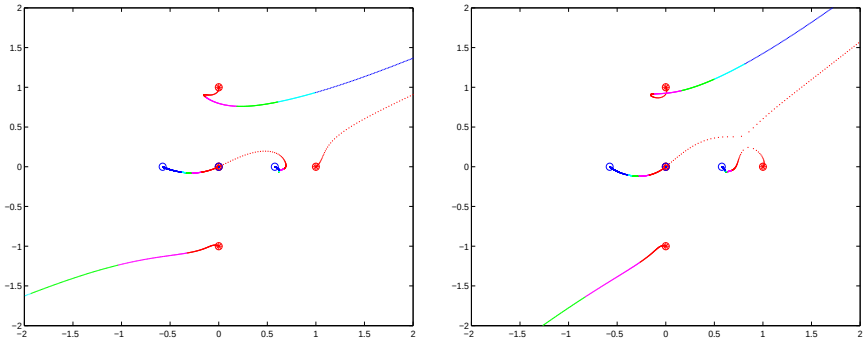
Here, $r = \operatorname{rank} E = 3$, $\det(V^T U) = -9 \neq 0$ *and* $0 \in \sigma(E)$ *is semi-simple with the geometric multiplicity* $g = 3$. *Then*

$$\pi(z) = z^3(z-1)(z^2+1), \quad \hat{\pi}(z) = -3z(3z^2-1),$$

and

$$\det Q(z) = (\pi(z))^2 \hat{\pi}(z).$$

$\sigma(A) \cap \hat{Z} = \{0\}$ *with* $m_0 = 3 > \hat{m}_0 = 1$. $\hat{Z} = F(A,E) \cup \{0\} = \{\pm 0.5774,\ 0\}$.



(a) $\theta = \pi/6$ (b) $\theta = \pi/4$

Figure 7.15: $t \mapsto \sigma(A)$ for $h = |t| \in [0,\ 300]$

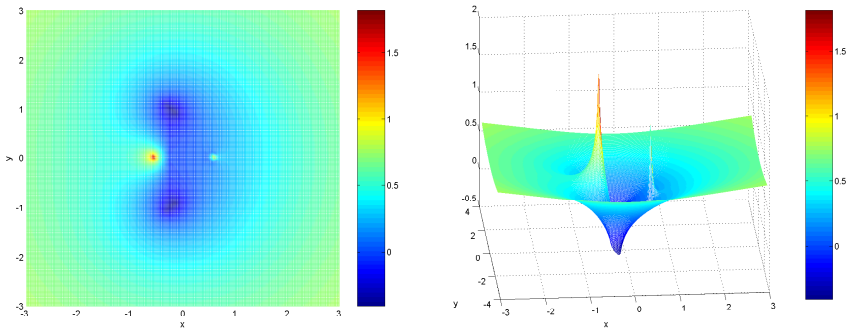
The Figures 7.15, (a) and (b) show that $\lambda = 0$ *as a common root of* $\pi(z)$ *and* $\hat{\pi}(z)$ *belongs to* Lim*: one copy of the triple* $0 \in \sigma(A)$ *is invariant, or else* $0 \in \sigma^i$ *.*

In Figures 7.16, (a) , (b) , we use the 2D*-frontier and the* 3D*-frontier portraits respectively to illustrate that the frontier portrait has* 2 *peaks at* $F(A, E) = \hat{Z}\backslash\{0\}$ *. A zoom near* $\pm i$ *shows that* ϕ_2 *has sinks near* $\pm i$ *. See Figures 7.17, (a) and (b) .*

Now, after simplifications

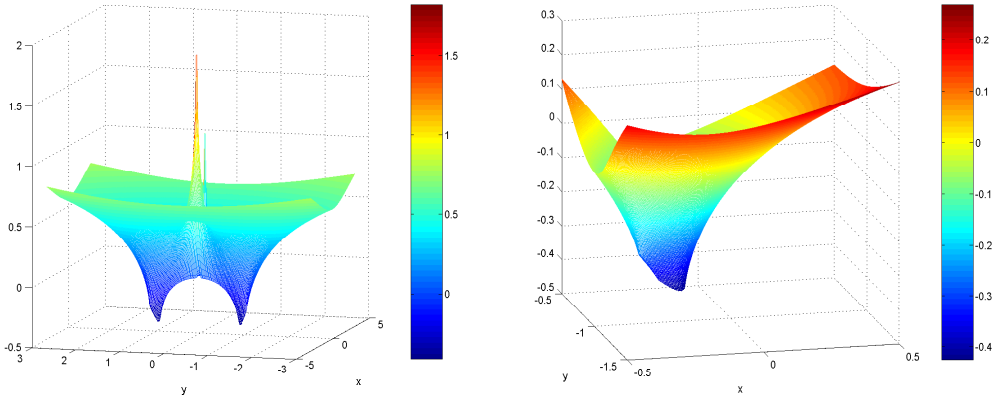
$$\alpha(z) = \frac{1}{-3z^4(3z^2-1)(z-1)(z^2+1)},\quad \beta(z) = \frac{z^2(z-1)(z^2+1)}{-3(3z^2-1)}.$$

$\beta(\lambda) = 0$ *for* $\lambda \in \sigma(A)\backslash\{0\}$ *, that is* $\vartheta(\lambda)$ *is not defined: the* 3 *eigenvalues* $\{1,\ \pm i\}$ *are partially observable. For* $\lambda = 0$ *, one has* $m_0 = 3 > \hat{m}_0 = 1$ *: 0 is also partially observable.*



(a) 2D-frontier portrait (b) 3D-frontier portrait from above

Figure 7.16: Frontier portrait $\phi_2 : z \mapsto \rho(M_z^{-1})$



(a) 3D-frontier portrait from below (b) Zoom around $\lambda = -i$

Figure 7.17: Frontier portrait $\phi_2 : z \mapsto \rho(M_z^{-1})$ from below

△

7.2.4 More numerical illustrations

Example 7.2.7 *Let A, U, V be the matrices defined in Example 7.1.1 with $r = 2$. Then we consider $B = A + e_2 e_2^T$ which is a modification of A of rank 1, with characteristic polynomial $\pi(z) = z^{11} - z^{10} + 1$.*

We consider $M_z = V^T(zI - B)^{-1}U$ and

$$Q(z) = V^T \mathrm{adj}(zI - B)U = \left(\begin{array}{c|c} -z^{10} - 1 & -z \\ \hline -z^{10} - z^8 + z^7 + \cdots + 1 & -z^9 \end{array} \right),$$

$$\det Q(z) = \pi(z) z(z+1)(z^2+1)(z^4+1), \quad \mathrm{tr} Q(z) = -(z^{10} + z^9 + 1),$$

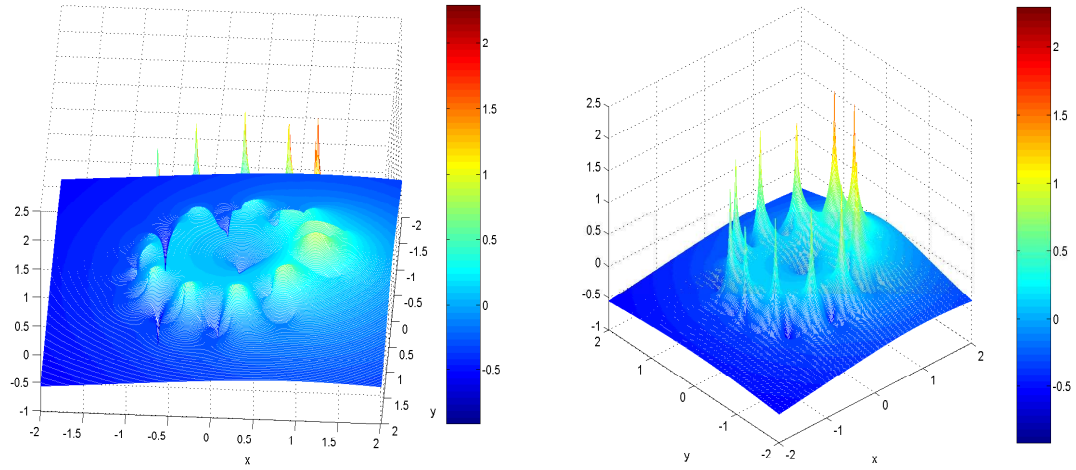
$$\hat{\pi}(z) = z(z+1)(z^2+1)(z^4+1),$$

where $\sigma(B) \cap \hat{Z} = \emptyset$ and $F(B, E) = \hat{Z}$.

One can see from the matrix polynomial $Q(z)$ that $\lim_{z \to \lambda} M_z = \lim_{z \to \lambda} \frac{Q(z)}{\pi(z)}$ does not exist for any $\lambda \in \sigma(B)$. Therefore the eigenvalues of B are normwise observable.

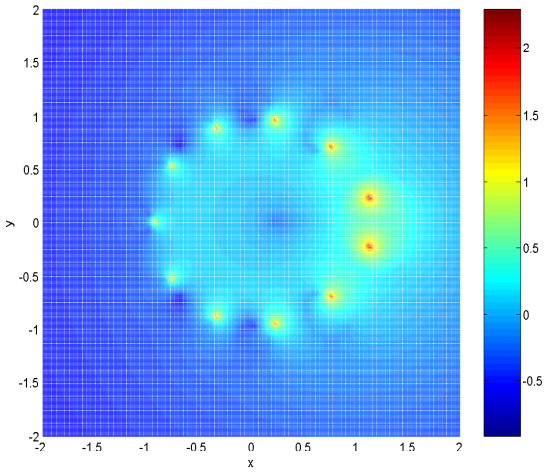
Figures 7.18, (a), (b) and (c) display the spectral portrait $\phi_1 : z \mapsto \rho(M_z)$ for (B, E). It has 11 peaks at the eigenvalues of B which indicate that they are at least partially observable. (See more in section 7.3 below).

The spectral portrait has 5 visible sinks: 2 near the roots of $z^2 + 1$, 2 near the two roots $\{-0.7071 \pm 0.7071i\}$ of $z^4 + 1$, and one near $z = 0$. See Figure 7.18, (a).



(a) 3D from below

(b) 3D from above



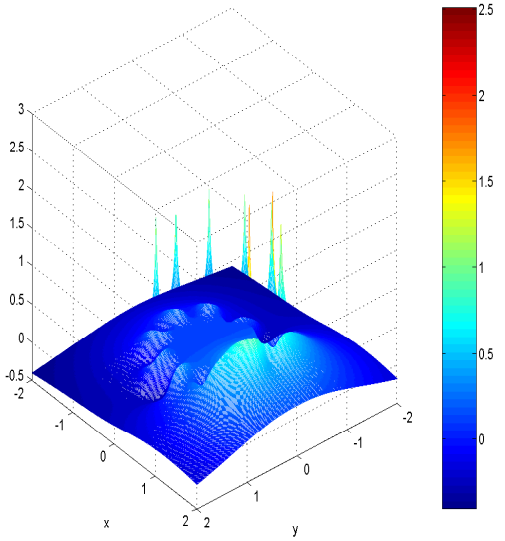
(c) 2D -spectral portrait

Figure 7.18: The spectral portrait $\phi_1 : z \mapsto \rho(M_z)$

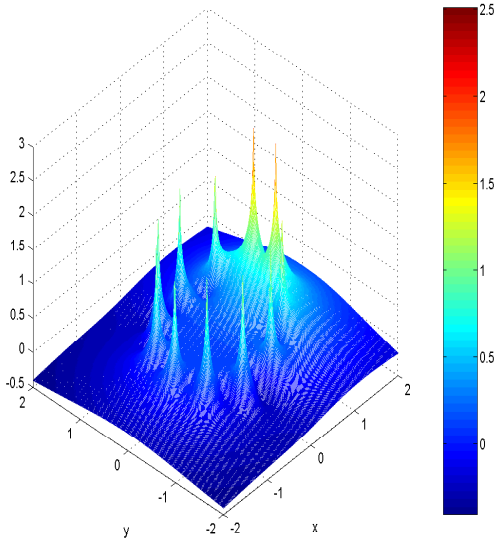
Look at the Figures 7.19, (b) and (c) to see that the normwise portrait $\phi_0 : z \mapsto \|M_z\|$ *has* 11 *peaks at the eigenvalues of* A. *It has no sink. See Figures 7.19, (a), and compare it with Figure 7.18, (a).*

The frontier portrait should display 8 *peaks at the frontier points in* $F(B, E) = \hat{Z} \cap re(B) = \hat{Z}$.

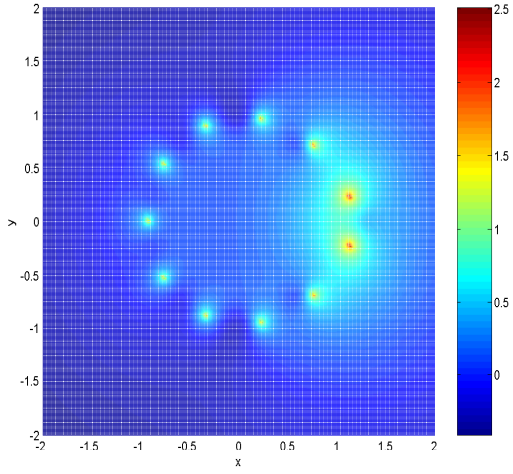
The values of $\alpha(z) = \frac{1}{\hat{\pi}(z)} = \beta(z)$ *are finite for the simple eigenvalues* $z = \lambda \in \sigma(B)$



(a) 3D from below



(b) 3D from above

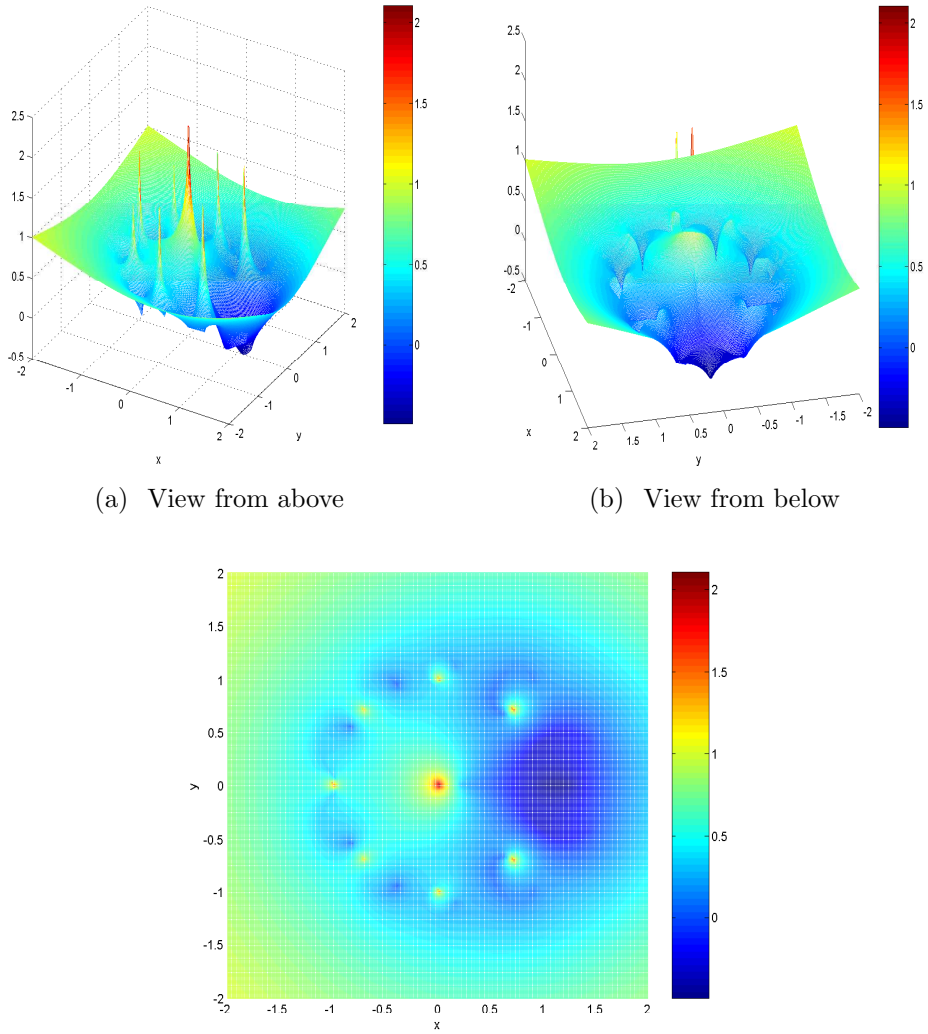


(c) 2D -normwise portrait

Figure 7.19: The normwise portrait $\phi_0 : z \mapsto \|M_z\|$

(with $\beta(\lambda) = 0$, $\alpha(\lambda) \neq 0$ *). Therefore* λ *is partially observable.*

Figures 7.20, (*a*) , (*b*) *and* (*c*) *display the frontier portrait* $\phi_2 : z \mapsto \rho(M_z^{-1})$ *for* (B, E) . *The* 8 *peaks at* $F(B, E)$ *are visible on Figure 7.20,* (*a*) . *On Figure 7.20,* (*b*) , *one can see the sinks near the eigenvalues of* A .



(a) View from above

(b) View from below

(c) 2D -frontier portrait

Figure 7.20: The frontier portrait $\phi_2 \; : \; z \mapsto \rho(M_z^{-1})$

△

Example 7.2.8 *Let A be the matrix B defined in Example 7.2.7. Let $E = UV^T$ with $U = [e \; e_2 \; e_4]$, $V = [e_{11} \; e_3 \; e_6]$ of rank $r = 3$, where $e = [1, \ldots, 1]^T$.* $\det(V^T U) = 0$. *Then we have*

$$\det Q(z) = z(z^4 + z^3 + z^2 + z + 1)(z^{11} - z^{10} + 1)^2 ,$$

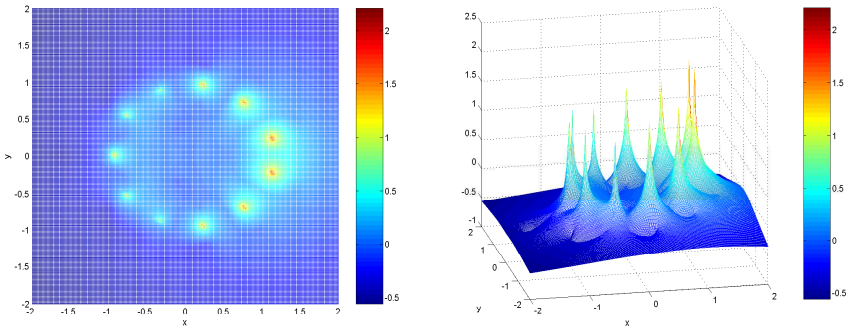
$$\pi(z) = z^{11} - z^{10} + 1 ,$$

$$\hat{\pi}(z) = z(z^4 + z^3 + z^2 + z + 1) .$$

Therefore $\sigma(A) \cap \hat{Z} = \emptyset$ for

$$\hat{Z} = F(A,E) = \{-0.8090 \pm 0.5878i,\ 0,\ 0.3090 \pm 0.9511i\}\,.$$

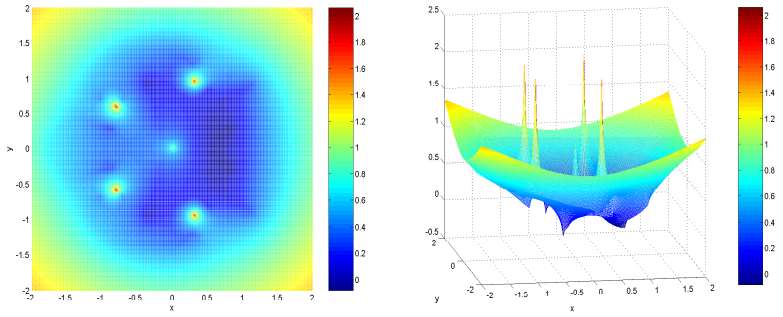
Figures 7.21, (a), (b) display the spectral portrait $\phi_1 : z \mapsto \rho(M_z)$ *for* (A,E). *It has* 11 *peaks at the eigenvalues of* A *which indicates that they are at least partially observable. There is no well on this portrait:* $C(A,E) = \emptyset$.



(a) 2D -spectral portrait (b) 3D -spectral portrait

Figure 7.21: The spectral portrait $\phi_1 : z \mapsto \rho(M_z)$

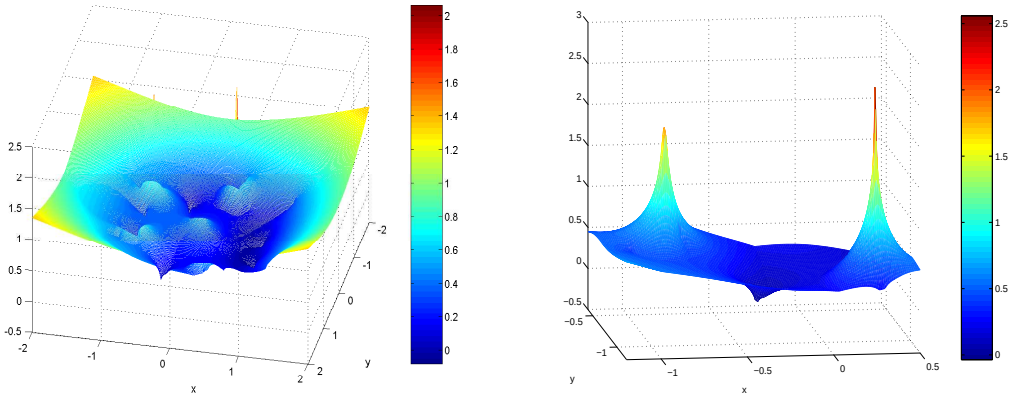
Figures 7.22, (a) and (b) display the frontier portrait for (A,E). *There are* 5 *peaks corresponding to the* 5 *frontier points in* $F(A,E)$*: the peak at* 0 *is less visible than the* 4 *others peaks.*



(a) 2D -frontier portrait (b) 3D -frontier portrait

Figure 7.22: The frontier portrait $\phi_2 : z \mapsto \rho(M_z^{-1})$

One sees local minima in finite precision between the peaks. See Figure 7.23 below. The local minima can be in $re(A)$ *or at partially observable eigenvalues. A zoom near each local minimum can provide numerical indication about the localization.*



(a) From below

(b) Zoom around $z = -0.8090 - 0.5878i$, and $z = 0.3090 - 0.9511i$

Figure 7.23: The frontier portrait $\phi_2 : z \mapsto \rho(M_z^{-1})$

△

Example 7.2.9 *Let A be the matrix B defined in Example 7.2.7. Let $E = UV^T$ with $U = [e\ e_2\ e_4\ e_7]$, $V = [e_{11}\ e_3\ e_6\ e_8]$ of rank $r = 4$, where $e = [1, \dots, 1]^T$ and* $\det(V^T U) = 0$. *Then we have*

$$\det Q(z) = z(z^2 + z + 1)(z^{11} - z^{10} + 1)^3,$$

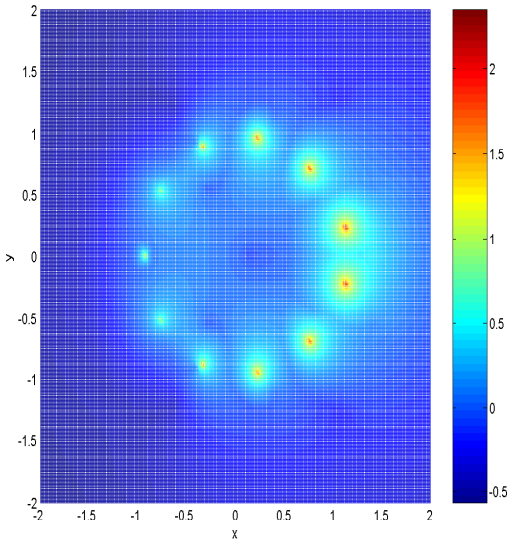
$$\pi(z) = z^{11} - z^{10} + 1,$$

$$\hat{\pi}(z) = z(z^2 + z + 1).$$

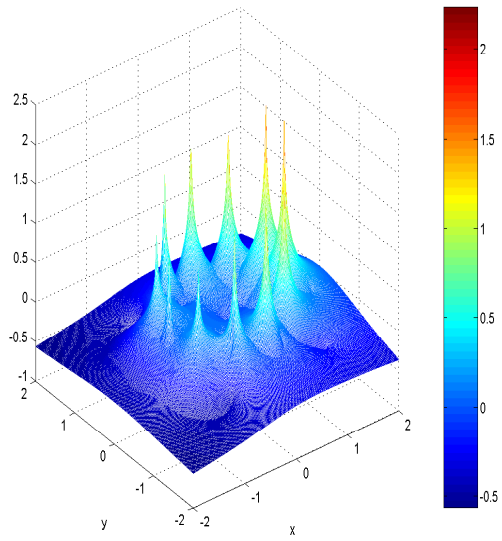
Therefore $\sigma(A) \cap \hat{Z} = \emptyset$ for $\hat{Z} = F(A, E) = \{\frac{-1}{2} - \frac{\sqrt{3}}{2}i,\ \frac{-1}{2} + \frac{\sqrt{3}}{2}i,\ 0\} = \{f_1,\ f_2,\ f_3\}$.

Figures 7.24, (a), (b) display the spectral portrait $\phi_1 : z \mapsto \rho(M_z)$ for (A, E). It has 11 peaks at the eigenvalues of A which indicates that they are at least partially observable. There is not any visible well on this portrait. Computation of $\rho(M_z) > 0$ for $z \in F(A, E)$ confirms that $C(A, E) = \emptyset$: $\rho(M_{f_k}) = 1.2660$ for $k = 1, 2$ and $\rho(M_{f_3}) = 1$. See Figure 7.24, (c).

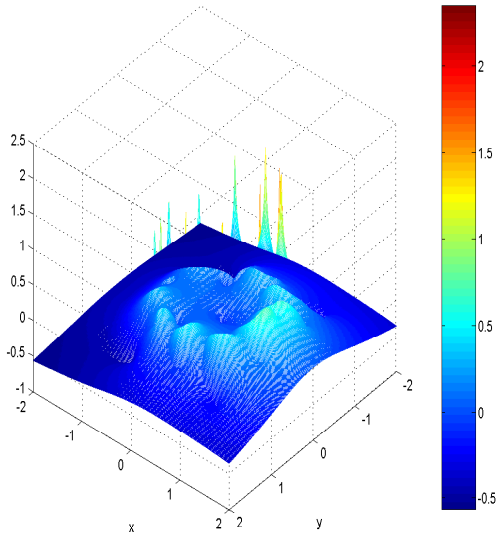
Figures 7.25, (a), (b) and (c) display the frontier portrait for (A, E). There are 3 peaks corresponding to the 3 frontier points in $F(A, E)$ (the peak at 0 is hard to see). There are sinks, some of them rather flat. See Figure 7.25, (c).



(a) 2D -spectral portrait

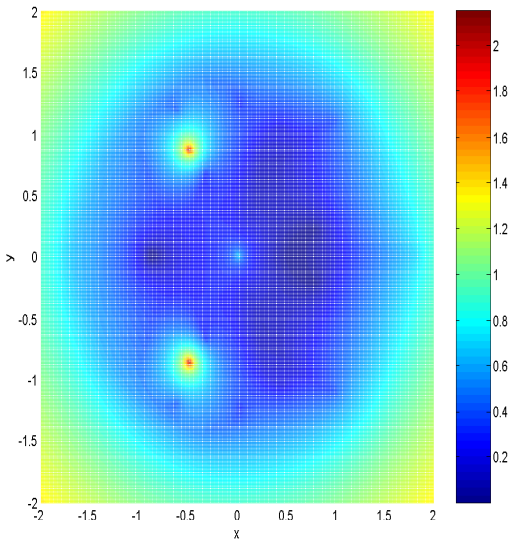


(b) 3D -spectral portrait

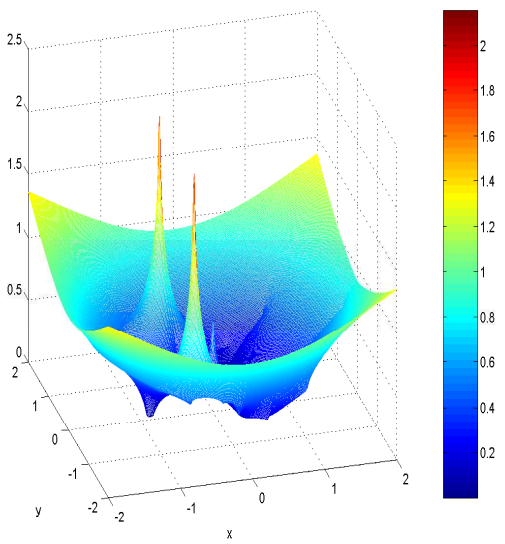


(c) 3D -spectral portrait from below

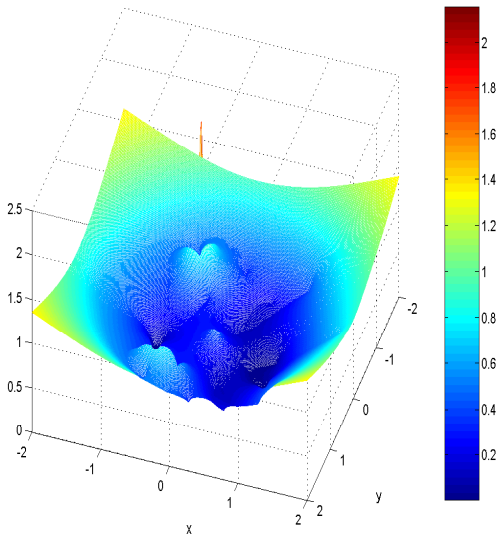
Figure 7.24: The spectral portrait $\phi_1 \ : \ z \mapsto \rho(M_z)$



(a) 2D -frontier portrait



(b) 3D -frontier portrait



(c) 3D -frontier portrait from below

Figure 7.25: The frontier portrait $\phi_2 \ : \ z \mapsto \rho(M_z^{-1})$

△

7.2.5 The intersection of spectral and frontier portraits

The line of balance

$$H = \left\{ z \in re(A) : \ \rho(M_z) = \rho(M_z^{-1}) \right\} \subset re(A)$$

has been introduced and studied in [18]. It is the projection onto $\mathbb{C}$ of the $3D$ curve $\mathcal{H}$ along which the spectral and frontier portraits intersect.

For any $z \in H$, the radii of convergence for $R(t, z)$ around 0 and ∞ are equal. For $r = 1$, and $z \in H$, then $|\mu_z| = 1$.

Figure 7.26 displays, for the Example 7.2.2, the spectral and frontier portraits intertwined. They are plotted separately on Figure 7.27 and 7.28 respectively, together with the intersection curve $\mathcal{H}$.

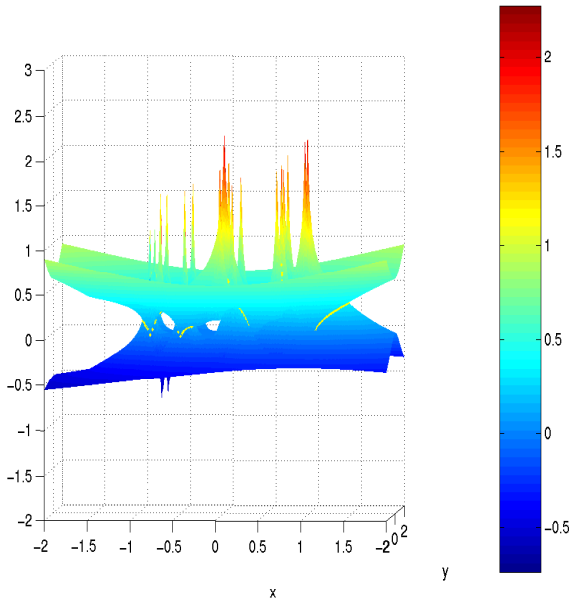


Figure 7.26: Spectral and frontier portraits intertwined for Example 7.2.2

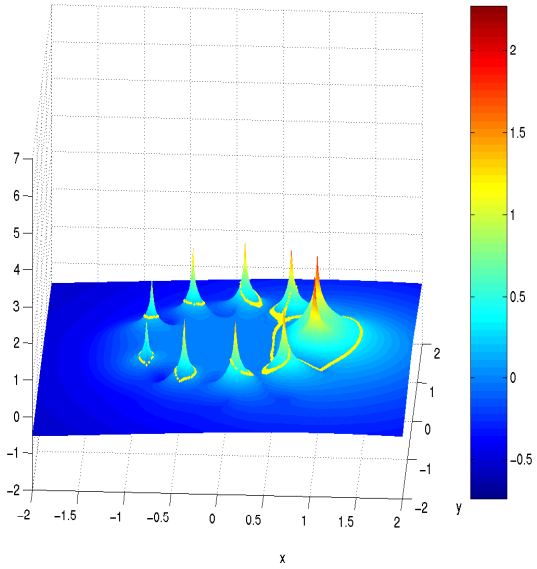


Figure 7.27: The spectral portrait with intersection curve $\mathcal{H}$

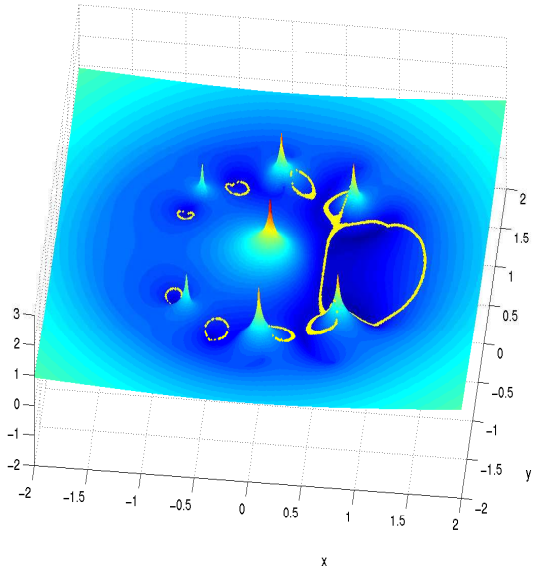


Figure 7.28: The frontier portrait with intersection curve $\mathcal{H}$

7.3 The case $r = \mathbf{rank}\, E = 2$

We study in detail the case $r = 2$ which can be treated explicitly.

7.3.1 E has rank 2

When E has rank 2, the matrices $Q(z)$ and M_z are of order 2 and their two eigenvalues are easily related to $\mathrm{tr}\, Q(z)$ and $\det Q(z)$: when $z \notin \sigma(A)$, we can use the characteristic polynomial

$$d^2(z) + \mathrm{tr}Q(z) + \det Q(z), \tag{7.3.1}$$

associated with $Q(z)$ to obtain

$$\mu_z^{\pm} = \frac{1}{2}\left[\frac{\mathrm{tr}Q(z)}{\pi(z)} \pm (\frac{\mathrm{tr}^2 Q(z)}{\pi(z)^2} - 4\frac{\hat{\pi}(z)}{\pi(z)})^{1/2}\right] = \frac{d^{\pm}(z)}{\pi(z)}$$
$$= \frac{1}{2}\frac{\mathrm{tr}Q(z)}{\pi(z)}\left[1 \pm \sqrt{1 - 4\frac{\det Q(z)}{\mathrm{tr}^2 Q(z)}}\right] = \frac{1}{2}\frac{\mathrm{tr}Q(z)}{\pi(z)}\left[1 \pm \sqrt{1 - a(z)}\right]$$

for $a(z) = 4\frac{\det Q(z)}{\mathrm{tr}^2 Q(z)}$. We assume that $\hat{\pi}(z) \not\equiv 0$.

Lemma 7.3.1 *[22] If* $\mathrm{tr}\, Q(\lambda) \neq 0$, *the frontier portrait has no well at* λ *which is exactly partially observable. If* $\lambda \in \hat{Z}$, *then* $\lambda \in \bar{\sigma}$ *and* $F(A, E)$ *can be extended into* $\mathbb{C}$ *at* λ.

Proof. When $\mathrm{tr}\, Q(z) \neq 0$, then

$$|\mu_z^+| = \left|\frac{1}{2}\frac{\mathrm{tr}Q(z)}{\pi(z)}\left[1 + \sqrt{1 - a(z)}\right]\right| \to \infty \quad \text{as} \quad a(z) \to a(\lambda) = 0.$$

On the other hand, we have

$$\mu_z^- = \frac{1}{2}\frac{\mathrm{tr}Q(z)}{\pi(z)}\left[1 - \sqrt{1 - a(z)}\right] \quad \to \quad \frac{\hat{\pi}(\lambda)}{\mathrm{tr}Q(\lambda)} \quad as \quad z \to \lambda$$

which is a finite limit.

If $\hat{\pi}(\lambda) = 0$, then $\lambda \in \hat{Z}$ and $\mu_\lambda^- = 0$: $\lambda \in \bar{\sigma}$.

□

Corollary 7.3.2 *[22] For* $\lambda \in \sigma(A)$ *such that* $\mathrm{tr}\, Q(\lambda) \neq 0$, *there is no well for* ϕ_2 *and* $\lambda \in \bar{\sigma}$ *if* $\lambda \in \hat{Z}$.

For a $\lambda \in \sigma(A)$ such that $\mathrm{tr}\, Q(\lambda) = 0$, it is possible that λ be $\sigma-$nonobservable, or $\sigma-$observable.

Lemma 7.3.3 *[22] When* $\hat{\pi}(z) \not\equiv 0$ *and* λ *is such that* $\mathrm{tr}\, Q(\lambda) = 0$, *there are* 2 *possibilities:*

i) if $\hat{m}_\lambda \geq m_\lambda \geq 1$, *then* $\lambda \in \hat{Z}$ *is* $\sigma-$ *nonobservable,*

ii) if $\hat{m}_\lambda < m_\lambda$, then λ is $\sigma-$ observable.

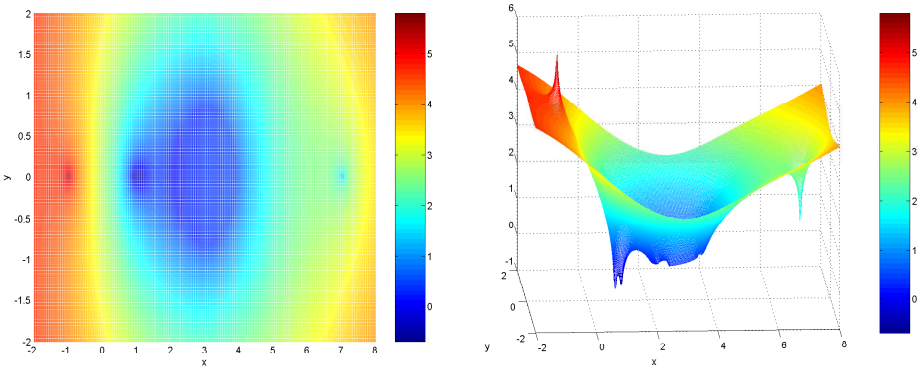
Corollary 7.3.4 *[22] When $\hat{m}_\lambda < m_\lambda$ then ϕ_2 has a well at λ. This happens for λ simple $\notin \hat{Z}$.*

We review the examples with $r = 2$ that have been proposed already.

1) In the Example 7.1.1, $\operatorname{tr} Q(-1) = 0$, and $\operatorname{tr} Q(\lambda) \neq 0$ for $\lambda \in \sigma(A)\backslash\{-1\}$. Therefore all eigenvalues $\lambda \in \sigma(A)\backslash\{-1\}$ are partially observable.

   $-1 \in \sigma(A)$ is such that $\hat{m}_{-1} = m_{-1} = 1$, therefore it is $\sigma-$ nonobservable. This was already proved. See Figures 7.8 and 7.28 for the spectral portrait and frontier portrait respectively.

2) In the Example 7.1.2, we have: $\operatorname{tr} Q(\lambda) \neq 0$ for all $\lambda \in \sigma(A)$. Therefore all eigenvalues are partially observable. See Figures 7.1 and 7.2. On Figures 7.29, (a) and (b), one can see 2D and 3D frontier portraits respectively. There are two sinks near $\lambda = 1$ and $\lambda = 7$. Also there is one peak only at the frontier point $z = -1 \in F(A, E) = \hat{Z} = \{-1\}$. Here $\hat{m}_\lambda = 0 < m_\lambda$ for any $\lambda \in \sigma(A)$.



(a) 2D− frontier portrait (b) 3D− frontier portrait

Figure 7.29: Frontier portrait $\phi_2 : z \mapsto \rho(M_z^{-1})$

3) Example 7.2.1, corresponds to the case $\hat{\pi}(z) \equiv 0$. Therefore $\det Q(z) \equiv 0$ and $Z(\det Q(z)) = \hat{Z} = \mathbb{C}$. $F(A, E) = re(A) \subset Z(\det Q(z))$ and $C(A, E) = \{0\}$.

4) In the Example 7.2.7, we have $\operatorname{tr} Q(\lambda) \neq 0$ for all $\lambda \in \sigma(A)$. Therefore all eigenvalues are partially observable.

We turn to two more examples with $E = UV^T$ of rank 2. A is the companion matrix for $\pi(z) = z^{11} + 1$ in upper Hessenberg form, $e = (1, \dots, 1)^T \in \mathbb{C}^{11}$.
In all the computer plots of the spectral field $t \mapsto \sigma(A(t))$, t is taken as the complex parameter $t = |t|e^{i\theta}$, where $h = |t| \in [0, 300]$ and θ is specified for each Figure. The eigenvalues of A appear as $\circledast$, the limit points in Lim appear as $\odot$.

7.3.2 (Σ) is satisfied: $G = V^TU$ has rank 2

Example 7.3.1 *$E = UV^T$ with $U = [e, f]$, $V = [e_{11}, e_3]$, $f = \sum_1^4 e_i$, so that $V^TU = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$. $0 \in \sigma(E)$ is semi-simple with $m = g = 9$. Then we have*

$$Q(z) = \left(\begin{array}{c|c} 1 + z + \cdots + z^{10} & 1 + z + z^2 + z^3 \\ \hline -(1 + z + \cdots + z^7) + z^8 + z^9 + z^{10} & -1 + z^8 + z^9 + z^{10} \end{array} \right),$$

$$\mathrm{tr}Q(z) = z(2z^9 + 2z^8 + 2z^7 + z^6 + \cdots + 1).$$

$$\det Q(z) = z(z^{11} + 1)(z^2 + z + 1)(z^6 + z^5 + \cdots + 1) \text{ has degree } 20,$$

$$\hat{\pi}(z) = z(z^2 + z + 1)(z^6 + z^5 + \cdots + 1) \text{ has degree } 9.$$

$F(A, E) = Zer(\hat{\pi}(z)) \cap re(A) = \hat{Z}$ consists of the 9 roots of $\hat{\pi}(z)$ which are 0 and the roots of 1 of order 3 and 7 complex ones, different from 1.

We know from theory that $\mathrm{Lim} = \sigma(\Pi)$ *where Π is a matrix of order 9. We conclude readily that* $\mathrm{Lim} = K(A, E) \subset re(A)$. *9 eigenvalues (out of 11) converge to the 9 kernel (or frontier) points.*

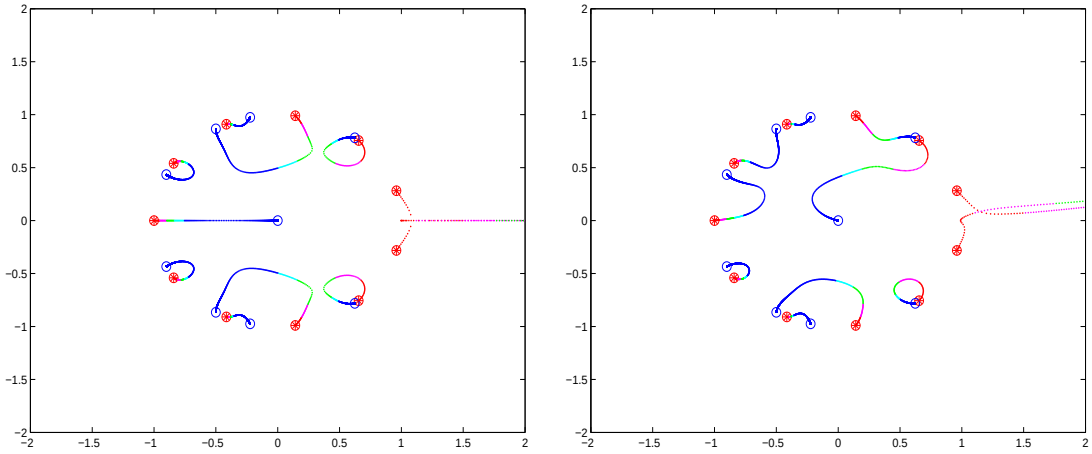
Moreover, $z = 0$ in $re(A)$ is a critical point since $\sigma_0 = \lim_{z \to 0} \sigma(M_z) = \{0, 0\}$, and $z^{1/2}$ is a factor of $\mu_z \in \sigma(M_z)$, $z \in re(A)$. At $z = 0$, $R(t, 0)$ is a polynomial in t of degree 2.

The spectral field is plotted for confirmation. See Figure 7.30, (a) ($\theta = 0$) and Figure 7.30, (b) ($\theta = \frac{\pi}{24}$) in which the escape of 2 eigenvalues to ∞ is visible. Two observations are in order:

1. *The slight change in θ from 0 to $\frac{\pi}{24} \sim 0.131$ provokes a drastic change in certain spectral maps $t \mapsto \lambda(t)$. For example, for t positive, there is a spectral ray connecting -1 to 0. For t slightly complex, the ray originating at -1 converges to a root of 1 of order 7, rather than to 0.*

2. *The ambiguities which arise from the perfect symmetry of Figure 7.30, (a) ($\theta = 0$) are easily resolved by making $\theta \neq 0$, as shown by Figure 7.30, (b).*

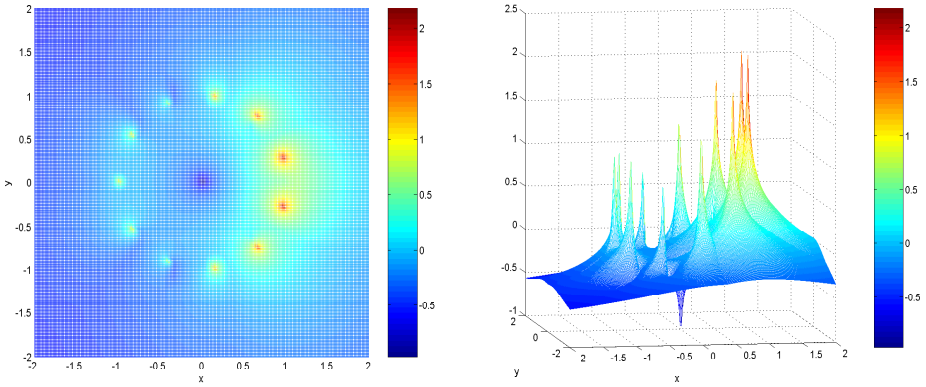
On Figures 7.31 and 7.32, one can see the spectral and frontier portraits respectively. Figures 7.31, (a) and (b) display 11 peaks corresponding to the eigenvalues of A



(a) $\theta = 0$ (b) $\theta = \frac{\pi}{24}$

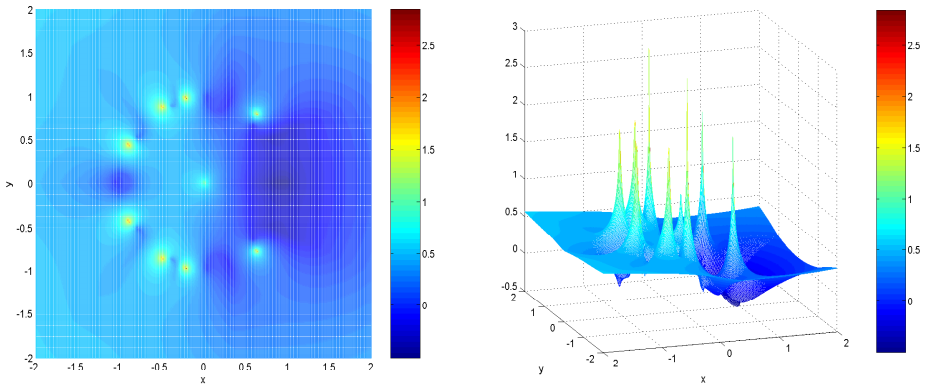
Figure 7.30: The map $t \mapsto \sigma(A(t))$ for $h = |t| \in [0,\ 300]$

and a well corresponding to the critical point $z = 0$. *Figures 7.32,* (a) *and* (b) *display* 9 *peaks corresponding to the* 9 *points in* $\mathrm{Lim} = F(A, E)$. *We observe that* $\mathrm{tr}\, Q(z) \neq 0$ *for* $\lambda \in \sigma(A)$ *: all eigenvalues are partially observable.*



(a) $2D$-spectral portrait (b) $3D$-spectral portrait

Figure 7.31: Spectral portrait $\phi_1 \ : \ z \mapsto \rho(M_z)$



(a) $2D$-frontier portrait (b) $3D$-frontier portrait

Figure 7.32: Frontier portrait $\phi_2 : z \mapsto \rho(M_z^{-1})$

△

7.3.3 $G = V^H U$ has rank 1

Example 7.3.2 *$E = UV^T$ with $U = [e, g]$, $V = [e_{11}, e_3]$, $g = \sum_1^4 e_i + e_{11}$, so that $V^T U = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$. $0 \in \sigma(E)$ is defective with 1 Jordan block of size 2.*

$$Q(z) = \left(\begin{array}{c|c} 1 + z + \cdots + z^{10} & 1 + z + z^2 + z^3 + z^{10} \\ \hline -(1 + z + \cdots + z^7) + z^8 + z^9 + z^{10} & -1 - z^7 + z^8 + z^9 + z^{10} \end{array} \right),$$

$$\mathrm{tr}Q(z) = z(z^2 + z + 1)(2z^7 + z^3 + 1)\,,$$

$$\det Q(z) = (z+1)z(z^{11} + 1)(z^2 - z + 1)(z^2 + z + 1)^2 \text{ is of degree } 19\,,$$

$$\hat{\pi}(z) = (z+1)z(z^2 - z + 1)(z^2 + z + 1)^2 \text{ has degree } 8\,.$$

We observe readily that, even though $(z+1)^2$ is a factor for $\det Q(z)$, $z+1$ is not a factor for $\mathrm{tr}Q(z)$. Hence, the factor $z+1$ does not cancel in the two eigenvalues $d(z)$ and in $\pi(z)$. However, one of the two eigenvalues of μ_z with minimum modulus is such that $\mu_{-1} = \lim_{z \to -1} \mu_z$ exists: $\lambda = -1$ is partially observable, but not $\sigma -$ observable. One of the 2 eigenvalues for $Q(z)$ is

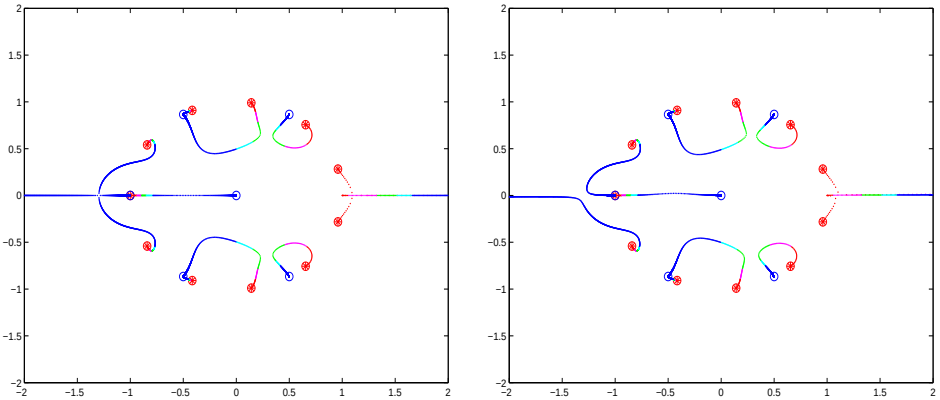
$$\begin{aligned} d(z) &= \frac{1}{2}\left[\mathrm{tr}Q(z) - \left[\mathrm{tr}^2 Q(z) - 4\,\mathrm{det}Q(z)\right]^{1/2}\right] \frac{\mathrm{tr}Q(z) + \left[\mathrm{tr}^2 Q(z) - 4\,\mathrm{det}Q(z)\right]^{1/2}}{\mathrm{tr}Q(z) + \left[\mathrm{tr}^2 Q(z) - 4\,\mathrm{det}Q(z)\right]^{1/2}} \\ &= -2\frac{\mathrm{det}Q(z)}{\mathrm{tr}Q(z) + \left[\mathrm{tr}^2 Q(z) - 4\,\mathrm{det}Q(z)\right]^{1/2}} \end{aligned}$$

For the corresponding μ_z, the factor $z+1$ in $\det Q(z)$ cancels with the one in $\pi(z)$. We conclude that $\lim_{z \to -1} \frac{d(z)}{z+1} = 0$: -1 is partially observable, corresponding

moreover to an eigenvalue $\mu_{-1} = 0 : -1 \in \sigma^f$. *It is interesting to see that* $\operatorname{tr} Q(-1) \neq 0$, *which confirms Lemma 7.3.1.*

$F(A,E) = Zer(\hat{\pi}(z)) \cap re(A)$ *contains the* 7 *roots of the polynomial* $z(z^2 - z + 1)(z^2 + z + 1)^2$. *By continuity,* $-1 \in \sigma(A)$ *can be added to* $F(A,E)$*:* $\bar{F}(A,E) = F(A,E) \cup \{-1\}$.

Because $z(z^2+z+1)$ *is a common factor for* $\operatorname{tr} Q(z)$ *and* $\det Q(z)$, *we conclude that* 0 *and* $\frac{1}{2}(-1 \pm i\sqrt{3})$ *are critical points in* $C(A,E)$. *Moreover,* $z^{1/2}(z^2 + z + 1)^{1/2}$ *is a factor for* $\mu_z \in \sigma(M_z)$, $z \in re(A)$.
We do not readily know if (G) *is satisfied for* $(0^2)^1(0^1)^8$. Lim *can be inferred from the computation of* $\sigma(A(t))$, *with* $t = |t|e^{i\theta}$, $h = |t| \in [0, 300]$.



(a) $\theta = 0$ (b) $\theta = 0.01$

Figure 7.33: The map $t \mapsto \sigma(A(t))$ for $h = |t| \in [0,\ 300]$

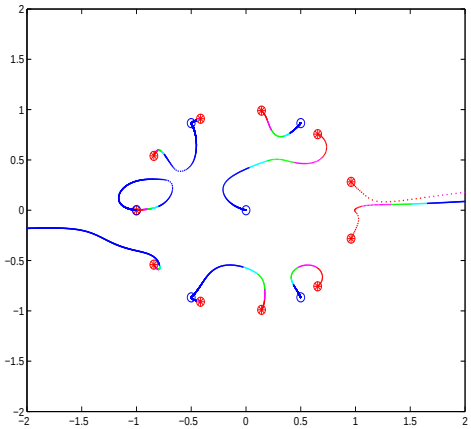


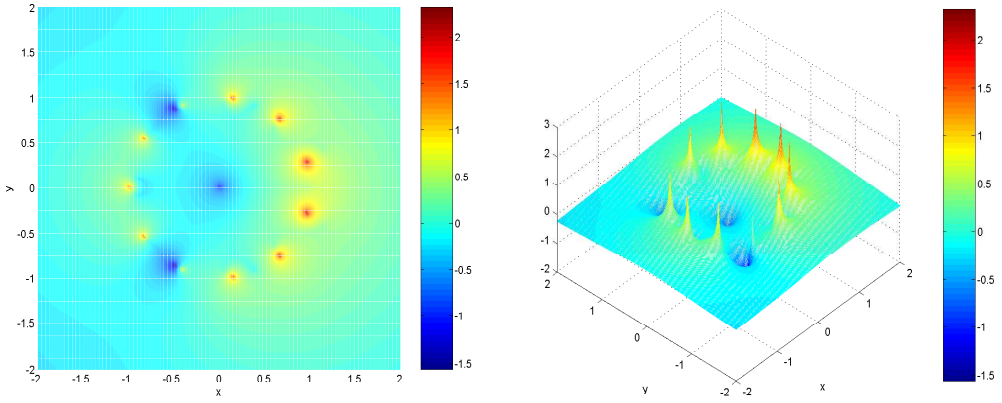
Figure 7.34: The map $t \mapsto \sigma(A(t))$ for $h = |t| \in [0,\ 300]$ and $\theta = \frac{\pi}{24}$

We observe on Figures 7.33, (a), (b) and Figure 7.34 that Lim *contains* 8 *points: the* 7 *frontier points plus the unobservable eigenvalue* $-1 \in \sigma^f$ *. This is a strong indication that* $(G) = (\mathrm{Li})$ *is satisfied* $(l_* = 8)$ *: there coexist* 7 *eigenvalues converging at rate* 1 *, and* 1 *converging at rate* > 1 *for* $E(s)$ *.*
This example has the following interesting features:

1. *The two limit points* $\frac{1}{2}(-1 \pm i\sqrt{3})$ *have multiplicity* 2 *: each of them is the limit of two different spectral rays originating from different eigenvalues.*

2. *Figure 7.34 exhibits a loop beginning and ending at* $\lambda = -1$ *for* $\theta = \pi/24$ *. The same phenomenon occurs for values of* $\theta \ (mod\ \pi)$ *ranging approximately between* 0.04π *and* 0.85π *. This is a rather* **rare event** *which is caused by the specific data* A, E *.*

On Figures 7.35 and 7.36, one can see the spectral portrait. Figures 7.35, (a) and (b) show the peaks at the eigenvalues of the matrix A *. On Figure 7.36, one can see* 3 *wells at* $C(A, E)$ *. On Figures 7.37, one can see the frontier portrait. Figures 7.37, (a) and (b) show* 8 *peaks at* $\mathrm{Lim} = F(A, E) \cup \{-1\} = \bar{F}(A, E)$ *.*



(a) $2D$-spectral portrait (b) $3D$-spectral portrait

Figure 7.35: Spectral portrait $\phi_1 \ : \ z \mapsto \rho(M_z)$

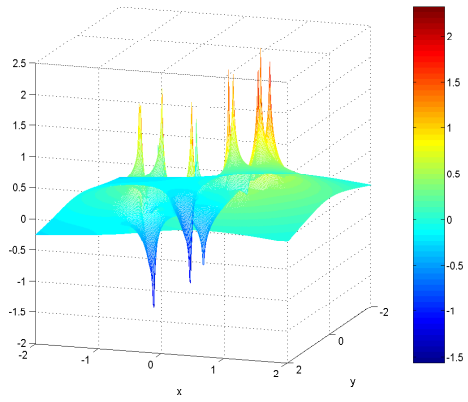
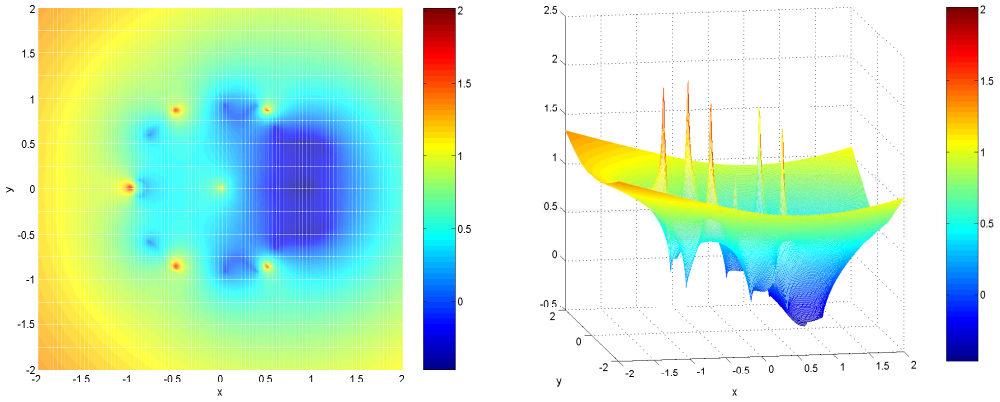


Figure 7.36: Spectral portrait $\phi_1 : z \mapsto \rho(M_z)$ from below



(a) $2D$-frontier portrait (b) $3D$-frontier portrait

Figure 7.37: Frontier portrait $\phi_2 : z \mapsto \rho(M_z^{-1})$

△

7.4 $\hat{\pi}(z) \equiv 0$

Under the condition (Δ), it is possible that $\hat{\pi}(z) \equiv 0$. In this case, $\hat{Z} = \mathbb{C}$, $F(A, E) = re(A)$. For the critical set which can be discrete or continuous, one has $C(A, E) \subseteq F(A, E) = re(A)$.

For any $z \in re(A)$, $\min_i |\mu_{iz}| = 0$. Therefore, by continuity, $\min_i |\mu_{i\lambda}| = 0$ for $|\mu_{i\lambda}| = \lim_{z \to \lambda} |\mu_{iz}|$. Hence $\bar{F}(A, E) = \mathbb{C}$.

However, it is possible that, if computed directly, $\min_i |\mu_{i\lambda}| \neq 0$. Indeed $\vartheta(z) = \frac{\hat{\pi}(z)}{\pi(z)}$ is $\equiv 0$ for $z \in re(A)$. But for $\lambda \in \sigma(A)$, $\vartheta(\lambda) = \frac{0}{0}$ is indeterminate,

it can be $\neq 0$ [22].

When $C(A,E)$ is discrete, then ϕ_1 has wells if $\bar{C}(A,E) \neq \emptyset$.

When $C(A,E)$ is continuous, then $C(A,E) = re(A)$, $\rho(M_z) = 0 \quad \forall\ z \in re(A)$. And $\sigma(A(t)) = \sigma(A) = \mathrm{Lim} = \sigma^i$. In this case, there is no intersection between the 2 planes defined by $\phi_1 \ : \ z \mapsto \rho(M_z) = 0$ and by $\phi_2 \ : \ z \mapsto \rho(M_z^{-1}) = \infty$.

7.5 A summary of the homotopic graphical toolkit and its use

We have defined two types of visualization tools:

1. in Chapter 6, we described the spectral rays and orbits derived from the spectral field $t \in \hat{\mathbb{C}} \mapsto \sigma(A(t)) \in \mathbb{C}^n$,

2. in Chapter 7, we described the three homotopic portraits associated with $z \in re(A) \mapsto M_z \in \mathbb{C}^{r\times r}$.

When $\hat{\pi}(z) \not\equiv 0$, these two different kinds of tools enable us to localize the points of interest in HD which are respectively:

i) the set Lim detected by the spectral rays and orbits,

ii) the $\|\cdot\|-$ observable eigenvalues (peaks of ϕ_0), the partially observable eigenvalues (peaks of ϕ_1), the $\sigma-$ observable eigenvalues (wells of ϕ_2), and the two sets $\bar{F}(A,E)$ (peaks of ϕ_2), and $\bar{C}(A,E)$ (wells of ϕ_1).

7.6 Computation of $\hat{Z}$ using MATLAB functions

It was mentioned in Chapter 3 that the *transfer matrix*

$$G(z) = C(zI - A)^{-1}B + D, \quad z \notin \sigma(A),$$

of the *state-space* equations

$$\left.\begin{array}{l} \dot{x}(t) = Ax(t) + Bu(t) \\ y(t) = Cx(t) + Du(t) \end{array}\right\}$$

in continuous-time system and discrete-time system in **Control Theory**, has the same structure as that of the matrix M_z in HD when $D = 0$, $C = V^H$, and $B = U$. In Matlab, the function *ss2tf* (resp. *tf2ss*) calculates the transfer matrix from state-space equations (resp. state-space equations from the transfer matrix). This means that, for quick MATLAB experimentation in HD when $\det Q(z) \not\equiv 0$, we can calculate the matrix polynomial $Q(z)$ by the function *ss2tf*. Then we can use

the function *polyeig* in Matlab to compute the roots of $\det Q(z)$ and discard $r-1$ copies of the roots of $\pi(z)$ to get the set $\hat{Z}$. However, this *indirect* computational route is not optimal. A more direct software approach, *specific* to HD, needs to be developed.

Chapter 8

Application of HD to Arnoldi's method

8.1 Introduction

Iterative processes are widely used in linear algebra for treating large sets of data. It is becoming more and more common that one iterative solver has to be embedded in an outer one: this is the case, for instance, for solving eigenproblems with inverse iterations or with a Krylov method with invert. Each step of the eigensolver requires the solution of a linear system which, if too large, must be solved in turn with an iterative method (inner step). It is a fact of *experience* that Krylov methods display an *extreme robustness* to large perturbations. No satisfactory explanation has been given so far of this remarkable property [13, 36, 50, 14, 24, 25].

In this chapter, we study the basic Arnoldi algorithm in the light of Homotopic Deviation theory in order to further our understanding of the nagging question : "why do Krylov methods work so well in practice ?".

8.2 Krylov subspace methods

Since their revival in the 70s and 80s, Krylov methods have been widely used worldwide to solve large scale problems such as

$$Ax = b \quad \text{and} \quad Ax = \lambda x,\ x \neq 0$$

which are the two basic problems associated with a large (often sparse) matrix A.

Their origin can be traced to C. Lanczos (hermitian and non hermitian tridiagonal form) and to W. E. Arnoldi (Hessenberg form). The notion of a subspace generated by powers of A applied to a vector is due to Krylov.

The Krylov subspace methods form an important class of methods available for computing eigenvalues and eigenvectors of large matrices. These techniques are

based on projection methods, both orthogonal and oblique, onto Krylov subspaces, i.e., subspaces spanned by the iterates of the simple power method [10]. What may appear to be a trivial extension of a very slow algorithm turns out to be one of the most successful methods for extracting eigenvalues of large matrices. A subspace of the form

$$\mathcal{K}^k(A, u_0) = \text{span}\{u_0, Au_0, \ldots, A^{k-1}u_0\},$$

is referred to as a Krylov subspace with starting vector u_0.

In such methods, the dimension of the approximation subspace is increased by one at each step of the approximation process. To keep the dimension manageable, one can attempt to force the starting vector u_0 to be more in the direction of the desired eigenvector. Alternatively, one can start with a set (block) of vectors instead of a single vector u_0 and this leads to the so-called block variants of these subspace methods. This gives rise to various iterative methods collectively referred to as Krylov methods. They include:

1. The symmetric Lanczos algorithm,
2. Arnoldi's method and its variations,
3. The nonsymmetric Lanczos algorithm.

In this chapter, we concentrate on the Arnoldi method. This method [9, 34, 49] is an orthogonal projection method onto the Krylov subspace $\mathcal{K}^k(A, u_0)$ of dimension $k \leq n$ for general matrices. The procedure was introduced in 1951 as a means of reducing a dense matrix into Hessenberg form. Arnoldi presented his method in this manner but hinted that the eigenvalues of the Hessenberg matrix obtained from a number k of steps smaller than n could provide accurate approximations to some eigenvalues of the original matrix. It was later discovered by Y. Saad that this strategy leads to an efficient technique for approximating eigenvalues of large sparse matrices.

8.2.1 The basic Arnoldi algorithm for the Hessenberg decomposition

The Arnoldi method is an orthogonal projection method onto a Krylov subspace. It starts with the Arnoldi procedure as described in Algorithm 1. The procedure can be understood as a *modified Gram-Schmidt* process for building an orthogonal basis of the Krylov subspace $\mathcal{K}^k(A, v)$.

The algorithm 1 will stop if the vector w computed in line (8) vanishes. The vectors $v_1, v_2, \ldots, v_k$ form an orthonormal system by construction and are called *Arnoldi*

Algorithm 1 Arnoldi Procedure

1: $v_1 = v/\|v\|_2$ for the starting vector v
2: **for** $j = 1, 2, \ldots, k$ **do**
3: $w := Av_j$
4: **for** $i = 1, 2, \ldots, j$ **do**
5: $h_{ij} = w^* v_i$
6: $w := w - h_{ij} v_i$
7: **end for**
8: $h_{j+1,j} = \|w\|_2$
9: **if** $h_{j+1,j} = 0$, **stop**
10: $v_{j+1} = w/h_{j+1,j}$
11: **end for**

vectors. It is shown that this system is a basis of the Krylov subspace $\mathcal{K}^k(A, v)$ [49, 10].

Now, we consider a fundamental relation between quantities generated by the algorithm 1. The following equality is readily derived:

$$Av_j = \sum_{i=1}^{j+1} h_{ij} v_i, \quad j = 1, 2, \ldots, k. \tag{8.2.1}$$

If we denote by V_k the $n \times k$ matrix with column vectors $v_1, v_2, \ldots, v_k$ and by H_k the Hessenberg matrix whose nonzero entries h_{ij} are defined by the algorithm, then the following relations hold:

$$AV_k = V_k H_k + h_{k+1,k} v_{k+1} e_k^H, \tag{8.2.2}$$

$$V_k^H A V_k = H_k. \tag{8.2.3}$$

Relation (8.2.3) follows from (8.2.2) by multiplying both sides of (8.2.2) by V_k^H and making use of the orthonormality of $\{v_1, v_2, \ldots, v_k\}$.

As was noted earlier the algorithm breaks down when the norm of w computed on line (8) vanishes at a certain step j. This happens if and only if the starting vector v is a combination of eigenvectors (i.e., the minimal polynomial of v_1 is of degree j). In addition, the subspace $\mathcal{K}_j$ is then invariant and the approximate eigenvalues and eigenvectors (if computed in exact arithmetic) are exact [34, 49].

The approximate eigenvalues $\lambda_i^{(k)}$ provided by the projection process onto $\mathcal{K}^k$ are the eigenvalues of the Hessenberg matrix H_k of order k. These are known as *Ritz values*. A *Ritz approximate eigenvector* associated with a Ritz value $\lambda_i^{(k)}$ is defined by $u_i^{(k)} = V_k y_i^{(k)}$, where $y_i^{(k)}$ is an eigenvector associated with the eigenvalue $\lambda_i^{(k)}$. A number of the Ritz eigenvalues, typically a small fraction of k, may constitute a good approximations for certain eigenvalues λ_i of A, and the quality of the approximation improves as k increases.

The original algorithm consists of increasing k until all desired eigenvalues of A are found. In theory, k should be increased until the value n. For large matrices, this becomes costly both in terms of computation and storage [10].

8.2.2 Implementation variants

The description of the Arnoldi procedure given earlier was based on the modified Gram-Schmidt process. Other orthogonalization algorithms could be used as well. One option is to reorthogonalize only when necessary. Whenever the final vector obtained at the end of the second loop in the above algorithm has been computed, a test is performed to compare its norm with the norm of the initial w (which is $\|Av_j\|_2$). If the reduction falls below a certain threshold (an indication that severe cancellation might have occurred), a second orthogonalization is made. It is known from a result by Kahan that more than two orthogonalizations are superfluous [47].

One of the most reliable orthogonalization techniques, from the numerical point of view, is the Householder algorithm [38]. This has been implemented for the Arnoldi procedure by Walker [53]. The Householder algorithm is numerically more reliable than the Gram-Schmidt or modified Gram-Schmidt versions, but it is also more expensive, requiring roughly the same storage as modified Gram-Schmidt but about twice as many operations. The Householder orthogonalization is a reasonable choice when developing general purpose, reliable software packages where robustness is a critical criterion.

8.2.3 Explicit Restarts

As was mentioned earlier, the standard implementations of the Arnoldi method are limited by their high storage and computational requirements as k increases. We suppose that we are interested in only one eigenvalue/eigenvector of A, namely, the eigenvalue of largest real part of A. Then one way to circumvent the difficulty is to *restart* the algorithm. After a run with k Arnoldi vectors, we compute the approximate eigenvector and use it as an initial vector for the next run with the Arnoldi method. This process, which is the simplest of this kind, is iterated to convergence to compute one eigenpair, see Algorithm 2.

Obtaining the residual norm, for Ritz pair, as the algorithm progresses is fairly inexpensive. Let $y_i^{(k)}$ be an eigenvector of H_k associated with the eigenvalue $\lambda_i^{(k)}$, and let $u_i^{(k)}$ be the Ritz approximate eigenvector $u_i^{(k)} = V_k y_i^{(k)}$. We have the relation

$$(A - \lambda_i^{(k)} I) u_i^{(k)} = h_{k+1,k} (e_k^T y_i^{(k)}) v_{k+1},$$

and therefore

$$\|(A - \lambda_i^{(k)} I) u_i^{(k)}\|_2 = h_{k+1,k} |e_k^T y_i^{(k)}|.$$

Thus the residual norm is equal to the absolute value of the last component of the eigenvector $y_i^{(k)}$ multiplied by $h_{k+1,k}$. The residual norms are not always indicative

of actual errors in $\lambda_i^{(k)}$, but can be helpful options in deriving stopping criterion especially under their relative form.

For computing other eigenpairs, and for improving the efficiency of the process, a number of strategies have been developed, which include deflation procedures and the implicit restarting strategy [10, 44].

Algorithm 2 Explicity Restarted Arnoldi Method

1: **Iterate:** Perform k steps of Algorithm 1.
2: **Restart:** Compute the approximate eigenvector $u_1^{(k)}$ associated with the rightmost eigenvalue $\lambda_1^{(k)}$.
3: If stopping criterion holds stop, else set $v_1 \equiv u_1^{(k)}$ and goto 1.

8.3 The incomplete Arnoldi decomposition

8.3.1 Definition

Let $A \in \mathbb{C}^{n\times n}$. For $1 < k < n$, the incomplete (or inexact) Arnoldi decomposition for A can be written as

$$(A - hU)V = VH_k,$$

where $V = [v_1, \ldots, v_k] \in \mathbb{C}^{n\times k}$ is the Krylov basis at step k, $H_k \in \mathbb{C}^{k\times k}$ is in Hessenberg form, $U = v_{k+1}v_k^H$ is nilpotent: $v_k^H v_{k+1} = 0$ implies $U^2 = 0$. And $h = h_{k+1\ k}$ denotes the $(k+1, k)$ element in H_{k+1}. We remark that h is real positive with a (classical or modified) Gram-Schmidt orthogonalisation strategy. With a Householder strategy, h can be complex. When $k = n$, the exact decomposition ($h = 0$) is completed with $AV = VH_n$, $V \in \mathbb{C}^{n\times n}$, where $H_n = V^H AV$ is one possible Hessenberg form for A.

8.3.2 Relation between $\sigma(A)$ and $\sigma(H_k)$, $1 < k < n$

The interpretation of the incomplete Arnoldi decomposition has been studied by E. Traviesas (1999) in the HD framework. It is shown in [26, 28] and [51], chapter 2, p. $41 - 42$ that

1. $\sigma(H_k) \subset \sigma(A - hU)$, $1 < k < n$,
2. $|h|$ represents the homotopic distance between $\sigma(A)$ and $\sigma(H_k)$,
3. the n eigenvalues of $A - hU$ lie on the homotopic $\frac{1}{|h|}$-level curve Γ defined by :

$$\Gamma = \{z \in \mathbb{C} - \sigma(A);\ \rho(U(A - zI)^{-1}) = \frac{1}{|h|}\}.$$

As a consequence, $|h|$ measures the backward error for the incomplete Arnoldi decomposition, which consists in replacing A by H_k [33].

8.3.3 Spectral consequences of $A = VH_nV^H$ ($k = n$)

In this paragraph, we summarize results given in [15, 28, 51]. Irreducible Hessenberg matrices are nonderogatory (one Jordan block per distinct eigenvalue) [34]. Therefore the Arnoldi decomposition on a derogatory matrix A yields a reduced Hessenberg matrix H_n (in exact arithmetic). The algorithmic assumption that H_n is irreducible implies that:

i) if H_n is simple, then A is simple with n distinct eigenvalues. Multiple eigenvalues escape.

ii) if H_n is defective, then A is nonderogatory: multiple Jordan blocks for the same eigenvalue are out of reach.

However, these limitations are only valid in *exact* arithmetic. They are bypassed by finite precision as experience tells us. In algorithmic practice, the goal is to force near reducibility to occur as soon as possible (that is for k small with respect to n). Understanding the process of numerical (or happy) breakdown (the finite precision counterpart of exact reducibility) is therefore crucial to the design of efficient stopping criteria [15, 28].

8.3.4 The Arnoldi residual, $1 < k < n$

Let (λ, y) denote an eigenpair for $H = H_k$, $1 < k < n$ and $h = h_{k+1\ k}$. This yields

$$Hy = \lambda y,\ y \in \mathbb{C}^k,$$

and

$$(A - \lambda I)Vy = V(H - \lambda I)y + hUVy, \quad Vy \in \mathbb{C}^n.$$

The pair (λ, Vy) is a pseudoeigenpair for A corresponding to the residual

$$r = (A - \lambda I)Vy = hUVy = hv_{k+1}(e_k^H y). \tag{8.3.1}$$

We set $y_k = e_k^H y$, the kth (and last) component for y. Equality (8.3.1) is valid in exact arithmetic, but not always in finite precision. Therefore we introduce the following distinction:

$$r_D = (A - \lambda I)Vy \text{ is the } \textit{direct residual} \text{ for } A,$$

$$r_A = (hy_k)v_{k+1} \text{ is the } \textit{Arnoldi residual}.$$

The backward error for A at (λ, Vy) is

$$BE(\lambda, Vy) = \frac{\|r_D\|}{\|A\|\|y\|} = \frac{|hy_k|}{\|A\|\|y\|}. \tag{8.3.2}$$

It is a fact of experience [23, 26], that the mathematical equality (8.3.2), which derives from (8.3.1), may not hold for iterations which follow the one for which the backward error reaches machine precision. Therefore the Arnoldi residual, of norm $|hy_k|$, is not a reliable indicator for *convergence*, once machine precision is reached for $\frac{|hy_k|}{\|A\|\|y\|}$.

The fact that numerically subtle events take place in the vicinity of convergence has been known for more than 2 decades [47, 23]. One remarks that the residual in 8.3.1 has an absolute formulation, suitable for a mathematical forward analysis of convergence. By comparison, the normalised residual in 8.3.2 is suited for a backward analysis of algorithmic convergence in finite precision.

In next section, we present these observations in HD framework (introduced in [17]) to take a fresh look at the phenomenon of near reducibility.

8.4 Spectral structure of an irreducible Hessenberg matrix

8.4.1 An inductive analysis for $1 < k < n$

$H \in \mathbb{C}^{k\times k}$ is a Hessenberg matrix of order k assumed to be irreducible. We consider the Hessenberg matrix of order $k+1$ defined by

$$H^+ = \left[\begin{array}{c|c} H & u \\ \hline 0\ h & a \end{array}\right]$$

where $u \in \mathbb{C}^k$, h and $a \in \mathbb{C}$. What can be said about the eigenstructure of H^+? We consider the eigenpair (μ, x) for H^+, with $x = (y^H, \alpha)^H \in \mathbb{C}^{k+1}$. $H^+x = \mu x$ implies

$$Hy + \alpha u = \mu y,$$
$$hy_k + a\alpha = \mu\alpha.$$

By assumption, H^+ is irreducible for $h \neq 0$, and x is the only eigendirection associated with μ. The discussion of whether $\alpha = 0$ or not in [17] shows that,

1. For $\alpha = 0$, the only possibility is that $h = 0$: H^+ is reducible.
2. For $\alpha = 1$, one has

$$(H - \mu I_k)y = -u,$$
$$hy_k = \mu - a.$$

When nonzero, the quantity hy_k measures the forward error for H^+ on the eigenvalue $\mu \in \sigma(H+)$ when it is approximated by a.

When $h \neq 0$, $hy_k = 0$ implies $y_k = 0$. If the matrix A is assumed nonderogatory, then H_k is irreducible for $k = 2, \ldots, n-1$. Therefore H^+ $(= H_{k+1})$ is irreducible and we only have to consider the case $\alpha = 1$. We conclude that when (μ, x) is an exact eigenpair for H^+, then

(i) $x = (y^H, 1)^H$, where y corresponds to one step of inverse iteration on H with $\mu \in \sigma(H^+)$ as an approximate eigenvalue for H, and $-u$ being the residual $(H - \mu I_k)y$, that is, $y = -(H - \mu I_k)^{-1}u$,

(ii) hy_k equals the forward error $\mu - a$. It can be zero either for $y_k = x_k = e_k^T x = 0$ or for $h = 0$,

(iii) $|h_{k+2k+1}| = \|(A - \mu I)Vx\|$ is the Arnoldi residual at step $k+1$, since $x_{k+1} = e_{k+1}^T x = 1$.

These observations show how the quantities h, a and u which represent new information about A not present in H are processed algorithmically to update the eigendecomposition of H into that of H^+, during the Arnoldi decomposition at step $k+1$.

8.4.2 h as a homotopy parameter

Consider the matrix

$$B = \left[\begin{array}{c|c} H & u \\ \hline 0 & a \end{array}\right]$$

of order $k+1$, with spectrum $\sigma(B) = \sigma(H) \cup \{a\}$, and $1 < k < n$. The matrix

$$H^+ = \left[\begin{array}{c|c} H & u \\ \hline 0\ h & a \end{array}\right]$$

can be written as $H^+ = B + hE = B(h)$ with $E = e_{k+1}e_k^T$. Note that $E^2 = 0$ because $e_k^T e_{k+1} = 0$: E is *nilpotent*. To study the dependence of the spectrum of H^+ on the parameter h, the framework of Homotopic Deviation (B, E), where E is rank one and nilpotent, is most natural [16, 27, 51]. The homotopy parameter h will be considered complex. Of particular interest will be the limits of the eigenvalues $\lambda(h)$ as $|h| \to \infty$.

8.5 $E = uv^H$, and $v^H u = 0$, $u, v \in \mathbb{C}^n$

8.5.1 $E = e_n e_{n-1}^T$

For $u, v \in \mathbb{C}^n$ such that $v^H u = 0$, consider the unitary basis

$$Q = \left[X, \frac{v}{\|v\|}, \frac{u}{\|u\|}\right] \text{ in } \mathbb{C}^n,$$

with $X^H X = I_{n-2}$, $v^H X = u^H X = 0$. Set $E = uv^H$ with $v^H u = 0$. The matrix $A(t) = A + tE$ is unitarily equivalent to $C(t) = C + tD$, with

$$D = \|u\|\|v\| e_n e_{n-1}^T = Q^H E Q.$$

We can therefore, without loss of generality in this section, restrict out attention to the deviation $E = e_n e_{n-1}^T$. From now on, in this Section, E is assumed to be $e_n e_{n-1}^T$. E is nilpotent, with only eigenvalue 0 with multiplicity n and structure $(0^1)^{n-2}(0^2)$. The Jordan chain associated with 0 double defective is (e_n, e_{n-1}): $Ee_n = 0$ and $Ee_{n-1} = e_n$.

8.5.2 The four sets of interest for (A, E), $E = e_n e_{n-1}^T$

Let P' be the orthogonal projection on $W_{n-2} = lin(e_1, \ldots, e_{n-2})$ which represents the eigenspace for E associated with 0 of ascent 1, multiplicity $n-2$.

$A_{n-2} = P' A P'$ represents the section (principal submatrix) of A of order $n-2$. We define the partitioning $(n-2, 2)$ of A as

$$\left[\begin{array}{c|c} A_{n-2} & R \\ \hline S & A_2 \end{array}\right]$$

with $R, S^H \in \mathbb{C}^{(n-2)\times 2}$ and A_2 of order 2. We assume that $\sigma(A) \cap \sigma(A_{n-2}) = \emptyset$. And we consider the family $A(t) = A + tE$, $t \in \mathbb{C}$.

The overview presented above tells us that four sets in $\mathbb{C}$ are useful to study the homotopic deviation process (A, E), where $r = 1$ and $E^2 = 0$ so that $g = n-1$, $g' = n-2$. These are

i) the set $\sigma(A)$ of n eigenvalues for A,

ii) the set Lim of limit points for $\sigma(A(t))$ which remain at finite distance when $|t| \to \infty$ such that $\sigma_\infty(A, E) = lim_{|t|\to\infty} \sigma(A(t)) = \{\infty, \mathrm{Lim}\}$,

iii) the set $C(A, E)$ of critical points z which are such that $\mu_z = 0$, hence $F_z^2 = 0$ (because $r = 1$, $C(A, E) = F(A, E)$),

iv) The set $\sigma(A_{n-2})$. The general theory [18] applied to E nilpotent with $r = 1$ entails that generically $\mathrm{Lim} \cap re(A) = C(A, E) = F(A, E)$ contains at most $n-2$ critical points. Nongenerically, it is possible that $C(A, E)$ is the continuous set $re(A)$, and that $\sigma(A) = \sigma_\infty(A, E)$ is invariant under $t \in \mathbb{C}$. We confirm some of these results by direct proof.

Proposition 8.5.1 *[17, 22] Any point in* Lim *which is not an eigenvalue of* A *is critical. If it is an eigenvalue of* A*, it is invariant.*

Proposition 8.5.2 *[17] If* A_2 *is not a lower triangle, exactly* 2 *eigenvalues of* $A(t)$ *escape to* ∞. *The remaining* $n-2$ *converge to* $\mathrm{Lim} = \sigma(\Omega)$, *where*

$$\Omega = A_{n-2} - \frac{1}{a_{n-1n}} uv^T,$$

with $u = (a_{1n}, \ldots, a_{n-2n})^T$ *and* $v^T = (a_{n-11}, \ldots, a_{n-1n-2})$.

8.5.3 The structure of F_z, for $z \notin \sigma(A)$

Because $-F_z = e_n e_{n-1}^T (A - zI)^{-1}$, e_n is the eigenvector associated with $\mu_z = -e_{n-1}^T (A - zI)^{-1} e_n$, which is nonzero when z is not critical. In the generic case, F_z is semi-simple: it has n independent eigenvectors.

When $z = \xi \notin \sigma(A)$ is critical, however, the structure of F_z changes from semi-simple ($z \neq \xi$) to defective and nilpotent: $F_\xi^2 = 0$: the eigenvector e_n is linked with another vector α by the Jordan chain of length 2: $F_\xi \alpha = e_n$. This creates, at the critical points, a computational dependency which is not present at a generic $z \notin \sigma(A)$. This dependency is easy to explicit in the case corresponding to the incomplete Arnoldi decomposition described in Section 8.4.2. This is the subject of Section 8.6.

8.6 Application of HD to the Arnoldi method

8.6.1 Three successive Hessenberg matrices constructed by the Arnoldi decomposition

In this section, we look at the convergence of Arnoldi algorithm from the point of view of Homotopic Deviation theory for the matrix of order $k+1$ ($3 \leq k+1 \leq n$),

$$B = \left[\begin{array}{c|c} H & u \\ \hline 0 & a \end{array}\right]$$

where H is assumed to be an *irreducible Hessenberg* matrix. In this case,

$$(B - zI_{k+1})^{-1} = \left[\begin{array}{c|c} (H - zI_k)^{-1} & w_z \\ \hline 0 & (a-z)^{-1} \end{array}\right]$$

with

$$w_z = -\frac{1}{a-z}(H - zI_k)^{-1} u \in \mathbb{C}^k.$$

We assume that $a \notin \sigma(H)$, that is B is nonderogatory. The matrix $H_{k+1} = H^+ = \left[\begin{array}{c|c} H_k & u \\ \hline 0\ h_{k+1\ k} & a \end{array}\right]$ for $H_k = H$ is the computed Hessenberg form of order $k+1$. The homotopy parameter is $h = h_{k+1\ k}$, and the deviation matrix

is $E = e_{k+1}e_k^T$: $B(h) = B + hE = H_{k+1}$. E is nilpotent ($E^2 = 0$) with rank $r = 1$, and $\sigma(E) = \{(0^1)^{k-1}, (0^2)\}$. This means that $0 \in \sigma(E)$ is defective and using the notation of chapter 3, we have $q = 2$, $n_q = 2$ with $r_q = 1$ where the multiplicities, m, g, and g' of $0 \in \sigma(E)$ satisfy $g' = k - 1 < g = k < m = k + 1$. Therefore $g' \geq 1$ for $k \geq 2$ and $(G) = (\mathrm{Li})$.

For k fixed, $1 < k < n$, we set $H^- = H_{k-1}$, $H = H_k$, $H^+ = H_{k+1}$: these are the three successive Hessenberg matrices constructed by the Arnoldi decomposition, of order $k-1$, k and $k+1$ respectively. We define $h^- = h_{k\ k-1}$ and $u = (\tilde{u}^H, u_k)^H$ where $\tilde{u} \in \mathbb{C}^{k-1}$ consists of the first $k-1$ entries of the vector u and $u_k = e_k^T u$ is the kth entry of u.

Since $H_k = H$ is irreducible, therefore $\sigma(H^-) \cap \sigma(H) = \emptyset$ and $h_{k\ k-1} \neq 0$ in particular. Therefore $\sigma(B) = \sigma(H) \cup \{a\}$. The eigenspace K' is $K' = \mathrm{lin}(e_1, \ldots, e_{k-1})$, and P' is the orthogonal projection on K'. Thus $\Pi' = H_{k-1} = H^-$, and

$$\Omega = H^- - \frac{h_{k\ k-1}}{u_k}\tilde{u}e_{k-1}^T = H^- - L\Gamma^{-1}R, \tag{8.6.1}$$

of order $k-1$ is defined for $u_k \neq 0$ where $L = \tilde{u}$, $\Gamma = (u_k)$ and $R = [0\ \ldots\ 0\ h^-]$. The matrix M_z reduces to the scalar $\mu_z = -e_k^T(B - zI_{k+1})^{-1}e_{k+1}$, for $z \notin \sigma(B)$.

8.6.2 The sets of interest for (B, E)

The section of B of order $(k+1)-2 = k-1$ is given by the irreducible Hessenberg matrix H^-. By assumption $h^- = h_{kk-1} \neq 0$ then $\sigma(H^-) \cap \sigma(H) = \emptyset$. Furthermore we assume that $\sigma(B) \cap \sigma(H^-) = \emptyset$, that is $a \notin \sigma(H^-)$. The four sets of paragraph 8.5.2 become respectively:

1. $\sigma(B) = \sigma(H) \cup \{a\}$, the spectrum of B,
2. $\sigma_\infty = \sigma_\infty(B, E) = \{\infty, \mathrm{Lim}\}$,
3. $C(B, E)$, the set of critical points,
4. $\sigma^- = \sigma(H^-)$.

We write $\sigma^+ = \sigma(H^+)$ for the spectrum of $H^+ = B + hE$, for any $h \neq 0$, $h \in C$. We set $\beta_z = (B - zI_{k+1})^{-1}e_{k+1}$ which represents the last column of $(B - zI_{k+1})^{-1}$. $\mu_z = -e_k^H\beta_z$ represents the only possibly non zero eigenvalue of $F_z = -e_{k+1}e_k^H(B - zI_{k+1})^{-1}$. Finally, because $r = 1$, $C(B, E) = F(B, E)$ in $re(B)$.

Proposition 8.6.1 *[17] If $z \in \sigma(H^+) = \sigma^+$, the vector β_z is an eigendirection for H^+ associated with the eigenvalue z, normalised such that $-he_k^H\beta_z = h\mu_z = 1$.*

We know from the general theory that $z \in re(B)$ is an eigenvalue of H^+ iff $h\mu_z = 1$. Proposition 8.5.1 applies: if $z \in \mathrm{Lim} \cap re(B)$, then $\mu_z = 0$, that is z is critical: it belongs to $\mathrm{Lim} \cap re(B) = C(B, E)$.

As a corollary of Proposition 8.5.2 for the generic case $u_k = h_{k,k+1} \neq 0$, we get the

Theorem 8.6.2 *[17] If $u_k = e_k^T B e_{k+1} \neq 0$, exactly 2 eigenvalues $\lambda(h)$ escape to ∞. The $k-1$ others converge to $\mathrm{Lim} = \sigma(\Omega)$, with*

$$\Omega = H^- - \frac{h^-}{u_k} \tilde{u} e_{k-1}^T,$$

such that

$$\|H^- - \Omega\| = |\frac{h^-}{u_k}| \ \|\tilde{u}\|.$$

We observe that for $u \neq 0$, $\frac{\|\tilde{u}\|}{|u_k|} = tan\Psi$, where is the acute angle between the directions spanned by $\tilde{u}$ and e_k. The condition $u_k \neq 0$ is equivalent to $0 \leq \Psi < \pi/2$, and $\|\tilde{u}\| = 0 \Longleftrightarrow \Psi = 0 \Longleftrightarrow \Omega = H^-$, since $h^- \neq 0$.

The computational significance of Theorem 8.6.2 should not be underestimated. It shows that, from an algorithmic point of view, the spectral information about A given by H^- when $|h|$ is large, can be as meaningful as the information given by H^+ for $|h|$ small. The robustness of the Arnoldi decomposition to large deviations stems from this powerful dual point of view.

Example 8.6.1 *[17] This example illustrates the Homotopic Deviation $H^+ = B(h) = B + hE$ of order $k+1 = 9$. The matrix $H - I$ of order $k = 8$ is taken to be Venice, the companion matrix defined in [27]. Its characteristic polynomial is $(x-1)^3(x-3)^4(x-7)$. Therefore the spectrum $\sigma(H) = \{2, 4, 8\}$ has the structure $(2^3)(4^4)(8^1)$.*

Let $E = e_9 e_8^T$, then B of order 9 is obtained by bordering H with $a = 9$, $u = (1, 2, 3, \ldots, 7, 8)^T$. The spectrum of B is that of H plus 9. The projection P is on $lin(e_1, \ldots, e_7)$, and

$$\Pi = PBP_{\restriction \mathrm{Ker}E} = H^- = \begin{bmatrix} 1 & & & 0 \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ 0 & & 1 & 1 \end{bmatrix},$$

a transposed Jordan block of order 7.

For $h \in \mathbb{C}$, the 9 maps $h \to \lambda_i(h) \in \sigma(H^+)$ $(i = 1, \ldots, 9)$ represent the spectral rays for the spectral field associated with (B, E). Figure 8.1 display the 9 spectral rays (computed by QR) for $\theta = 0$, $|t| = h \in [0, \ 7 \times 10^6] \subset \mathbb{R}^+$. The eigenvalues

of B corresponding to $h = 0$ are denoted by red $\circledast$ and the the elements of Lim$= \sigma(\Omega)$, when exist, are shown by $\odot$.

According to the Theorem 8.6.2, it is expected that exactly 2 spectral rays escape to infinity ($\pm\infty$) because $u_8 = 8 \neq 0$. Indeed, this the case with two eigenvalues diverging: the ray originating at 8 or 4 (resp. 9) escape to $-\infty$ (resp. $+\infty$). The remaining 7 rays converge to Lim$= \sigma(\Omega)$ with $\Omega = H^- - \frac{1}{8}\tilde{u}e_7^T$, for $\tilde{u} = [1, 2, \ldots, 7]^T$.

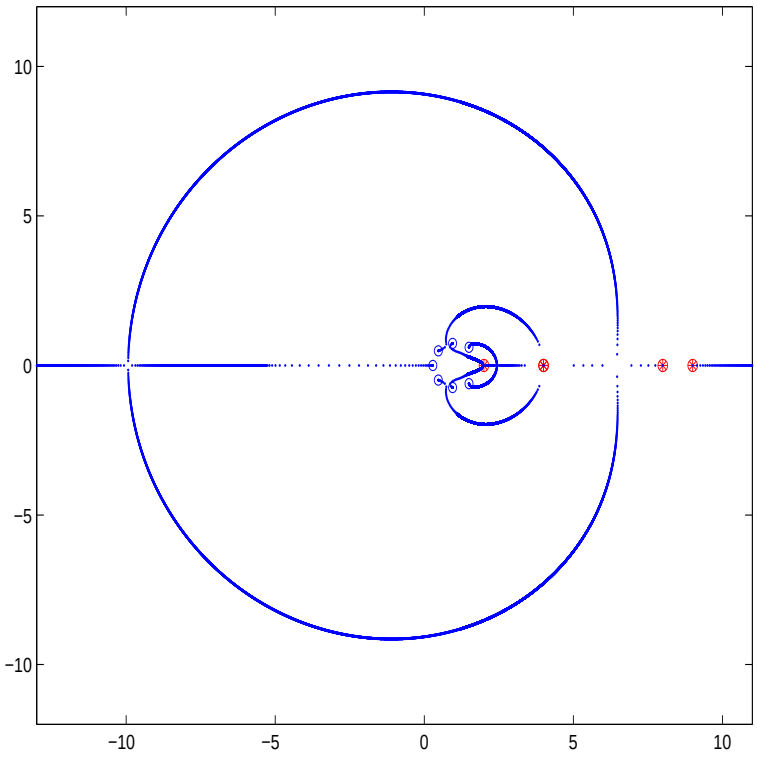


Figure 8.1: $0 \leq h \leq 7 \times 10^6$, $h \in \mathbb{R}^+$, and $e_8^T u = 8 \neq 0$

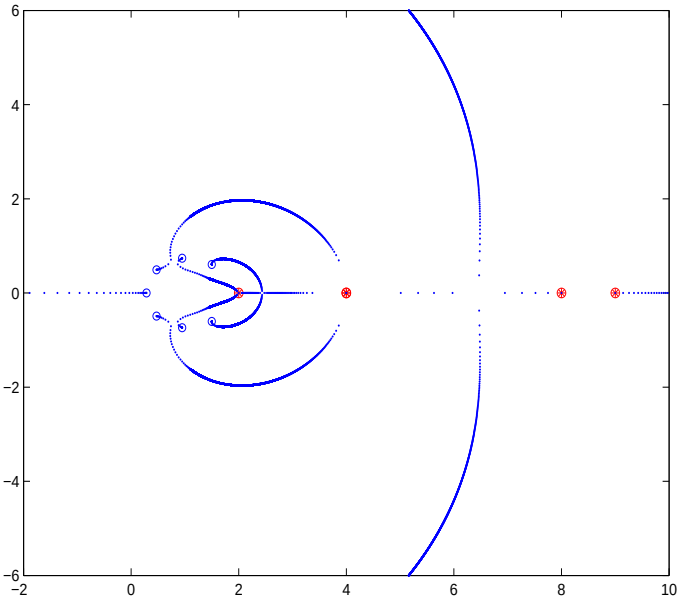


Figure 8.2: Zoom for $0 \leq h \leq 7 \times 10^6$, $h \in \mathbb{R}^+$, and $e_8^T u = 8 \neq 0$

△

8.6.3 The non generic case $u_k = 0$

Let us consider

$$\tilde{\Pi}(z) = \left[\begin{array}{c|c} u_k & 0 \ \ldots \ 0 \ h^- \\ \hline \tilde{u} & H^- - zI \end{array}\right] = \left[\begin{array}{c|c} \Gamma & R \\ \hline L & \Pi^{'} - zI_{g'} \end{array}\right]. \tag{8.6.2}$$

Then using the formula (1.6.8), we have

$$\tilde{q}(z) = \det\tilde{\Pi}(z) = u_k \det(H^- - zI_{g'}) - h^- e_{k-1}^T \mathrm{adj}(zI_{g'} - H^-)\tilde{u}, \tag{8.6.3}$$

which is a scalar polynomial in z of degree $\leq g^{'} = k - 1$.

When $u_k = h_{k,k+1} = 0$, then Ω does not exist, but we can use the formula (8.6.2) and the determinant $\tilde{q}(z)$ in (8.6.3) as follows.

Proposition 8.6.3 *We assume that $\tilde{q}(z)$ defined in (8.6.3) is $\not\equiv 0$.*

1. *When $h_{k,k+1} = u_k = 0$ and $h_{k-1,k+1} \neq 0$, exactly 3 eigenvalues of $B(h)$ go to ∞ as $|t| \to \infty$.*

2. *When $h_{k,k+1} = u_k = u_{k-1} = h_{k-1,k+1} = 0$ and $h_{k-2,k+1} \neq 0$, exactly 4 eigenvalues of $B(h)$ go to ∞ as $|t| \to \infty$.*

3. *The same can be said when more and more successive values of* $h_{i,k+1}$ *are zero for* $i = k, \ldots, 1$.

Proof. When $u_k = h_{k,k+1} = 0$, then an application of Laplace formula on the last column of $\tilde{\Pi}(z)$, yields

$$\tilde{q}(z) = (-1)^{(k-1)+k} u_{k-1} p_{k-2}(z) + (-1)^{(k-2)+k} u_{k-2} p_{k-3}(z) + \ldots \qquad (8.6.4)$$
$$+ \cdots + (-1)^{2+k} u_2 p_1(z) + (-1)^{1+k} u_1 p_0(z).$$

where $p_j(z)$ is a polynomial in z of degree j for $j = 0, \ldots, k-2$. Now, the expansion (8.6.4) shows that

1. When $h_{k,k+1} = u_k = 0$ and $u_{k-1} = h_{k-1,k+1} \neq 0$, then $\tilde{q}(z)$ is a polynomial in z of degree $k-2$, therefore exactly $k-2$ eigenvalues of $B(h)$ remain in finite distance and 3 eigenvalues go to ∞ as $|t| \to \infty$.

2. When $h_{k,k+1} = u_k = u_{k-1} = h_{k-1,k+1} = 0$ and $h_{k-2,k+1} \neq 0$, then $\tilde{q}(z)$ is of degree $k-3$, thus exactly $k-3$ eigenvalues of $B(h)$ remain in finite distance and 4 eigenvalues go to ∞ as $|t| \to \infty$.

3. The same argument can be used when more and more successive values of $h_{i,k+1}$ are zero for $i = k, \ldots, 1$.

□

Example 8.6.2 *If we modify the last entry of the vector* u *in the example 8.6.1 to be* $u = (1, 2, 3, \ldots, 6, 7, 0)^T$ *, then according to the proposition 8.6.3, we expect* 3 *escaping rays, because* $u_8 = 0$ *. This is supported by the Figures 8.3 and 8.4.*

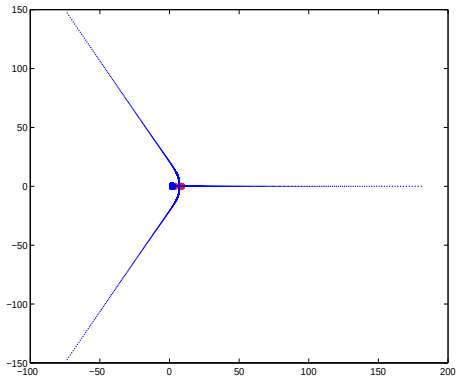


Figure 8.3: $0 \leq h \leq 7 \times 10^6$, $h \in \mathbb{R}^+$, $u_8 = e_8^T u = 0$ and $u_7 = e_7^T u = 7 \neq 0$

If we let h *become extremely large until* $h = 10^{306}$ *, we can still see* 3 *escaping rays. See Figure 8.5.*

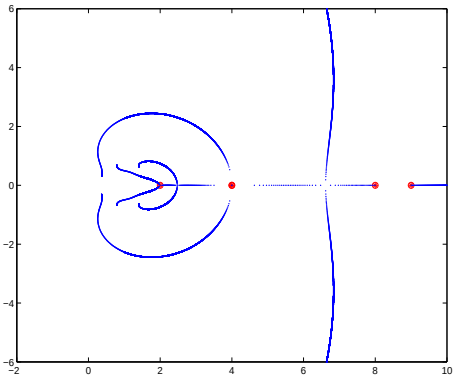


Figure 8.4: Zoom in $[-2, 10] \times [-6, 6]$ for $0 \leq h \leq 7 \times 10^6$, $h \in \mathbb{R}^+$, $u_8 = e_8^T u = 0$ and $u_7 = e_7^T u = 7 \neq 0$

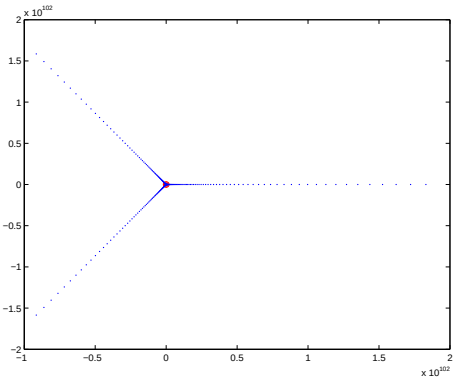


Figure 8.5: $0 \leq h \leq 10^{306}$, $h \in \mathbb{R}^+$, $u_8 = e_8^T u = 0$ and $u_7 = e_7^T u = 7 \neq 0$ with the scale factor 10^{102}

△

8.6.4 What happens when $u_k \to 0$?

Theory tell us that under the condition of the Theorem 8.6.2 where $u_k \neq 0$, exactly two eigenvalues of H^+ escape to ∞ as $|h| \to \infty$. Now, we are interested in the case when $u_k = h_{k,k+1}$ is small and tends to 0.

To this end, let us denote the $k-1$ remaining eigenvalues of H^+ by $\{\lambda_1, \cdots, \lambda_{k-1}\}$. Applying the Hadamard-Gershgorin's theorem 1.3.2 to the matrix

$$\Omega = H^- - \frac{h^-}{u_k} e_{k-1}^T = \begin{bmatrix} h_{11} & h_{12} & \dots & \dots & h_{1,k-2} & (h_{1,k-1} - \frac{h^-}{u_k} u_1) \\ h_{21} & h_{22} & \dots & \dots & h_{2,k-2} & (h_{2,k-1} - \frac{h^-}{u_k} u_2) \\ 0 & \ddots & & & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & h_{k-1,k-2} & (h_{k-1,k-1} - \frac{h^-}{u_k} u_{k-1}) \end{bmatrix},$$

we get that for $i = 1, \cdots, k-2$,

$$\lambda_i \in D_1 = Disk[h_{i,i},\ (\Sigma_{j=i-1}^{k-2} |h_{i,j}|) + |h_{i,k-1} - \frac{h_{k,k-1}}{h_{k,k+1}} h_{i,k+1}|\ ], \tag{8.6.5}$$

and also for $i = k-1$

$$\lambda_{k-1} \in D_2 = Disk[(h_{k-1,k-1} - \frac{h_{k,k-1}}{h_{k,k+1}} h_{k-1,k+1}),\ |h_{k-1,k-2}|], \tag{8.6.6}$$

where $Disk[ce, ra] = \{z \in \mathbb{C};\ |z - ce| \leq ra\}$. For a fixed value of the radius $|h_{k-1,k-2}|$, the inclusion (8.6.6) shows how the position of λ_{k-1} depends on (or is sensitive to) the values of $h_{k-1,k-1}$ $, h_{k,k-1}, h_{k-1,k+1}$ and especially to $u_k = h_{k,k+1}$.

Example 8.6.3 *Let us set $u = (1, 2, \dots, 6, 7, u_k)^T$ for $u_k = 10^{-5}$ in the example 8.6.1. In this case there still exist 2 escaping eigenvalues and 7 eigenvalues which stay at finite distance as $h \to \infty$, but the magnitude of one of the finite limit become very large. As we can see on the Figure 8.6 for $u_k = 10^{-5}$, we must take $h \in [0,\ \ 7 \times 10^{19}]$ to see that there are exactly 2 escaping rays. ($h_M \sim 10^{20}$).*

This is still true for smaller values such as $u_k = 10^{-15}$, $u_k = 10^{-20}, \dots, u_k = 10^{-30}$: in each case the magnitude of one of the finite limits becomes very large and one has to increase the value of h_M in $h \in [0,\ h_M]$ to see the convergence take place. The values of u_k with the corresponding h_M are listed in the Table 8.1.

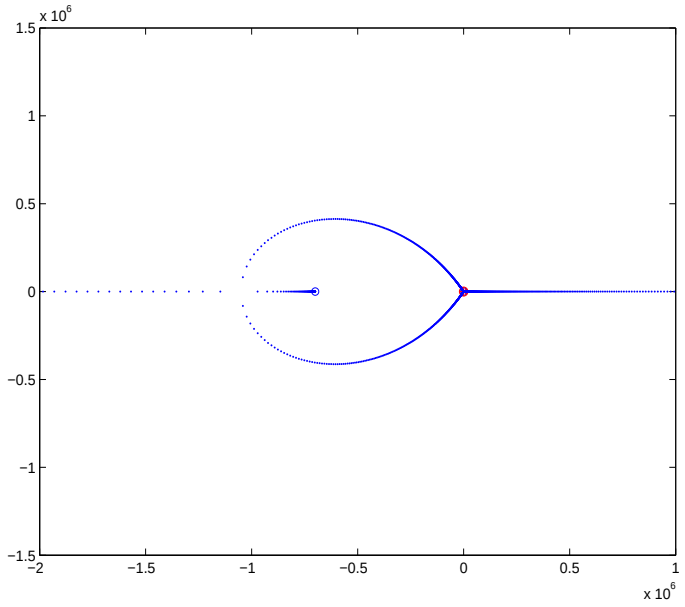


Figure 8.6: $0 \leq h \leq 7 \times 10^{19}$, $h \in \mathbb{R}^+$, $u_k = e_8^T u = 10^{-5}$ with the scale factor 10^6

u_k	10^{-5}	10^{-15}	10^{-20}	10^{-30}
h_M	10^{20}	10^{50}	10^{64}	10^{95}

Table 8.1: Correspondence $u_k \mapsto h_M$, for $\theta = 0$

△

8.7 What have we learnt about Arnoldi?

We have not really progressed in understanding why inherently *algorithmic* methods such as Krylov behave in practice like *direct* methods.

The fact that Krylov-type methods are practically finite algorithms is well illustrated by their robustness to large perturbations. Such a robustness is not exhibited by truly asymptotic methods such as Newton-type methods [24, 25]. As one proceeds towards convergence, the admissible perturbations must be of decreasing norm in Newton-type methods.

By scrutinizing Arnoldi in the light of HD, we have added credit to the marvelous experimental properties of Krylov. But we are still far from being able to *prove* anything meaningful.

Conclusions and perspective

This is the end of our tour of Homotopic Deviation theory. What have we learnt about the linear coupling $A(t) = A + tE$? The essential lesson is that most of the relevant algebraic information is given by the $r \times r$ matrix M_z defined for $z \in re(A)$. The information related to $R(t,z) = (A(t) - zI)^{-1}$ is readily derived from the augmented matrix

$$\hat{A}(z) = \begin{bmatrix} zI - A & -U \\ V^H & 0 \end{bmatrix}$$

where U and V in $\mathbb{C}^{n\times r}$ are given by the SVD of the deviation matrix $E = UV^H$, $r = \text{rank}\, E$, $1 \leq r \leq n$.

The information related to $\text{Lim} \subseteq \lim_{|t|\to\infty} \sigma(A(t))$ is more difficult to get for $r < n$. This indicates a tight spectral coupling between E and A when E is singular. This coupling depends heavily on the Jordan structure of $0 \in \sigma(E)$ when 0 is defective. Because in HD the parameter t belongs to the completed complex plane $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$, one can see HD as the *completion* of the perturbation theory for algebraic singularities which started with Puiseux in 1850 [48] and was later developed in the 20^{th} Century by Baumgärtel [11], Chatelin [32], Kato [42], Lidskii [45], and Wilkinson [55], to mention a few names.

Specific homotopic notions such as frontier, critical and limit points add new flavours to the original notion of singularities in Matrix Algebra. At such points, various radically different behaviours take place. And we have uncovered the unexpected property that these behaviours are extremely *robust to finite precision*. Computer simulations reproduce faithfully the mathematical reality in a variation range for $|t|$ which is many orders of magnitude larger than what we have been accustomed to in classical Numerical Analysis.

Even though we do not fully comprehend to-day the role of these ideas to draw a complete picture of Computation, we understand enough to have confidence that such a role will be essential in furthering our computational understanding of non-linearities.

Personal contributions

I have presented the first complete account of Homotopic Deviation theory which is available in written form. The main sources of information for my research have been the various technical reports, theses and papers published by members of the Qualitative Computing group established at Cerfacs since its origin, 20 years ago. The fundamental results were scattered in various places and supports. When I arrived at Cerfacs in August 2003, I was assigned two goals for my thesis:

1. Write a coherent survey of the theory already available, and advance it whenever possible.

2. Develop computer simulations to study the effect of finite precision on the convergence to limit points.

These two tasks have been fulfilled in the following way (after a 3 months "stage" to learn French):

1. Part I of the thesis manuscript is a self-contained exposition of the theory. It incorporates the following three personal major contributions:

   a. Work with the rational form $\frac{1}{\pi(z)}Q(z)$ for M_z.

   b. Suggest the formula

   $$\det Q(z) = (\pi(z))^{r-1} q(z),$$

   where $q(z)$ is a polynomial of degree $\leq n - r$.

   c. Use the concept of frontier points to analyze the structure of the matrix pencil family $z \mapsto (A - zI) + tE$, where z is a complex parameter. This leads to a novel algorithm to compute the eigenvalues of the pencil, based on the SVD of E, which is extremely cost effective when E has *low* rank.

2. The numerical software developments in Chapters 6 and 7 of Part II of the manuscript are entirely personal. They are my own contribution to the development of efficient visualization tools for HD. In particular, I worked with extremely large or small magnitude for the parameter t, to explore the limits of robustness for homotopic computation.

Bibliography

[1] M. Ahmadnasab. Computing the generalized eigenvalues of $A + tE$ by means of the SVD of E, when $\operatorname{rank} E \ll \operatorname{rank} A$ which is maximum. 2007, in preparation.

[2] M. Ahmadnasab. Homotopic backward analysis for matrix eigenvalues. Presentation in CERFACS May 2005 and Université Versailles Saint-Quentin en Yvelines. 5 July 2005 .

[3] M. Ahmadnasab. Parameter analysis of the structure of regular matrix pencils by homotopic deviation theory. 6th International Congress on Industrial and Applied Mathematics, Zurich, Switzerland 16-20 July 2007.

[4] M. Ahmadnasab and F. Chaitin-Chatelin. Backward analysis for matrix eigenvalues. 2ème Congrès National de Mathématiques Appliquées et Industrielles, Evian - France 23 - 27 May 2005, Evian - France.

[5] M. Ahmadnasab, F. Chaitin-Chatelin, and N. Megrez. Homotopic deviation in the light of algebra. Technical Report TR/PA/05/05, CERFACS, Toulouse, France, 2005.

[6] M. Akian, R. Bapat, and S. Gaubert. Generic asymptotics of eigenvalues and min-plus algebra. INRIA RR-5104, 2004.
Also arXiv:math.SP/0402090, 2004.

[7] M. Akian, R. Bapat, and S. Gaubert. Perturbation of eigenvalues of matrix pencils and optimal assignment problem. *C.R. Acad. Sci. Paris*, **Ser. I339**, pp.103-108, 2004.
Also arXiv: math.SP/0402438, 2004.

[8] R. Alam and S. Bora. Effect of linear perturbation on spectra of matrices. *Linear Algebra and its Applications*, **368**:329–342, 2003.

[9] W. E. Arnoldi. The principle of minimized iteration in the solution of the matrix eigenvalue problem. *Quart. Appl. Math.*, **9**:17–29, 1951.

[10] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst. **Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide**. SIAM, Philadelphia, 2000.

[11] H. Baumgärtel. **Analytic Perturbation Theory for Matrices and Operators**. Akademie Verlag Berlin, 1984.

[12] F. S. V. Bazán. Private conversations. CERFACS, Summer-Fall 2004.

[13] A. Bouras. *Contrôle de convergence des solveurs emboités pour le calcul de valeurs propres avec inversion*. PhD thesis, Université Toulouse I and CERFACS, TH/PA/00/77 CERFACS Toulouse France, 2000.

[14] A. Bouras and V. Frayssé. Inexact matrix-vector products in Krylov methods for solving linear systems: A relaxation strategy. *SIAM J. Matrix Analysis and Applications*, **26**(3):660–678, 2005.

[15] F. Chaitin-Chatelin. Comprendre les méthodes de Krylov en précision finie : le programme du Groupe Qualitative Computing au CERFACS. Technical Report TR/PA/00/11, CERFACS, Toulouse, France, 2000.

[16] F. Chaitin-Chatelin. About singularities in inexact computing. Technical Report TR/PA/02/106, CERFACS, Toulouse, France, 2002.

[17] F. Chaitin-Chatelin. The Arnoldi method in the light of Homotopic Deviation theory. Technical Report TR/PA/03/15, CERFACS, Toulouse, France, 2003.

[18] F. Chaitin-Chatelin. Computing beyond analyticity: Matrix algorithms in inexact and uncertain computing. Technical Report TR/PA/03/110, CERFACS, Toulouse, France, 2003.

[19] F. Chaitin-Chatelin. On Lidskii's algorithm to quantify the first order terms in the asymptotics of a defective eigenvalue, Part I. Technical Report TR/PA/04/129, CERFACS, Toulouse, France, 2004.

[20] F. Chaitin-Chatelin. On Lidskii's algorithm to quantify the first order terms in the asymptotics of a defective eigenvalue, Part II. Technical Report TR/PA/05/04, CERFACS, Toulouse, France, 2005.

[21] F. Chaitin-Chatelin. The dynamics of matrix coupling with an application to Krylov methods. In **Proceedings of NAA2004, Rousse: Bulgaria**, Li Z et al. (eds), *Lecture Notes in Computer Science*, volume **3401** pp. 14-24, Springer Verlag, Berlin 2005.
Also Technical Report TR/PA/04/29 CERFACS Toulouse, France, 2004.

[22] F. Chaitin-Chatelin. A spectral analysis of the link matrix M_z, that is the matrix rational function $z \mapsto \frac{1}{\det(zI-A)} V^H \mathrm{adj}(zI-A)U$ as $z \to \lambda \in \sigma(A)$: the role of the linking polynomial $\hat{\pi}(z)$. CERFACS Working Notes, Toulouse, France, 2007.

[23] F. Chaitin-Chatelin and V. Frayssé. **Lectures on Finite Precision Computations**. SIAM, 1996.

[24] F. Chaitin-Chatelin and T. Meškauskas. Inner-outer iterations for mode solvers in structural mechanics: application to the code-aster. Contract Rep. FR/PA/01/85, CERFACS, 2001.

[25] F. Chaitin-Chatelin, T. Meškauskas, and M. van Gijzen. Inner-outer iterations for power method with Chebyshev acceleration in neutronics. Contract Rep. CR/PA/02/56, CERFACS, 2002.

[26] F. Chaitin-Chatelin, V. Toumazou, and E. Traviesas. Accuracy assessment for eigencomputations: variety of backward errors and pseudospectra. *Linear Algebra and its Applications*, **309**:73–83, 2000.
Also Technical Report TR/PA/99/03, CERFACS, Toulouse, France, 1999.

[27] F. Chaitin-Chatelin and E. Traviesas. Homotopic perturbation - unfolding the field of singularities of a matrix by a complex parameter: a global geometric approach. Technical Report TR/PA/01/84, CERFACS, Toulouse, France, 2001.

[28] F. Chaitin-Chatelin, E. Traviesas, and L. Plantié. Understanding Krylov methods in finite precision. In **Proceedings of NAA2000, Rousse: Bulgaria**, L. Vulkov, J. Wasviewski, P. Yalamov (eds), *LNCS*, volume **1988**, pp. 187-197, Springer Verlag 2000.
Also Technical Report TR/PA/00/40 CERFACS Toulouse, France, 2000.

[29] F. Chaitin-Chatelin and E. Traviesas-Cassan. Qualitative Computing. Chapter **5** in **Accuracy and Reliability in Scientific Computing**, B. Einarsson (ed.), SIAM, Philadelphia, 2005.
Also Technical Report TR/PA/02/58 CERFACS, Toulouse, France, 2002.

[30] F. Chaitin-Chatelin and M. B. van Gijzen. Homotopic deviation with an application to computational acoustics. Technical Report TR/PA/04/05, CERFACS, Toulouse, France, 2004.

[31] F. Chaitin-Chatelin and M. B. van Gijzen. Analysis of parameterized quadratic eigenvalue problems in computational acoustics with homotopic deviation theory. *Numerical Linear Algebra with Applications*, **13**:487–512, 2006.

[32] F. Chatelin. **Spectral Approximation of Linear Operators**. Academic Press, New York, 1983.

[33] F. Chatelin. **Valeurs propres de matrices**. Masson, Paris, 1988.

[34] F. Chatelin. **Eigenvalues of matrices**. Wiley, Chichester, 1993.

[35] T. C. Chen. **Linear system theory and design**. Oxford University Press, 1999.

[36] V. Frayssé. *The power of backward error analysis*, HDR, Institut National Polytechnique de Toulouse and CERFACS, Report TH/PA/00/65 CERFACS, Toulouse, France, 2000.

[37] F. R. Gantmacher. **The theory of matrices**. Vol. I and II, Chelsea Publishing Company, New York, 1960.

[38] Gene H. Golub and Charles F. Van Loan. **Matrix Computations**. John Hopkins University Press, 1983.

[39] M. W. Ho, P. Saunders, and S. Fox. A new paradigm for evolution. *New Scientist*, pp. 41-43, February 27, 1986.

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

[41] A. Ilahi. *Validation du calcul sur ordinateur: application de la théorie des singularités algébriques.* PhD thesis, Université Toulouse I and CERFACS, TH/PA/98/31 CERFACS Toulouse France, 1998.

[42] T. Kato. **Perturbation Theory for Linear Operators**. Springer Verlag, New York, 1965.

[43] P. Lancaster and M. Tismenetsky. **Theory of Matrices**. Academic Press, New York, 1987.

[44] R. B. Lehoucq, and D. C. Sorensen. Deflation techniques within an implicitly restarted Arnoldi iteration. *SIAM J. Matrix Anal. Appl.*, **17**:789–821, 1996.

[45] V. B. Lidskii. Perturbation theory of non-conjugate operators. *U.S.S.R Comput. Math. Phys.*, **1**:73–85, 1965.

[46] J. Moro, J. Burke, and M. Overton. On the Lidskii-Vishik-Lyusternik perturbation theory for eigenvalues of matrices with arbitrary Jordan structure. *SIAM J. Matrix Analysis and Applications*, **18**:793–817, 1997.

[47] N. B. Parlett. **The Symmetric Eigenvalue Problem**. Prentice Hall, Englewood Cliffs, 1980.

[48] V. Puiseux. Recherches sur les fonctions algébriques. *J. Math. Pures Appl.*, **51**:465–581, 1850.

[49] Y. Saad. **Numerical methods for large eigenvalue problems**. Halsted Press, 1992.

[50] G. L. G. Sleijpen, J. van den Eshof, and M. B. van Gijzen. Restarted GMRES with inexact matrix-vector products. In **Proceedings of NAA2004, Rousse: Bulgaria**, *LNCS*, volume **3401** pp. 494-501, Springer Verlag, 2005.

[51] E. Traviesas. *Sur le déploiement du champ spectral d'une matrice*. PhD thesis, Université Toulouse I and CERFACS, TH/PA/00/30 CERFACS Toulouse France, 2000.

[52] L. N. Trefethen and D. Bau III. **Numerical Linear Algebra**. SIAM, 1997.

[53] H. F. Walker. Implementation of the GMRES method using Householder transformations. *SIAM J. Sci. Statist. Comput.*, **9**:152–163, 1988.

[54] J. H. Wilkinson. **Rounding Errors in Algebraic Processes**. vol. **32**, Her Majesty's Stationary Office, London, 1963.

[55] J. H. Wilkinson. **The Algebraic Eigenvalue Problem**. Oxford University Press, 1988.