UNIVERSITÀ DEGLI STUDI DI TRIESTE

UNIVERSITÀ DEGLI STUDI DI TRIESTE

XXXIV CICLO DEL DOTTORATO DI RICERCA IN

Scienza della Terra, Meccanica dei Fluidi e Matematica
Interazioni e Metodiche

Relative conditioning of linear systems of ODEs with respect to perturbation in the matrix of the system and in the initial value

Settore scientifico-disciplinare: **MAT/08**

DOTTORANDA
Asma Farooq

COORDINATORE
PROF. Stefano Maset

SUPERVISORE DI TESI
PROF. Stefano Maset

ANNO ACCADEMICO 2020/2021

[This page intentionally left blank]

Abstract

The thesis is about how perturbations in the initial value y_0 or in the coefficient matrix A propagate along the solutions of n-dimensional linear ordinary differential equations (ODE)

$$\begin{cases} y'(t) = Ay(t), \ t \geq 0, \\ y(0) = y_0, \end{cases}$$

where $A \in \mathbb{R}^{n\times n}$ and $y_0 \in \mathbb{R}^n$ and $y(t) = e^{tA}y_0$ is the solution of the equation.

The paper [59] considers a perturbation analysis when the initial value y_0 is perturbed to $\tilde{y}_0$ with relative error

$$\varepsilon = \frac{\|\tilde{y}_0 - y_0\|}{\|y_0\|},$$

where $\|\cdot\|$ is a vector norm on $\mathbb{R}^n$. Due to the perturbation in the initial value, the solution $y(t) = e^{tA}y_0$ is perturbed to $\tilde{y}(t) = e^{tA}\tilde{y}_0$ with relative error

$$\delta(t) = \frac{\left\|e^{tA}\tilde{y}_0 - e^{tA}y_0\right\|}{\|e^{tA}y_0\|}.$$

In other words, the paper studies the (relative) conditioning of the problem

$$y_0 \mapsto e^{tA}y_0.$$

It describes the relation between the error ε and the error $\delta(t)$ by three condition numbers namely: the condition number with the direction of perturbation, the condition number independent of the direction of perturbation and the condition number not only independent of the specific direction of perturbation but also independent of the specific initial value. How these condition numbers behave over a long period of time is an important aspect of the study.

We remark that in literature any relative error perturbation analysis considers the matrix exponential e^{tA}, not the matrix exponential as applied to y_0, namely the vector quantity $e^{tA}y_0$. Moreover, it is missing a study of how the conditioning depends on the time t.

In the thesis, we move towards perturbations in the matrix as well as componentwise relative errors, rather than normwise relative errors, for perturbations of the initial

value. The contents of the thesis have given rise to the two papers [27] and [26].

About the first topic of the thesis (whose contents are in [27]), we look over how perturbations propagate along the solution of the ODE, when it is the coefficient matrix A rather than the initial value that perturbs. In other words, the interest is to study the conditioning of the problem

$$A \mapsto e^{tA} y_0.$$

In case when the matrix A perturbs to $\tilde{A}$, the relative error is given by

$$\epsilon = \frac{\left|\left|\left| \tilde{A} - A \right|\right|\right|}{|||A|||},$$

where $|||\cdot|||$ is a matrix norm, and the relative error in the solution of the ODE is given by

$$\xi(t) = \frac{\left\| e^{t\widetilde{A}} y_0 - e^{tA} y_0 \right\|}{\|e^{tA} y_0\|}.$$

The aim is to describe the relation between ϵ and $\xi(t)$. We introduce three condition numbers similarly to [59]. The analysis of the condition numbers is done for a normal matrix A and by making use of 2-norm . We give very useful upper and lower bounds on these three condition numbers and we study their asymptotic behavior as time goes to infinity. We quote here one of our results about the condition number $K(t, A, y_0)$ independent of the direction of perturbation. The result is about the asymptotic behavior in a generic situation for the initial value: if y_0 has a nonzero projection on the eigenspaces of the rightmost eigenvalues, then

$$K(t, A, y_0) \sim \|A\|_2 \, t, \ t \to +\infty.$$

In the paper [59], the conditioning of the problem $y_0 \mapsto e^{tA} y_0$ has been studied by considering a normwise relative error. There could be cases when someone is interested in the relative errors

$$\delta_l(t) = \frac{|\tilde{y}_l(t) - y_l(t)|}{|y_l(t)|}, \quad l = 1, \ldots, n,$$

of the perturbed solution components. These componentwise relative errors $\delta_l(t)$ could be very different from the normwise relative error $\delta(t)$. Indeed we can have a small $\delta(t)$, but some large componentwise relative error $\delta_l(t)$. Vice versa, when all the componentwise relative errors $\delta_l(t)$ are small, $\delta(t)$ is also small.
With the motivation that componentwise relative errors give more information than the normwise relative error, we make a componentwise relative error analysis, which is the other topic of this thesis (whose contents are in [26]).

We consider perturbations in initial value y_0 with normwise relative error ε and the relative error in the components of the solution of the equation given by the $\delta_l(t)$. The interest is to study, for the l-th component, the conditioning of the problem

$$y_0 \mapsto y_l(t) = e_l^T e^{tA} y_0,$$

where e_l^T is the l-th vector of the canonical basis of $\mathbb{R}^n$. We make this analysis for a diagonalizable matrix A, diagonalizability being a generic situation for the matrix A. We give two condition numbers: a condition number with the direction of perturbation and a condition number independent of the direction of perturbation.

We state here one of our results describing, in a generic situation for y_0, the long-time behavior of the condition number $K_l(t, A, y_0)$ independent of the direction of perturbation. Suppose that A diagonalizable has a unique real eigenvalue λ_1 of multiplicity one, or a unique pair λ_1 and $\lambda_2 = \overline{\lambda_1}$ of complex conjugate eigenvalues of multiplicity one, as rightmost eigenvalues. Let v be an eigenvector of λ_1 and let w be the first row of $W = V^{-1}$, V being the matrix of the eigenvectors with v as first column. Assume $v_l \neq 0$. If $wy_0 \neq 0$, then

$$K_l(t, A, y_0) \to \frac{\|w\| \, \|y_0\|}{|wy_0|}, \quad t \to +\infty,$$

when the rightmost eigenvalue is the real eigenvalue and

$$K_l(t, A, y_0) \sim \frac{\left\| \mathrm{Re}\left(\mathrm{e}^{\sqrt{-1}\omega_1 t} v_l w \right) \right\| \, \|y_0\|}{\left| \mathrm{Re}\left(\mathrm{e}^{\sqrt{-1}\omega_1 t} v_l w \right) y_0 \right|}, \quad t \to +\infty,$$

when the rightmost eigenvalues are the complex conjugate pair.

Dedication

Dedicated to my dear husband Dr. Shafqat Ali,
my mother and my son Maaz Ali.

Acknowledgement

This PhD thesis has the contributions of several people. First of all, I would like to pay my deepest gratitude to my advisor, *Prof. Stefano Maset,* for sharing his knowledge, patient guidance, encouragement, and advice throughout this dissertation. I feel extremely lucky to have a supervisor who has always been there for my help and who cared so much about my work.

I would like also to extend my gratitude to *Prof. Vanni Noferini* for interesting discussion, constructive suggestions, and collaboration in the last few months of my PhD.

I am grateful to all my colleagues at the University of Trieste, who helped me in my research activities.

My gratitude also goes to the student secretariat office for helping and assisting me during the last four years.

Some special words of gratitude go to all my friends here in Italy and back in Pakistan for being consistently emotional and moral support.

Above all, I would like to express my deep love and gratitude to my dear husband *Dr. Shafqat Ali*. Thanks, for being my biggest support and a source of great motivation throughout these years. Thanks for believing in me, being always along with me and for being my best friend. I owe you everything.

Lastly, I pay my deepest gratitude to my mother who has been consistent support throughout my life. Without her support, I would not have reached where I am today. My inlaws, my brother *Umar* my sister *Humaira* my dear son *Maaz*, cute Irhaa and Salar and other family members I am very grateful to all of you for being there for me. I would not have been able to pursue my dreams without your support.

Asma Farooq, Trieste, Italy 2022

Contents

Chapter 1

Introduction

Numerical analysts have developed a large number of algorithms for solving problems numerically, not having or difficult to compute analytic solutions. Execution of tedious and cumbersome calculations have become possible with advancement of the digital computers. However, due to the limited memory of computers it is only possible to store a finite precision of real numbers i.e we can only give approximations. For example, we can store $\frac{1}{3} = 0.33333\ldots$, up to a finite number of bits. Representation of floating points by finite precision causes errors. Such errors pill up to a greater extend when a large number of computations are made. Hence these errors concern serious attention otherwise they can lead to irrelevant results.
This issue was initially considered by known mathematician Alan Turing in (1948) (see [76]). Errors can be of different types like round off errors, truncation errors, discretization errors, modeling errors and input errors. After being introduced with errors the next important question arises in mind is "How to measure the error"? Is the measured error "big" or "small"? Expressions like "big" or "small" error gives rise to the definition of relative error since we can only declare a value big or small if we compare it with some other quantity. This is exactly the definition of relative error, as it compares magnitude of absolute error to the magnitude of true value.
Let $x \in \mathbb{R}$ with $x \neq 0$ and let $\tilde{x}$ be the approximation of x, the relative error is given by

$$RelErr(x, \tilde{x}) = \frac{AbsErr(x, \tilde{x})}{|x|},$$

where

$$AbsErr(x, \tilde{x}) = |\tilde{x} - x|,$$

is the absolute error. We care much about the relative error rather than absolute error since relative error being a dimensionless quantity conveys more meaning. For $x \in \mathbb{R}^n$, we can extend definition of relative error in two ways.
Componentwise relative error: For $x \in \mathbb{R}^n$ we consider the relative error of each component of x. We define the relative error of x, where x is such that $x_i \neq 0$ for $i = 1, \ldots, n$, as

$$RelErr(x, \tilde{x}) = \max_{i=1,\ldots,n} \frac{|\tilde{x}_i - x_i|}{|x_i|}. \tag{1.0.1}$$

Normwise relative error: For a norm $\|.\|$ defined on $\mathbb{R}^n$ we can mimic the definition for the scalar case as

$$RelErr(x, \tilde{x}) = \frac{\|\tilde{x} - x\|}{\|x\|}.$$

The next question is how these errors affect the computation of a problem?
Suppose that a problem is describe by a function

$$g : \mathbb{R}^m \to \mathbb{R}^n.$$

Further, suppose that the computation of the function g is affected by some input perturbation. In response to this input perturbation, does the output perturb a little or it blow up? Studying these questions is known as sensitivity analysis.
In this thesis we are going to study the relation between the relative error in the input data and the relative error in the output of the problem, when the problem is described by a n-dimensional linear ordinary differential equation whose solution involves the matrix exponential function.

The rest of this chapter presents succinct definitions which are fundamental for the whole thesis.

1.1 Condition Numbers

Suppose that X and Y are normed spaces and consider a function

$$f : X \mapsto Y, \quad f(x) = y.$$

Let the input x be perturbed to $\tilde{x}$ (by finite precision or by some other reason) and let

$$RelErr(x, \tilde{x}) = \frac{\|\tilde{x} - x\|}{\|x\|}$$

be the normwise relative error of the input.
In correspondence with the input data the output data also perturbs with the following relative error

$$RelErr(y, \tilde{y}) = \frac{\|\tilde{y} - y\|}{\|y\|}.$$

How large the magnification of the error is? That is how the perturbation in input data x affects the output y? We can provide an answer to this question by comparing both relative errors, i.e by considering the ratio

$$\frac{RelErr(y, \tilde{y})}{RelErr(x, \tilde{x})}.$$

If this ratio is not large the problem is well conditioned. On the other hand, the bigger this ratio, the more sensitive with respect to input perturbations the problem in hand

is. To have more information we bound the error in input by a small number ϵ and we consider the worst case among all such errors bounded by ϵ. Mathematically, we consider

$$\sup_{RelErr(x,\tilde{x})\leq\epsilon} \frac{RelErr(y,\tilde{y})}{RelErr(x,\tilde{x})}.$$

Taking one step further, we consider ϵ arbitrarily small by taking the limit as ϵ tending to 0. This gives rise to the definition of condition number:

$$cond_{abs}(f,x) = \lim_{\epsilon\mapsto 0} \sup_{RelErr(x,\tilde{x})\leq\epsilon} \frac{RelErr(y,\tilde{y})}{RelErr(x,\tilde{x})}.$$

The nature of the condition number was introduced by the Alan Turing [76] and later on by John von Neumann and Herman H. Goldstine [79] in order to understand the accuracy of the solution of linear systems when solved by some computing machine. After these fundamental articles condition numbers have been vastly investigated. Peter Bürgisser and Felipe Cucker mentioned in their book [11]:
A combined search by Mathscinet and Zentralblatt shows more than 800 *articles with expression "condition numbers".*
For a worth reading of this ubiquitous topic see ([15, 20, 36, 49, 53, 64, 65, 69, 85]).
Formal definitions of the absolute and relative condition number are

$$cond_{abs}(f,x) = \lim_{\epsilon\mapsto 0} \sup_{\|\Delta x\|\leq\epsilon} \frac{\|f(x+\Delta x)-f(x)\|}{\|\Delta x\|}$$

and

$$cond_{rel}(f,x) = \lim_{\epsilon\mapsto 0} \sup_{\|\Delta x\|\leq\epsilon\|x\|} \frac{\|f(x+\Delta x)-f(x)\|\,\|x\|}{\|\Delta x\|\,\|f(x)\|}. \tag{1.1.1}$$

Clearly, the condition number gives the worst possible magnification of error and quite obviously it depends upon the norm used.
Observe that

$$cond_{rel}(f,x) = cond_{abs}(f,x)\frac{\|x\|}{\|f(x)\|}$$

However, it is much more difficult to calculate relative condition number rather than absolute condition number.
In [65], we find expressions of the condition numbers in terms of the norm of Fréchet derivative:

$$cond_{abs}(f,x) = \left\|f'(x)\right\|$$

and

$$cond_{rel}(f,x) = \frac{\|f'(x)\|\,\|x\|}{\|f(x)\|}, \tag{1.1.2}$$

where the linear map $f'(x) : X \mapsto Y$ is the Fréchet derivative. In computing condition numbers, either relative or absolute, the actual difficulty is to estimate $\|f'(x)\|$. In [41, Ch. 3], Nicholas J.Higham has provided algorithms to estimate such quantities for matrices. In general condition numbers are expensive to compute. Being a key concept

in numerical linear algebra many numerical analysts focus on providing an inexpensive estimate and bounds on condition numbers . In this thesis we will study condition numbers and their long term behavior for a problem modeled by linear ordinary differential equations. However, our focus is a qualitative study of the condition numbers rather than the computational aspects.

1.2 Matrix Exponential

A rigorous notion of a function of matrix can be given in several ways. For instance, let $f(x)$ be a polynomial function. We can define this function for a matrix argument simply by replacing x by A in the expression for $f(x)$. For a detailed study see [18, 29, 32, 41, 43, 56, 66].
Given a function $f(x)$ that is expressed by a power series, i.e. $f(x) = \sum_{k=0}^{\infty} a_k x^k$, we can extend it to matrix arguments. In this aspect, much more attention is needed on convergence issue. We can define f for a matrix input if the spectrum (set of eigenvalues) of A lies within the radius of convergence of the power series. For example, the logarithm function

$$\log(I + A) = A - \frac{A^2}{2!} + \frac{A^3}{3!} - \dots$$

is defined for $\rho(A) < 1$, where $\rho(A)$ is the spectral radius of the matrix A. The condition $\rho(A) < 1$ is analogous to $|x| < 1$ valid for the scalar version and it guarantees convergence of the series. Another example is the matrix exponential function,

$$e^A = I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \dots, \tag{1.2.1}$$

which is defined for any matrix A and it is our main focus because of its role in linear differential equations. The matrix exponential function has attracted attention of many authors in past few decades (see [3, 12, 35, 41, 47, 62, 70, 73, 77, 87]). Its computation can be done by dozen of methods and a comprehensive overview of such efforts has been given by [61]. See also [31].
The definition of the matrix exponential function given by (1.2.1) is equivalent to the following one

$$e^A = \lim_{n \mapsto \infty} \left(I + \frac{A}{n} \right)^n, \tag{1.2.2}$$

(see [74]).
For more interesting representations of matrix exponential see [41, Ch. 10].
Unlike the scalar version of the exponential, in general for two matrices A and B of the same dimension we have $e^{A+B} \neq e^A e^B$. Indeed, the following theorem holds.

Theorem 1.2.1. *For $A, B \in \mathbb{C}^{n \times n}$, $e^{A+B} = e^A e^B$ if and only if $AB = BA$.*

For a proof see [41, Ch. 10].
The matrix exponential e^{A+B} has been studied by a number of authors. For example

Trotter in [75] gave a product formula for e^{A+B}. Gantmacher [25] drives some useful results for e^{A+B}. We refer interesting readers to see [35, 57, 78, 83].

For a $n \times n$ diagonalizable matrix A with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, any function of matrix f as applied to A is given by

$$f(A) = \sum_{i=1}^{n} f(\lambda_i) P_i,$$

where P_i is the projection matrix on the eigenspace corresponding to the eigenvalue λ_i. In particular, for the exponential matrix we have

$$e^A = \sum_{i=1}^{n} e^{\lambda_i} P_i,$$

which is of our interest.

1.3 Linear Ordinary Differential Equations(ODEs)

Ordinary differential equations are a powerful tool to model real world problems effectively, see[6, 13, 14, 23, 44, 58, 60] . This thesis considers the following linear n-dimensional ODE

$$\begin{cases} y'(t) = Ay(t), \ t \geq 0, \\ y(0) = y_0, \end{cases} \tag{1.3.1}$$

where $A \in \mathbb{R}^{n \times n}$, whose solution is $y(t) = e^{tA} y_0$ and analyses the error in the solution when the matrix A or the initial value y_0 are perturbed.
The error analysis when the matrix is perturbed requires the knowledge of Fréchet derivative of the map $A \mapsto e^{tA}$. By the following explicit expression for $e^{t(A+B)}$:

$$e^{t(A+B)} = e^{tA} + \int_0^t e^{(t-s)A} B e^{sA} ds + O(\|B\|^2), \tag{1.3.2}$$

found in [8], we see that such Fréchet derivative is the linear operator $L(t, A, \cdot) : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$ given by

$$L(t, A, B) = \int_0^t e^{(t-s)A} B e^{sA} ds. \tag{1.3.3}$$

Suppose that the matrix A of the linear ODE is perturbed to $\tilde{A}$ or the initial value y_0 is perturbed to $\tilde{y}_0$. Let ϵ be the relative error in the matrix perturbation:

$$\epsilon = RelErr(A, \tilde{A}) = \frac{|||\tilde{A} - A|||}{|||A|||}, \tag{1.3.4}$$

where $|||.|||$ is a matrix norm and let

$$\varepsilon = RelErr(y_0, \tilde{y}_0) = \frac{\|\tilde{y}_0 - y_0\|}{\|y_0\|}, \tag{1.3.5}$$

be the relative error in the initial value, where $\|.\|$ is a vector norm. In response to a perturbation in matrix, the solution $y(t)$ of the ODE perturbs to $\tilde{y}(t)$. The Relative error in the solution is given by $\delta(t)$

$$\delta(t) = RelErr(y(t), \tilde{y}(t)) = \frac{\|\tilde{y}(t) - y(t)\|}{\|y(t)\|}, \tag{1.3.6}$$

Our focus in this thesis is to explore the relation between ϵ or ε and $\delta(t)$.

This relative error analysis has not been considered in literature where some authors only consider absolute errors of perturbed solution arising from perturbed initial value. For example [45] gives the following bound for the absolute error:

$$\|\tilde{y}(t) - y(t)\| \leq M(t)e^{t\alpha(A)} \|\tilde{y}_0 - y_0\|, \quad t \geq 0,$$

where $\alpha(A)$ is the spectral abscissa, i.e the maximum real part of eigenvalues of A, and $M(t)$ grows polynomially with t.
Another bound on the absolute error is the following one:

$$\|\tilde{y}(t) - y(t)\| \leq e^{t\mu(A)} \|\tilde{y}_0 - y_0\|, \quad t \geq 0,$$

where $\mu(A)$ is logarithmic norm of A defined as

$$\mu(A) = \lim_{h \mapsto 0^+} \frac{|||I + hA||| - 1}{h},$$

where $|||.|||$ is the matrix norm induced by the vector norm. Perturbation analysis of the matrix exponential e^{tA} is also given by [4, 32, 41, 42]. However, these papers ignored the role of initial value in their analysis, that is part of our study.

As mentioned by [11], one can consider relative error in two ways, componentwise and normewise. The errors ϵ, ε, and $\delta(t)$ are normwise relative errors. In this thesis we also consider the componentwise relative errors

$$\delta_l(t) = \frac{|\tilde{y}_l(t) - y_l(t)|}{|y_l(t)|}, \; l = 1, \ldots, n, \tag{1.3.7}$$

of the perturbed solution $\tilde{y}(t)$.
The norms play a central role in perturbation theory by associating a single number to a $m \times n$ matrix. That conveys perturbation results immediately. Saying, without norm the perturbation theory would not be as rich as it is today, is not wrong. Being such an important tool we can not ignore the role of norm in perturbation theory but still there

are a few failures of norms. We address one of them, since the norm provides an overall size of perturbation but it disregards how the size is allocated among each element. However, this could be useful and required information in case when data is sparse or badly-scaled. To overcome this drawback of norms *"componentwise perturbation analysis"* has became popular among researchers (see [9, 17, 67, 84]). Nicholas J. Higham in his survey of componentwise perturbation theory [40] provides a brief review of efforts made for componentwise perturbation theory. In componentwise perturbation theory we can consider two types of condition numbers mixed and componentwise condition number, see [11] and for a linear system see [37, 85]. Mixed and the componentwise condition number, for the structured matrices are given by [30, 84], for Moore-Penrose inverse and the linear least square problem are given by [17] and for symmetric algebraic Riccati equation are given by [86].

However, normwise and componentwise relative errors are related in the following manner.

Remark 1.3.1. *For a vector $u \in \mathbb{R}^n$ not having any zero component, perturbed to a vector $\tilde{u} \in \mathbb{R}^n$, we have*

$$\frac{\|\tilde{u}-u\|}{\|u\|} \leq \max_{i=1,\cdots,n} \frac{|\tilde{u}_i - u_i|}{|u_i|}$$

and

$$\frac{|\tilde{u}_i - u_i|}{|u_i|} \leq \frac{\|u\|}{|u_i|}.\frac{\|\tilde{u}-u\|}{\|u\|}, \quad i = 1, \cdots, n.$$

Above bounds holds for a norm that satisfies

$$|v_i| \leq \|v\|, \ v \in \mathbb{R}^n \text{ and } i = 1, \cdots, n,$$
$$\|v\| = \|(|v_1|, \cdots, |v_n|)\|, \quad v \in \mathbb{R}^n$$
$$\|v\| \leq \|w\|, \ v, w \in \mathbb{R}^n \text{ such that } |v_i| \leq |w_i| \text{ and } i = 1, \cdots, n.$$

These conditions are satisfied by $p-$norms.

1.4 Contents the of Thesis

The thesis is arranged in the following manner.

The contents of the second chapter are based upon a paper by S.Maset [59], it sets the stage for the rest of chapters. The chapter describes the relation between the error ε in (1.3.5) on the initial value and the error $\delta(t)$ in (1.3.6) on the solution, by three condition numbers namely: the condition number with direction of perturbation, the condition number independent of direction of perturbation and the condition number not only independent of the specific direction of perturbation but also independent of the specific initial value. How these condition numbers behave in a long period of time

is an important aspect of the chapter. The analysis of these three condition numbers is done in case of an n-dimensional ODE (1.3.1) described by a normal matrix.
The novelty of the thesis is given in the third and fourth chapters: The third chapter is devoted to study how perturbations propagate along the solution of linear ODE, when these perturbations are considered in the matrix. In other words, it studies the relation between the error ϵ in (1.3.4) on the matrix and the error $\delta(t)$ in (1.3.6) on the solution. As in the previous chapter 2, the analysis is done for a normal matrix and we introduce three similar condition numbers. We give very useful upper and lower bounds on these three condition numbers and we study their asymptotic behavior as time goes to infinity. To verify our analysis, a number of numerical tests are part of this chapter.

In the fourth chapter of this thesis we make a componentwise perturbation analysis for the ODE (1.3.1) by taking into account perturbations in the initial value. In this componentwise perturbation analysis we study the relation between the errors $\delta_l(t)$, $l = 1, \ldots, n$, in (1.3.7) for the solution and the normwise relative error ε in (1.3.5) for the initial value. Here the analysis is done for a diagonalizable matrix not only for normal matrix.

Finally, we have the conclusion section. Which gives a brief summary of the whole thesis and a few observations for possible arising queries.

Chapter 2

Perturbations in the initial value: a normwise relative error analysis

This chapter is about the perturbation analysis of the linear ODE (1.3.1) when it is the initial value that goes under perturbation. i.e we study the conditioning of the problem

$$y_0 \mapsto e^{tA} y_0. \tag{2.0.1}$$

As mentioned in the introductory chapter, the contents of this chapter are based upon the paper [59]. The chapter gives the error analysis of the linear ODE (1.3.1) defined by a normal matrix A by introducing three condition numbers. Namely, a condition number with a specific direction of perturbation of the initial value, a condition number independent of the direction of perturbation and a condition number independent of the specific initial value. Then it studies the asymptotic behavior of these three condition numbers.

Conditioning studies concern relative errors rather than absolute errors. As already mentioned in Chapter 1, the relative error is a dimensionless quantity and being a dimensionless quantity it conveys more meaning. In order to illustrate the different behaviors of the relative and the absolute errors, we consider the simple case of a scalar ODE:

$$\begin{cases} y'(t) = ay(t), \ t \geq 0, \\ y(0) = y_0 \end{cases}$$

where a is any real number. The solution of the equation is given by $y(t) = e^{at} y_0$. The vector $y_0 \in \mathbb{R}$ is a non-zero initial value. Suppose that y_0 is perturbed to $\widetilde{y}_0$. The perturbation in the initial value results in a perturbation in the solution $y(t)$, which takes the form $\widetilde{y}(t) = e^{at}\widetilde{y}_0$. First, we measure the error in the solution by the absolute error and we have

$$|\widetilde{y}(t) - y(t)| = |e^{at}\widetilde{y}_0 - e^{at} y_0| = e^{at}|\widetilde{y}_0 - y_0| \tag{2.0.2}$$

where $|\tilde{y}_0 - y_0|$ is the absolute error in the initial value.
By equation (2.0.2), we can observe that the absolute error can be increasing or decreasing with time t, depending upon the sign of a .
Next, we measure the error in the solution by the relative error

$$\delta(t) = \frac{|\tilde{y}(t) - y(t)|}{|y(t)|}$$

of this solution. The above expression makes sense because $y(t) = e^{at}y_0 \neq 0$ since $y_0 \neq 0$.
On the other hand, let

$$\varepsilon = \frac{|\tilde{y}_0 - y_0|}{|y_0|}.$$

be the relative error in initial value. Now

$$\delta(t) = \frac{|e^{at}\tilde{y}_0 - e^{at}y_0|}{|e^{at}y_0|} = \frac{|\tilde{y}_0 - y_0|}{|y_0|} = \varepsilon. \tag{2.0.3}$$

Equation (2.0.3) shows that the relative error does not change with time for any coefficient a. By looking at the equation (2.0.2) and (2.0.3), we can conclude that the absolute and the relative errors behave quite differently.
Next example illustrates the difference of behavior of the errors in case of a system of ODEs rather than a scalar ODE.

Example 2.0.1. *Consider the following linear ODE*

$$\begin{cases} y'(t) = Ay(t), \ t \geq 0, \\ y(0) = y_0, \end{cases} \tag{2.0.4}$$

with the symmetric matrix

$$A = \begin{bmatrix} -2.505 & 2.495 \\ 2.495 & -2.505 \end{bmatrix},$$

and the initial value

$$y_0 = (1, -1)\,,$$

and the perturbed initial value

$$\tilde{y}_0 = y_0 + (0.01, 0.01)\,.$$

Making use of the euclidean norm $\|.\|_2$, the relative error in the initial value is given by

$$\epsilon = \frac{\|\tilde{y}_0 - y_0\|_2}{\|y_0\|_2}$$

and the relative error in the solution is given by

$$\delta(t) = \frac{\|\tilde{y}(t) - y(t)\|_2}{\|y(t)\|_2}.$$

The absolute error in $\|\tilde{y}(t) - y(t)\|_2$ *the solution satisfies*

$$\|\tilde{y}(t) - y(t)\|_2 \leq e^{-0.01t} \|\tilde{y}_0 - y_0\|_2 , \tag{2.0.5}$$

where $\lambda_1 = -0.01$ *is the maximum eigenvalue of the matrix A. In figure* 2.1 *and* 2.2*, we see the behavior of the relative error and the absolute error, respectively, for* $t \in [0, 3]$*. We use the MATLAB function expm to compute values of* $y(t)$ *and* $\tilde{y}(t)$*. The figure* 2.1 *shows an explosion in time of the relative error. On the other hand, in the figure* 2.2*, we can observe that the absolute error has a slow decrease in time as predicted by* (2.0.5)*.*

Even though the matrix A is considered stable because both its eigenvalues are negative, it is stable only in the sense of the absolute error, but at the same time it is unstable for the relative error. This gives a better understanding of the difference of behaviors of two errors. Thus, the propagation in time of the relative error in the initial data needs to be explored.

Now, we introduce the three condition numbers for the problem (2.0.1). We consider the more general situation of a linear problem

$$u \mapsto v = Bu \tag{2.0.6}$$

where B is $n \times n$ real matrix. Let $\|.\|$ be a vector norm on $\mathbb{R}^n$. With the same symbol we denote the induced matrix norm on $\mathbb{R}^{n\times n}$. Suppose that the data $u \neq 0$ is perturbed to $\tilde{u}$. In response to this perturbation in the data the solution v ends up with $\tilde{v} = B\tilde{u}$. We specify the perturbed data in the following manner

$$\tilde{u} = u + \varepsilon \|u\| \hat{z}_0,$$

where the unit vector

$$\hat{z}_0 = \frac{\tilde{u} - u}{\|\tilde{u} - u\|},$$

is the direction of perturbation. The relative error in the data is given by

$$\varepsilon = \frac{\|\tilde{u} - u\|}{\|u\|} \tag{2.0.7}$$

and the relative error in the solution is

$$\delta = \frac{\|\tilde{v} - v\|}{\|v\|}, \tag{2.0.8}$$

assuming that $v \neq 0$.
The relation between ε and δ is given by

$$\delta = \frac{\|B\hat{z}_0\| \|u\|}{\|Bu\|}\varepsilon = \frac{\|B\hat{z}_0\|}{\|B\hat{u}\|}\varepsilon,$$

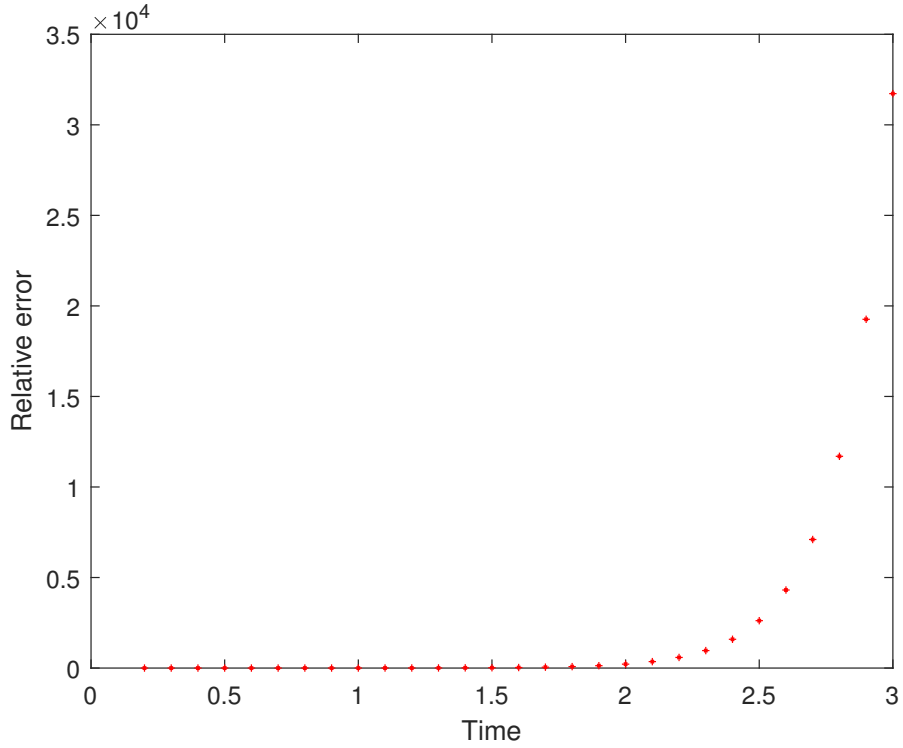

Figure 2.1: Relative error in solution of equation.

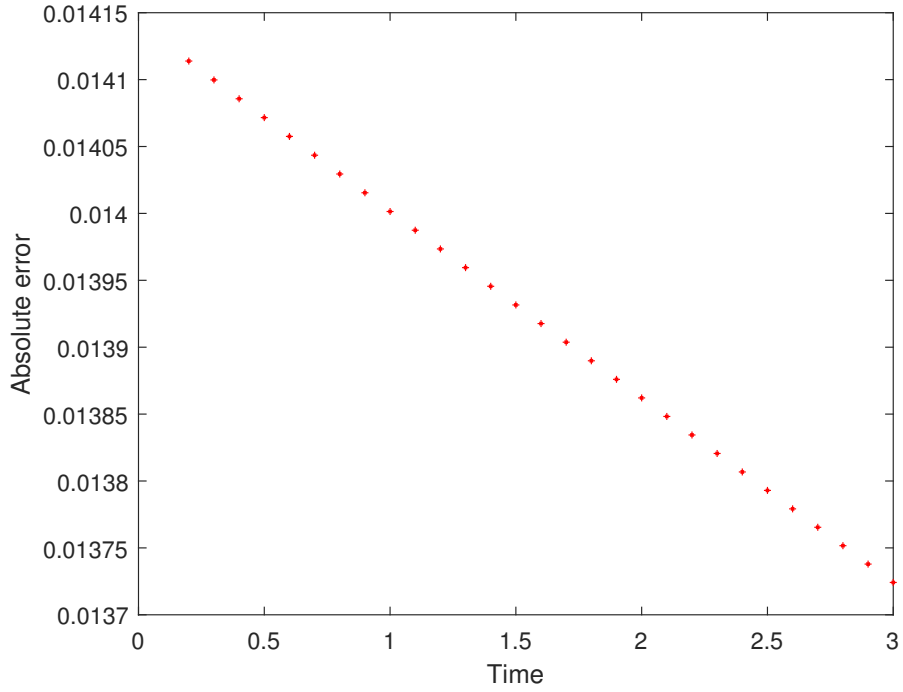

Figure 2.2: Absolute error in solution of equation.

where $\hat{u} = \frac{u}{\|u\|}$.
We rewrite the above expression as

$$\delta(t) = \kappa(B, u, \hat{z}_0)\varepsilon, \tag{2.0.9}$$

where

$$\kappa(B, u, \hat{z}_0) = \frac{\|B\hat{z}_0\|}{\|B\hat{u}\|}$$

is called the condition number with direction of perturbation of the problem (2.0.6). It is not always the case that we have information on the specific direction of perturbation. In this scenario, we can specify the relation between δ and ε in the following manner

$$\delta \leq \kappa(B, u)\varepsilon, \tag{2.0.10}$$

where

$$\kappa(A, u) = \max_{\|\hat{z}\|=1} \kappa(B, u, \hat{z}_0) = \frac{\|B\|}{\|B\hat{u}\|}$$

is called the condition number of the problem (2.0.6) (see[11] for the definition of the condition number of a general problem). Note that the equality in (2.0.10) holds if the specific direction of perturbation $\hat{z}_0$ satisfies

$$\|B\| = \max_{\|\hat{z}_0\|=1} \|B\hat{z}_0\| = \|B\hat{z}_0\|. \tag{2.0.11}$$

Moreover, we also have

$$\delta(t) \leq \kappa(B)\varepsilon, \tag{2.0.12}$$

where

$$\kappa(B) = \max_{\|u\|=1} \kappa(B, u) = \|B\| \left\|B^{-1}\right\|.$$

is called the condition number independent of the data of the problem (2.0.6). We can see that the equality in (2.0.12) holds if the data u satisfies $u = bx$, where $x \in \mathbb{R}^n$ is such that

$$\left\|B^{-1}\right\| = \frac{\left\|B^{-1}x\right\|}{\|x\|}$$

and the direction of perturbation $\hat{z}_0$ satisfies (2.0.11).

Note that the problem (2.0.1) is the problem (2.0.6) with $B = e^{tA}$. Hence the three condition numbers take the following form:

- $$K(t, A, y_0, \hat{z}_0) = \kappa(e^{tA}, y_0, \hat{z}) = \frac{\left\|e^{tA}\hat{z}_0\right\|}{\|e^{At}\hat{y}_0\|} \tag{2.0.13}$$

  is the condition number with direction of perturbation of the problem (2.0.1).

- $$K(t, A, y_0) = \kappa(e^{tA}, y_0) = \frac{\|e^{tA}\|}{\|e^{tA}\hat{y}_0\|} \tag{2.0.14}$$

  is the condition number the problem (2.0.1).

- $$K(t, A) = \kappa(e^{tA}) = \|e^{tA}\| \, \|e^{-tA}\| \tag{2.0.15}$$

  is the condition number independent of the data of the problem (2.0.1).

Now our purpose is to analyze these condition numbers for a normal matrix. Before starting the analysis it is useful to know some basic definitions and properties of a normal matrix.

2.1 Normal matrices and their properties

There are several equivalent ways to define a normal matrix. A few comprehensive surveys [22, 24, 34] on normal matrices show a large number of equivalent definitions. For example, [34] gives a list of 70 conditions. Each condition is equivalent to the basic definition (given below). Later [24] compiled a list of 20 more conditions. The best known definition of a normal matrix is the following one.

Definition A matrix $A \in \mathbb{C}^{n\times n}$ is normal if $A^*A = AA^*$, where A^* is the conjugate transpose of the matrix A. For A real, A is normal if $A^TA = AA^T$, where A^T is the transpose of A.

- A normal matrix has all eigenvalues non-defective (an eigenvalue is non-defective if the algebraic and geometric multiplicities are the same), i.e the matrix is diagonalizable.
- Moreover, a normal matrix is unitary diagonalizable i.e there exist orthonormal basis of $\mathbb{C}^n$ of eigenvectors of the matrix.

For a worth reading of the topic see [28, 34, 43, 72].
The class of the real normal matrices contains sub-classes of many important matrices like the orthogonal matrices, symmetric matrices and shifted skew-symmetric matrices. Remind that a shifted skew-symmetric matrix is a matrix A of the form

$$A = B + cI_n$$

where $c \in \mathbb{R}$ and B is skew-symmetric i.e $B^T = -B$. All eigenvalues of a shifted skew-symmetric matrix have same real parts.

Let $A \in \mathbb{R}^{n\times n}$ be a diagonalizable matrix. The matrix A, whose distinct eigenvalues are $\lambda_1, \lambda_2, \ldots, \lambda_p$, can be written as

$$A = \sum_{i=1}^{p} \lambda_i P_i \tag{2.1.1}$$

where $P_i = u_i u_i^*$, $i = 1, \ldots, p$, is the orthogonal projection on the eigenspace of the eigenvalue λ_i, u_i being the eigenvector relevant to the eigenvalue λ_i.
More generally, for any complex analytic function f, we have

$$f(A) = \sum_{i=1}^{p} f(\lambda_i) P_i. \tag{2.1.2}$$

The next proposition gives the 2-norm of $f(A)u$, where $u \in \mathbb{R}^n$, in the case where A is, in addition a normal matrix.

Proposition 2.1.1. *We have*

$$\|f(A)u\|_2^2 = \sum_{i=1}^{p} |f(\lambda_i)|^2 . \|P_i u\|_2^2 , \quad u \in \mathbb{R}^n. \tag{2.1.3}$$

Proof. For $v \in \mathbb{C}^n$, we have

$$\|v\|_2^2 = \langle v, v\rangle = v^* v,$$

where $\langle \cdot, \cdot \rangle$ is the scalar product on $\mathbb{C}^n$.
Thus

$$\|f(A)u)\|_2^2 = \langle \sum_{i=1}^{p} f(\lambda_i) P_i u, \sum_{k=1}^{p} f(\lambda_k) P_k u\rangle = \sum_{i=1}^{p}\sum_{k=1}^{p} \overline{f(\lambda_i)} f(\lambda_k) \langle P_i u, P_k u\rangle$$

Hence, by taking advantage of the fact that $P_i u$ and $P_k u$ with $i \neq k$, $i, k = 1, \ldots, p$, are orthogonal when A is a normal matrix, we obtain

$$\|f(A)u)\|_2^2 = \sum_{i=1}^{p} \overline{f(\lambda_i)} f(\lambda_i) \langle P_i u, P_i u\rangle = \sum_{i=1}^{p} |f(\lambda_i)|^2 \|P_i u\|_2^2 .$$

□

2.2 Condition numbers for a normal matrix

Up to end of this chapter, we consider the matrix A as normal. We denote the spectrum of the matrix A by Λ

$$\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_p\},$$

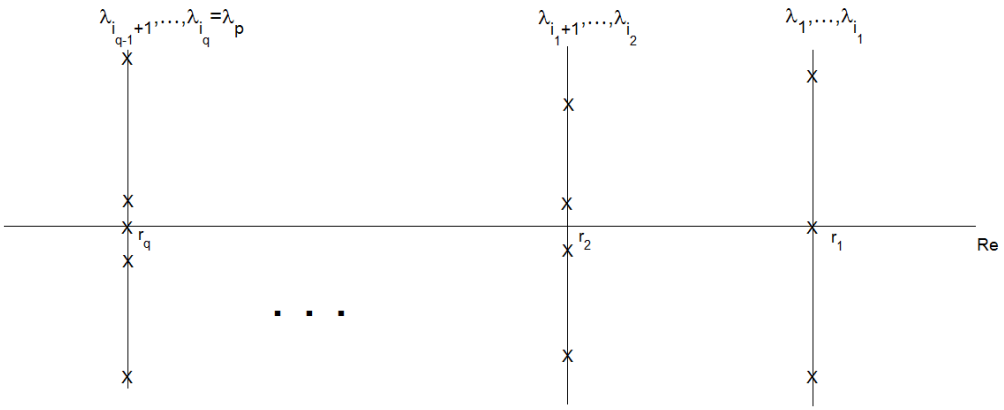

Figure 2.3: Spectrum of A partitioned by decreasing real parts.

where $\lambda_1, \ldots, \lambda_p$ are the distinct eigenvalues.
We partitioned the spectrum Λ by decreasing real parts, into subsets $\Lambda_1, \ldots \Lambda_q$,

$$\Lambda_j := \{\lambda_{i_{j-1}+1}, \lambda_{i_{j-1}+2}, \ldots, \lambda_{i_j}\}, \quad j = 1, \ldots, q,$$

where $0 = i_0 < i_1 < \cdots < i_q = p$, we have

$$\mathrm{Re}(\lambda_{i_{j-1}+1}) = \mathrm{Re}(\lambda_{i_{j-1}+2}) = \cdots = \mathrm{Re}(\lambda_{i_j}) = r_j, \ j = 1, \ldots, q,$$

with

$$r_1 > r_2 > \cdots > r_q.$$

See the Figure 3.

For $i = 1, \ldots, p$, as in the previous subsection, P_i denotes the projection on the eigenspace of the eigenvalue λ_i. Observe that P_i is an orthogonal projection.
Moreover, for $j = 1, \ldots, q$, let

$$Q_j = \sum_{\lambda_i \in \Lambda_j} P_i,$$

be the orthogonal projection on the sum of the eigenspces of eigenvalues in Λ_j.
Observe that, for $u \in \mathbb{R}^n$,

$$\|Q_j u\|_2^2 = \sum_{\lambda_i \in \Lambda_j} \|P_i u\|_2^2, \quad j = 1, \ldots, q \tag{2.2.1}$$

and

$$\sum_{j=1}^{q} \|Q_j u\|_2^2 = \|u\|_2^2. \tag{2.2.2}$$

The next theorem gives the condition numbers $K(t, A, y_0, \hat{z}_0), K(t, A, y_0)$ and $K(t, A)$ for the 2−norm. Since we are using 2-norm, we denote the condition numbers as $K_2(.)$ instead of $K(.)$.

Theorem 2.2.1. *Suppose that the initial value $y_0 \in \mathbb{R}^n$ of the ODE*

$$\begin{cases} y'(t) = Ay(t), \ t \geq 0, \\ y(0) = y_0, \end{cases} \tag{2.2.3}$$

is not zero and it is perturbed to $\tilde{y}_0$ with the direction of perturbation $\hat{z}_0$ and relative error

$$\epsilon = \frac{\|\tilde{y}_0 - y_0\|_2}{\|y_0\|_2}$$

The three condition numbers of the problem (2.0.1) *are given by*

$$K_2(t, A, y_0, \hat{z}_0) = \frac{\sqrt{\|Q_1\hat{z}_0\|_2^2 + \sum_{j=2}^{q} e^{2(r_j - r_1)t} \|Q_j\hat{z}_0\|_2^2}}{\sqrt{\|Q_1\hat{y}_0\|_2^2 + \sum_{j=2}^{q} e^{2(r_j - r_1)t} \|Q_j\hat{y}_0\|_2^2}}, \tag{2.2.4}$$

$$K_2(t, A, y_0) = \frac{1}{\sqrt{\|Q_1\hat{y}_0\|_2^2 + \sum_{j=2}^{q} e^{2(r_j - r_1)t} \|Q_j\hat{y}_0\|_2^2}}, \tag{2.2.5}$$

$$K_2(t, A) = e^{(r_1 - r_q)t}. \tag{2.2.6}$$

Moreover,

- *If $\hat{z}_0$ lies in the sum of eigenspaces of eigenvalues in Λ_1, then equality between condition numbers* (2.2.4) *and* (2.2.5) *holds.*
- *If y_0 lies in sum of eigenspaces of eigenvalues in Λ_q, then equality between condition numbers* (2.2.5) *and* (2.2.6) *holds. In addition, if $\hat{z}_0$ lies in the sum of eigenspaces of eigenvalues in λ_1, we have equality between all three condition numbers.*

Proof. In case of the matrix exponential function, the equation (2.1.3) gives

$$\left\|e^{tA}u\right\|_2^2 = \sum_{i=1}^{p} |e^{\lambda_i t}|^2 . \|P_i u\|_2^2 = \sum_{i=1}^{p} e^{2Re(\lambda_i)t} \|P_i u\|_2^2 .$$

By making use of the equation (2.2.1), we get

$$\left\|e^{tA}u\right\|_2^2 = \sum_{j=1}^{q} e^{2r_j t} \sum_{\lambda_i \in \Lambda_j} \|P_i u\|_2^2 = \sum_{j=1}^{q} e^{2r_j t} \|Q_j u\|_2^2 .$$

Substituting the above expression in the equation (2.0.13), we get

$$K_2(t, A, y_0, \hat{z}_0)^2 = \frac{\left\|e^{tA}\hat{z}_0\right\|^2}{\left\|e^{tA}\hat{y}_0\right\|^2} = \frac{\sum_{j=1}^{q} e^{2r_j t} \|Q_j\hat{z}_0\|_2^2}{\sum_{j=1}^{q} e^{2r_j t} \|Q_j\hat{y}_0\|_2^2}$$

or

$$K_2(t,A,y_0,\hat{z}_0)^2 = \frac{\sum_{j=1}^{q} e^{2(r_j-r_1)t}\,\|Q_j\hat{z}_0\|_2^2}{\sum_{j=1}^{q} e^{2(r_j-r_1)t}\,\|Q_j\hat{y}_0\|_2^2} = \frac{\|Q_1\hat{z}_0\|_2^2 + \sum_{j=2}^{q} e^{2(r_j-r_1)t}\,\|Q_j\hat{z}_0\|_2^2}{\|Q_1\hat{y}_0\|_2^2 + \sum_{j=2}^{q} e^{2(r_j-r_1)t}\,\|Q_j\hat{y}_0\|_2^2}.$$

Now, we give an expression for $K(t,A,y_0)_2$. We have, for any direction of perturbation $\hat{z}_0$,

$$\|Q_1\hat{z}_0\|_2^2 + \sum_{j=2}^{q} e^{2(r_j-r_1)t}\,\|Q_j\hat{z}_0\|_2^2 \le \sum_{j=1}^{q} \|Q_j\hat{z}_0\|_2^2 = 1.$$

by (2.2.2). Hence

$$K_2(t,A,y_0)^2 \le \frac{1}{\|Q_1\hat{y}_0\|_2^2 + \sum_{j=2}^{q} e^{2(r_j-r_1)t}\,\|Q_j\hat{y}_0\|_2^2}.$$

In addition, if $\hat{z}_0$ lies in the sum of eigenspaces of eigenvalues in Λ_1, i.e $Q_j\hat{z}_0 = 0$ for all $j = 2,\ldots,q$, we have

$$\|Q_1\hat{z}_0\|_2^2 + \sum_{j=2}^{q} e^{2(r_j-r_1)t}\,\|Q_j\hat{z}_0\|_2^2 = \|Q_1\hat{z}_0\|_2^2 = 1.$$

So, we get

$$K_2(t,A,y_0)^2 = \frac{1}{\|Q_1\hat{y}_0\|_2^2 + \sum_{j=2}^{q} e^{2(r_j-r_1)t}\,\|Q_j\hat{y}_0\|_2^2}.$$

Now, to give the expression for $K_2(t,A)$, observe that, for any initial value y_0,

$$\|Q_1\hat{y}_0\|_2^2 + \sum_{j=2}^{q} e^{2(r_j-r_1)t}\,\|Q_j\hat{y}_0\|_2^2 \ge e^{2(r_q-r_1)t}\sum_{j=1}^{q} \|Q_j\hat{y}_0\|_2^2 = e^{2(r_q-r_1)t}.$$

Hence,

$$K_2(t,A)^2 \le e^{2(r_q-r_1)t}.$$

In addition, if y_0 lies in the sum of eigenspaces of eigenvalues in Λ_q. i.e $Q_j\hat{y}_0 = 0$ for all $j = 1,\ldots,q-1$, we have

$$\|Q_1\hat{y}_0\|_2^2 + \sum_{j=2}^{q} e^{2(r_j-r_1)t}\,\|Q_j\hat{y}_0\|_2^2 = e^{2(r_q-r_1)t}\sum_{j=1}^{q} \|Q_j\hat{y}_0\|_2^2 = e^{2(r_q-r_1)t}.$$

So, we get

$$K_2(t,A)^2 = e^{2(r_q-r_1)t}.$$

□

2.3 Asymptotic behavior of condition numbers

We begin the asymptotic analysis by considering the condition number (2.2.5). Observe that the condition numbers is an increasing function of t and $K_2(0, A, y_0) = 1$. Its asymptotic behavior is given by

$$K_2(t, A, y_0) = \frac{1}{\sqrt{\|Q_1\hat{y}_0\|_2^2 + \sum_{j=2}^{q} e^{2(r_j - r_1)t} \|Q_j\hat{y}_0\|_2^2}} \sim \frac{e^{(r_1 - r_{j^*})t}}{\|Q_{j^*}\hat{y}_0\|_2}, \quad t \to +\infty,$$

where j^* is the minimum index $j = 1, \ldots, q$ such that $Q_j\hat{y}_0 \neq 0$ and the notation

$$a(t) \sim b(t), \quad t \to +\infty,$$

means

$$\lim_{t \to +\infty} \frac{a(t)}{b(t)} = 1.$$

We can observe an exponential increase of the condition number in case of $j^* > 1$.
Now, we give the asymptotic behavior of the condition number (2.2.4). We rewrite $K_2(t, A, y_0, \hat{z}_0)$ in the following manner

$$K_2(t, A, y_0, \hat{z}_0) = K_2(t, A, y_0).R_2(t, A, \hat{z}_0),$$

where

$$R(t, A, \hat{z}_0) = \sqrt{\|Q_1\hat{z}_0\|_2^2 + \sum_{j=2}^{q} e^{2(r_j - r_1)t} \|Q_j\hat{z}_0\|_2^2}.$$

Observe that $R(t, A, \hat{z}_0)$ is a decreasing function of t with $R(0, A, \hat{z}_0) = 1$. The asymptotic behavior of $R(t, A, \hat{z}_0)$ is given by

$$R(t, A, \hat{z}_0) \sim e^{(r_{j^{**}} - r_1)t} \|Q_{j^{**}}\hat{z}_0\|_2, \quad t \to +\infty,$$

where, j^{**} is the minimum index $j = 1, \ldots, q$, such that $Q_{j^{**}}\hat{z}_0 \neq 0$. Ultimately, the asymptotic behavior of $K_2(t, A, y_0, \hat{z}_0)$ is given by

$$K_2(t, A, y_0, \hat{z}_0) \sim e^{(r_{j^{**}} - r_{j^*})t} \frac{\|Q_{j^{**}}\hat{z}_0\|_2}{\|Q_{j^*}\hat{y}_0\|_2}, \quad t \to \infty.$$

We can observe an exponential growth of the condition number in case $j^{**} < j^*$ and an exponential decay in case $j^{**} > j^*$.

2.4 The case $q = 1$

Observe that

$$K_2(t, A, \hat{y}_0, \hat{z}_0) = 1,$$

when $q = 1$, in other words, when

$$Re(\lambda_1) = Re(\lambda_2) = \cdots = Re(\lambda_p).$$

In this case, we have $\delta(t) = \varepsilon$ for all $t \geq 0$. Such a scenario occurs in case of skew-symmetric matrices. This is because of the fact that skew-symmetric matrices have pure imaginary eigenvalues. In such a case, all the condition numbers are exactly equal to 1:

$$K_2(t, A, y_0, \hat{z}) = K_2(t, A, y_0) = K_2(t, A) = 1.$$

It is quite straight forward that $\delta(t) = \varepsilon$ for all $t \geq 0$, holds for a skew-symmetric matrix A. In fact, in this case e^{tA} is an orthogonal matrix and then

$$\delta(t) = \frac{\left\| e^{tA} \left(\tilde{y}_0 - y_0 \right) \right\|_2}{\| e^{tA} y_0 \|_2} = \frac{\| \tilde{y}_0 - y_0 \|_2}{\| y_0 \|_2} = \varepsilon.$$

Chapter 3

Perturbations in the matrix

After the analysis of how perturbations in the initial value propagates along the solution of linear ordinary differential equations, we are now interested to look over the same question when it is the coefficient matrix A in

$$\begin{cases} y'(t) = Ay(t),\ t \geq 0, \\ y(0) = y_0, \end{cases} \tag{3.0.1}$$

where $A \in \mathbb{R}^{n\times n}$ and $y_0 \in \mathbb{R}^n$, rather than the initial value that perturbs. This chapter is devoted to give such a perturbation analysis. In short, the interest is to study the conditioning of the problem

$$A \mapsto e^{tA}y_0. \tag{3.0.2}$$

The main feature of the chapter is the perturbation analysis of the equation (3.0.1) when the matrix A is normal. Moreover, the chapter deals with the asymptotic behavior of this perturbation. Suppose that the matrix $A \neq 0$ is perturbed to $\widetilde{A}$. The relative error is given by

$$\epsilon = \frac{|||\widetilde{A} - A|||}{|||A|||}, \tag{3.0.3}$$

where $|||\cdot|||$ is a generic matrix norm on $\mathbb{R}^{n\times n}$. The solution $y(t) = e^{tA}y_0$ is perturbed to $\widetilde{y}(t) = e^{t\widetilde{A}}y_0$ with relative error

$$\xi(t) = \frac{\left\|e^{t\widetilde{A}}y_0 - e^{tA}y_0\right\|}{\|e^{tA}y_0\|}, \tag{3.0.4}$$

where $\|\cdot\|$ is a generic vector norm. Observe that $\xi(t)$ is well defined for $y_0 \neq 0$ since in this case $e^{tA}y_0 \neq 0$. Moreover we have $\xi(0) = 0$.
Now, we want to explore the relation between ϵ and $\xi(t)$. As already mentioned in the first chapter, there are papers (see[2, 45, 51, 54, 81]) in the literature dealing with the conditioning of the problem

$$A \mapsto e^{tA}. \tag{3.0.5}$$

In other words, they deal with the effect in e^{tA}, which is a matrix quantity; of a perturbation in A. These papers do not consider the role of the initial value, as they consider the following relative error

$$\frac{\left|\left|\left| e^{t\tilde{A}} - e^{tA} \right|\right|\right|}{|||e^{tA}|||},$$

rather then $\xi(t)$. On the other hand, our focus is to study the conditioning of the problem (3.0.2), that takes into account the role of the initial value as well. The conditioning of the problem (3.0.2), attained less attention in the literature. To the best of our knowledge there are two papers [1, 19] dealing with the problem (3.0.2), but the focus of these papers is on computational aspects rather than a qualitative analysis. For example, the paper [1], in order to analyze an algorithm for computing $e^{tA}Y_0$, where Y_0 is a matrix, considered the conditioning of the problem $(A, Y_0) \mapsto f(tA)Y_0$ (relevant to Frobenius norms), where f is matrix function, and obtained a bound for it. The purpose of the paper [19], is to develop algorithms for studying the conditioning of the problem $(t, A, y_0) \mapsto f(t, A)y_0$. In the present thesis, we are going to analyze the conditioning of the problem (3.0.2) and to study how it depends upon time t and the initial value y_0. Similar to the case of perturbation in the initial value, we give three condition numbers.

The chapter is structured in following manner and it is substantially the contents of the paper [27]. In section 3.1, we begin our analysis with the introduction of the condition numbers of the problem (3.0.2), for a general matrix and for general vector and matrix norms $\|\cdot\|$ and $|||\cdot|||$. These condition numbers are given in terms of the Frechét derivative. Hence, to give an introduction of Frechét derivative to the readers, a brief note on Frechét derivative appears at first. The section 3.2 gives the analysis of the condition numbers for a normal matrix and for $\|\cdot\| = \|\cdot\|_2$ and $|||\cdot||| = \|\cdot\|_2$ (i.e $|||\cdot|||$ is the spectral matrix norm, namely the matrix norm induced by the $2-$vector norm). The asymptotic behavior of the condition numbers is given in section 3.3. To testify our analysis we give a few numerical tests in section 3.4. The conclusion of the chapter is given in section 3.5.

3.0.1 The Fréchet Derivative

As we have seen in (1.1.2), the Fréchet derivative plays an important role in the definition of condition numbers (see [3, 4, 38, 41, 46, 63]). The idea of Frechét derivative is to extend the notion of derivative of a real valued function of a single variable to Banach spaces (see [7, 33, 55]). The definition of the Frechét derivative is the following one.
Definition Let U and V be two Banach spaces. The Fréchet derivative of $f : U \mapsto V$ at $X \in U$ is defined as a linear map $L(X, \cdot) : U \mapsto V$ such that

$$f(X + E) - f(X) - L(X, E) = o(\|E\|), \quad E \in U \text{ and } \|E\| \to 0.$$

As an easy example of Fréchet derivative, we consider $U = V = \mathbb{R}^{n\times n}$ and $f(X) = X^2$. We have

$$f(X + E) - f(X) = E^2 + XE + EX$$

and we get $L(X, E) = XE+EX$. Note that it is not always straightforward to determine the Fréchet derivative.
In the case where f is the exponential function, i.e $f(X) = e^{tX}$, $X \in U = V = \mathbb{R}^{n\times n}$, the Fréchet derivative is given by

$$L(t, X, E) = \int_0^t e^{(t-s)X} E e^{sX} ds. \tag{3.0.6}$$

as we have seen in (1.3.3). Here is the proof of it.

Proof. Recall that $y(t) = e^{tA}$, if and only if $y(t)$ satisfies the initial value problem (3.0.1). Now suppose that

$$X(t) = e^{tA} + \int_0^t e^{(t-s)A} E e^{s(A+E)} ds \tag{3.0.7}$$

Differentiation equation (3.0.7) reveals

$$X'(t) = (A + E)X(t).$$

Since $X(0) = I$, so

$$X(t) = e^{t(A+E)} = e^{tA} + \int_0^t e^{(t-s)A} B e^{s(A+E)} ds.$$

By using the above expression inside the integral gives

$$e^{t(A+E)} = e^{tA} + \int_0^t e^{A(t-s)} E e^{sA} ds + o(\|E\|^2).$$

Hence, by the definition of the Fréchet derivative we get

$$L(t, X, E) = \int_0^t e^{(t-s)X} E e^{sX} ds.$$

□

Since the Fréchet derivative (3.0.6) is a linear operator on $\mathbb{R}^{n\times n}$, we have the Kronecker form of this Fréchet derivative.

$$vec(L(t, X, E)) = M(t, A)vec(E) \tag{3.0.8}$$

for some matrix $M(t, A) \in \mathbb{R}^{n^2\times n^2}$, where $vec(C)$, where $C \in \mathbb{R}^{n\times n}$, is the the vector of $\mathbb{R}^{n^2}$ obtained by stacking the columns of C, starting from the first column to the last. For more details on the Fréchet derivative see ([39, 41, 63]).

3.0.2 Condition numbers

We specify the perturbed matrix in (3.0.1) as

$$\widetilde{A} = A + \epsilon |||A||| \widehat{B}, \tag{3.0.9}$$

where the matrix

$$\widehat{B} = \frac{\widetilde{A} - A}{|||A|||}$$

is the *direction of the perturbation* and $\left|\left|\left| \widehat{B} \right|\right|\right| = 1$ holds.

We define

$$\overline{K}(t, A, y_0, \widehat{B}) := \lim_{\epsilon \to 0} \frac{\xi(t)}{\epsilon}, \tag{3.0.10}$$

where $\xi(t)$ and ϵ are given in (3.0.4) and (3.0.3) respectively, as the *condition number with direction of perturbation* of the problem (3.0.2).

The next Theorem gives an expression for such a condition number.

Theorem 3.0.1. *We have*

$$\overline{K}\left(t, A, y_0, \widehat{B}\right) = \frac{\left\| L\left(t, A, \widehat{B}\right) \widehat{y}_0 \right\| |||A|||}{\|e^{tA} \widehat{y}_0\|}, \tag{3.0.11}$$

where

$$L\left(t, A, \widehat{B}\right) = \int_0^t e^{(t-s)A} \widehat{B} e^{sA} ds$$

is the Fréchet derivative (3.0.6) *and*

$$\widehat{y}_0 := \frac{y_0}{\|y_0\|}.$$

Proof. We have

$$e^{t\widetilde{A}} y_0 - e^{tA} y_0 = \left(e^{t(A+E)} - e^{tA}\right) y_0,$$

where

$$E = \epsilon |||A||| \widehat{B}.$$

Since (see [41])

$$e^{t(A+E)} - e^{tA} = L(t, A, E) + O(|||E|||^2), \ E \to 0,$$

where

$$E \mapsto L(t, A, E) = \int_0^t e^{(t-s)A} E e^{sA} ds$$

is the Frechét derivative of the map $A \mapsto e^{tA}$, we obtain

$$\xi(t) = \frac{\left\| L\left(t, A, \widehat{B}\right) y_0 \right\| \, |||A|||}{\|e^{tA} y_0\|} \epsilon + O\left(\epsilon^2\right), \ \varepsilon \to 0,$$

and (3.0.11) follows. □

We define

$$\overline{K}(t, A, y_0) := \sup_{\substack{\widehat{B} \in \mathbb{R}^{n \times n} \\ |||\widehat{B}|||=1}} \overline{K}\left(t, A, y_0, \widehat{B}\right) \tag{3.0.12}$$

as the *condition number* of the problem (3.0.2). We have

$$\overline{K}(t, A, y_0) = \frac{\|\mathcal{L}(t, A, y_0)\|}{\|e^{tA} y_0\|}, \tag{3.0.13}$$

where $\mathcal{L}(t, A, y_0) : \mathbb{R}^{n \times n} \mapsto \mathbb{R}^{n \times n}$ is the linear operator given by

$$\mathcal{L}(t, A, y_0)\widehat{B} = L(t, A, \widehat{B}), \quad B \in \mathbb{R}^{n \times n}$$

and $\|\mathcal{L}(t, A, y_0)\|$ is the operator norm relevant to the norms $\|\cdot\|$ and $\mathbb{R}^n$ and $|||\cdot|||$ on $\mathbb{R}^{n \times n}$. By using the Kronecker form (3.0.8) we get

$$L(t, A, \widehat{B}) y_0 = (y_0^T \otimes I_n) vec(L(t, A, \widehat{B})) = (y_0^T \otimes I_n) M(t, A) \mathrm{vec}(\widehat{B}),$$

where $\otimes$ is the Kronecker product. So, we get

$$\|\mathcal{L}(t, A, y_0)\| = \|(y_0^T \otimes I_n) M(t, A)\|_2 \ \text{ if } \|\cdot\| = \|\cdot\|_2 \text{ and } |||\cdot||| = \|\cdot\|_F$$

and

$$\begin{aligned} &\frac{1}{\sqrt{n}} \|(y_0^T \otimes I_n) M(t, A)\|_2 \le \|\mathcal{L}(t, A, y_0)\| \le \sqrt{n} \|(y_0^T \otimes I_n) M(t, A)\|_2 \\ &\|(y_0^T \otimes I_n) M(t, A)\|_1 \le \|\mathcal{L}(t, A, y_0)\| \le n \|(y_0^T \otimes I_n) M(t, A)\|_1 \\ &\quad \text{if } \|\cdot\| = \|\cdot\|_1 \text{ and } |||\cdot||| = \|\cdot\|_1. \end{aligned} \tag{3.0.14}$$

The condition number (3.0.13) corresponds to the standard definition of condition number of a general problem (see [11]) and it is the same condition number considered in the papers [1] and [19]. The paper [19] used (3.0.14) for estimating the condition number.

Finally, we define

$$\overline{K}(t, A) := \sup_{\substack{y_0 \in \mathbb{R}^n \\ y_0 \neq 0}} \overline{K}(t, A, y_0) \tag{3.0.15}$$

as the *condition number independent of the data* of the problem (3.0.2).

3.1 Analysis for A normal

From now on, we consider A normal, $\|\cdot\| = \|\cdot\|_2$ and $|||\cdot||| = \|\cdot\|_2$. We write the condition numbers $\overline{K}(\cdot)$ defined in the previous section as $\overline{K}_2(\cdot)$.

We partition the spectrum $\Lambda := \{\lambda_1, \ldots \lambda_p\}$ of A (see Figure 1) in the subsets Λ_j, $j = 1, \ldots, q$, in the same manner as in chapter 2.
Recall that in chapter 2 we have introduced, for $i = 1, \ldots, p$, the orthogonal projection P_i on the eigenspace of the eigenvalue λ_i and, for $j = 1, \ldots, q$, the orthogonal projection

$$Q_j = \sum_{\lambda_i \in \Lambda_j} P_i.$$

on the sum of the eigenspaces of eigenvalues in Λ_j.

3.1.1 The condition number $\overline{K}_2\left(t, A, y_0, \widehat{B}\right)$ with direction of perturbation

The next theorem provides an expression for $\overline{K}_2\left(t, A, y_0, \widehat{B}\right)$.

Theorem 3.1.1. *We have*

$$\overline{K}_2\left(t, A, y_0, \widehat{B}\right) = \frac{\sqrt{\sum_{j=1}^{q}\left(e^{(r_j - r_1)t}\left\|Q_j\left(\int_0^t e^{-sA}\widehat{B}e^{sA}ds\right)\widehat{y}_0\right\|_2\right)^2}}{\sqrt{\sum_{j=1}^{q}\left(e^{(r_j-r_1)t}\left\|Q_j\widehat{y}_0\right\|_2\right)^2}}\|A\|_2 \tag{3.1.1}$$

and for the numerator in (3.1.1) we have

$$\begin{aligned}&\sqrt{\sum_{j=1}^{q}\left(e^{(r_j - r_1)t}\left\|Q_j\left(\int_0^t e^{-sA}\widehat{B}e^{sA}ds\right)\widehat{y}_0\right\|_2\right)^2}\\ &= \left\|\sum_{\lambda_i\in\Lambda}\sum_{\lambda_k\in\Lambda} C(t,\lambda_i,\lambda_k)P_i\widehat{B}P_k\widehat{y}_0\right\|_2,\end{aligned} \tag{3.1.2}$$

where, for $\lambda_i \in \Lambda_j$, with $j \in \{1, \ldots, q\}$, and $\lambda_k \in \Lambda$,

$$C(t, \lambda_i, \lambda_k) := e^{(r_j - r_1)t}\int_0^t e^{(\lambda_k - \lambda_i)s}ds. \tag{3.1.3}$$

Proof. By recalling (3.0.11), we write

$$\overline{K}_2\left(t, A, y_0, \widehat{B}\right) = \frac{\left\| e^{tA} \int_0^t e^{-sA} \widehat{B} e^{sA} ds \widehat{y}_0 \right\|_2}{\|e^{tA}\widehat{y}_0\|_2} \|A\|_2 .$$

Since A is normal, we have, for $u \in \mathbb{R}^n$,

$$\|e^{tA}u\|_2 = \left\| \sum_{\lambda_i \in \Lambda} e^{t\lambda_i} P_i u \right\|_2 = \sqrt{\sum_{\lambda_i \in \Lambda} \left(|e^{t\lambda_i}| \, \|P_i u\|_2 \right)^2} = \sqrt{\sum_{j=1}^{q} \left(e^{r_j t} \, \|Q_j u\|_2 \right)^2} .$$

Thus

$$\left\| e^{tA} \int_0^t e^{-sA} \widehat{B} e^{sA} ds \widehat{y}_0 \right\|_2 = \sqrt{\sum_{j=1}^{q} \left(e^{r_j t} \left\| Q_j \left(\int_0^t e^{-sA} \widehat{B} e^{sA} ds \right) \widehat{y}_0 \right\|_2 \right)^2}$$

and

$$\left\| e^{tA} \widehat{y}_0 \right\|_2 = \sqrt{\sum_{j=1}^{q} \left(e^{r_j t} \, \|Q_j \widehat{y}_0\|_2 \right)^2}$$

Hence,

$$\overline{K}_2(t, A, y_0, \hat{B}) = \frac{\sqrt{\sum_{j=1}^{q} \left(e^{(r_j - r_1)t} \left\| Q_j \left(\int_0^t e^{-sA} \widehat{B} e^{sA} ds \right) \widehat{y}_0 \right\|_2 \right)^2}}{\sqrt{\sum_{j=1}^{q} \left(e^{(r_j - r_1)t} \, \|Q_j \widehat{y}_0\|_2 \right)^2}} \|A\|_2 .$$

By the orthogonality of the projections Q_j, $j = 1, \ldots, q$, we get

$$\begin{aligned} &\sqrt{\sum_{j=1}^{q} \left(e^{(r_j - r_1)t} \left\| Q_j \int_0^t e^{-sA} \widehat{B} e^{sA} ds \widehat{y}_0 \right\|_2 \right)^2} \\ &= \left\| \sum_{j=1}^{q} e^{(r_j - r_1)t} Q_j \int_0^t e^{-sA} \widehat{B} e^{sA} ds \widehat{y}_0 \right\|_2 . \end{aligned}$$

Now, by decomposing the matrices e^{-sA} and e^{sA} as

$$e^{-sA} = \sum_{i=1}^{p} e^{-\lambda_i s} P_i \quad \text{and} \quad e^{sA} = \sum_{k=1}^{p} e^{\lambda_k s} P_k, \tag{3.1.4}$$

We obtain (3.1.2). □

The next proposition concerns the functions $C(t, \lambda_i, \lambda_k)$ defined in (3.1.3).

Proposition 3.1.1. *Let $j, l \in \{1, \ldots, q\}$, let $\lambda_i \in \Lambda_j$ and let $\lambda_k \in \Lambda_l$. Moreover, let*

$$\lambda_i = r_j + \sqrt{-1}\omega_i \quad \text{and} \quad \lambda_k = r_l + \sqrt{-1}\omega_k$$

be the cartesian forms of the complex numbers λ_i and λ_k, where $\sqrt{-1}$ denotes the imaginary unit.

If $j \leq l$, then

$$|C(t, \lambda_i, \lambda_k)| \leq e^{(r_j - r_1)t} t.$$

If $j \geq l$, then

$$|C(t, \lambda_i, \lambda_k)| \leq e^{(r_l - r_1)t} t.$$

If $\lambda_i \neq \lambda_k$, then

$$C(t, \lambda_i, \lambda_k) = \frac{e^{(r_l - r_1)t} e^{\sqrt{-1}(\omega_k - \omega_i)t} - e^{(r_j - r_1)t}}{\lambda_k - \lambda_i}.$$

If $\lambda_i = \lambda_k$, then

$$C(t, \lambda_i, \lambda_k) = e^{(r_j - r_1)t} t.$$

Proof. We have

$$\left| \int_0^t e^{(\lambda_k - \lambda_i)s} ds \right| \leq \int_0^t e^{(r_l - r_j)s} ds$$

and then

$$|C(t, \lambda_i, \lambda_k)| \leq e^{(r_j - r_1)t} \int_0^t e^{(r_l - r_j)s} ds \leq e^{(r_j - r_1)t} t$$

for $j \leq l$ and

$$|C(t, \lambda_i, \lambda_k)| \leq e^{(r_j - r_1)t} \int_0^t e^{(r_l - r_j)s} ds \leq e^{(r_j - r_1)t} e^{(r_l - r_j)t} t \leq e^{(r_l - r_1)t} t.$$

for $j \geq l$.

If $\lambda_i \neq \lambda_k$, we have

$$\int_0^t e^{(\lambda_k - \lambda_i)s} ds = \frac{e^{(\lambda_k - \lambda_i)t} - 1}{\lambda_k - \lambda_i}$$

and then

$$\begin{aligned} C(t, \lambda_i, \lambda_k) &= e^{(r_j - r_1)t} \frac{e^{(\lambda_k - \lambda_i)t} - 1}{\lambda_k - \lambda_i} \\ &= e^{(r_j - r_1)t} \frac{e^{(r_l - r_j)t} e^{\sqrt{-1}(\omega_k - \omega_i)t} - 1}{\lambda_k - \lambda_i} \\ &= \frac{e^{(r_l - r_1)t} e^{\sqrt{-1}(\omega_k - \omega_i)t} - e^{(r_j - r_1)t}}{\lambda_k - \lambda_i}. \end{aligned}$$

If $\lambda_i = \lambda_k$, we have

$$\int_0^t e^{(\lambda_k - \lambda_i)s} ds = t$$

and then

$$C(t, \lambda_i, \lambda_k) = e^{(r_j - r_1)t} t.$$

□

Remark 3.1.1. *Let $j, l \in \{1, \ldots, q\}$, let $\lambda_i \in \Lambda_j$ and let $\lambda_k \in \Lambda_l$. The previous proposition shows that:*

- *if $j > 1$ and $l > 1$, then $C(t, \lambda_i, \lambda_k)$ vanishes as $t \to +\infty$;*
- *if ($j = 1$ or $l = 1$) and $\lambda_i \neq \lambda_k$, then $C(t, \lambda_i, \lambda_k)$ is a bounded function of $t \geq 0$ and it does not vanish as $t \to +\infty$;*
- *if ($j = 1$ or $l = 1$) and $\lambda_i = \lambda_k$, i.e. $j = l = 1$ and $\lambda_i = \lambda_k$, then $C(t, \lambda_i, \lambda_k) = t$.*

3.1.2 The condition number $\overline{K}_2(t, A, y_0)$

The next theorem gives lower and upper bounds for $\overline{K}_2(t, A, y_0)$.

Theorem 3.1.2. *We have the lower bounds*

$$\overline{K}_2(t, A, y_0) \geq \|A\|_2 \, t \tag{3.1.5}$$

and

$$\overline{K}_2(t, A, y_0) \geq \frac{\max\limits_{\lambda_i, \lambda_k \in \Lambda} |D(t, \lambda_i, \lambda_k)| \, \|P_k \widehat{y}_0\|_2}{\sqrt{\sum\limits_{j=1}^{q} \left(e^{(r_j - r_1)t} \, \|Q_j \widehat{y}_0\|_2 \right)^2}} \, \|A\|_2 \,, \tag{3.1.6}$$

where

$$D(t, \lambda_i, \lambda_k) := \begin{cases} C(t, \lambda_i, \lambda_k) \ \text{ if } \lambda_k \text{ is real} \\ \\ \frac{\sqrt{2}}{2} \left(C(t, \lambda_i, \lambda_k) + C(t, \lambda_i, \overline{\lambda_k}) \right) \ \text{ if } \lambda_k \text{ is not real.} \end{cases}$$

Here $\overline{\lambda_k}$ denotes the complex conjugate of λ_k.

Moreover, we have the upper bound

$$\overline{K}_2(t, A, y_0) \leq \frac{\sqrt{\sum\limits_{\lambda_i \in \Lambda} \sum\limits_{\lambda_k \in \Lambda} |C(t, \lambda_i, \lambda_k)|^2 \, \|P_k \widehat{y}_0\|^2}}{\sqrt{\sum\limits_{j=1}^{q} \left(e^{(r_j - r_1)t} \, \|Q_j \widehat{y}_0\|_2 \right)^2}} \, \|A\|_2 \,. \tag{3.1.7}$$

Proof. For the first lower bound (3.1.5), consider numerator of (3.1.1),

$$\sqrt{\sum_{j=1}^{q}\left(e^{(r_j-r_1)t}\left\|Q_j\left(\int_0^t e^{-sA}\widehat{B}e^{sA}ds\right)\widehat{y}_0\right\|_2\right)^2}$$

substitute $\widehat{B}=I$ and making use of (3.1.6), we get

$$\sqrt{\sum_{j=1}^{q}\left(e^{(r_j-r_1)t}\sum_{\lambda_i\in\Lambda}\sum_{\lambda_k\in\Lambda}\left\|\left(\int_0^t e^{(\lambda_i-\lambda_k)s}dsP_i.P_k\right)\widehat{y}_0\right\|_2\right)^2}$$

Since $P_i$andP_k are orthogonal, only surveying terms are where $i=k$. For $i=k$

$$\int_0^t e^{(\lambda_i-\lambda_k)s}ds=t.$$

We are left with

$$\sqrt{\sum_{j=1}^{q}\left(e^{(r_j-r_1)t}\left\|Q_j\widehat{y}_0\right\|_2\right)^2}\left\|A\right\|_2 t$$

Substituting in (3.1.1), we get the result.

Now, we prove the second lower bound (3.1.6). We show that

$$\sup_{\substack{\widehat{B}\in\mathbb{R}^{n\times n}\\ \left\|\widehat{B}\right\|_2=1}}\left\|\sum_{\lambda_i\in\Lambda}\sum_{\lambda_k\in\Lambda}C(t,\lambda_i,\lambda_k)P_i\widehat{B}P_k\widehat{y}_0\right\|_2\geq\max_{\lambda_a,\lambda_b\in\Lambda}\left|D\left(t,\lambda_a,\lambda_b\right)\right|\left\|P_b\widehat{y}_0\right\|_2 \tag{3.1.8}$$

holds for (3.1.1)-(3.1.2). Fix $\lambda_a,\lambda_b\in\Lambda$ with $P_by_0\neq 0$. We consider the four cases:

A λ_a and λ_b are real;

B λ_a is not real and λ_b is real;

C λ_a is real and λ_b is not real;

D λ_a and λ_b are not real.

When λ_a is not real, let $\lambda_{\overline{a}}$, where $\overline{a}\in\{1,\ldots,p\}\setminus\{a\}$, be the eigenvalue which is the complex conjugate of λ_a. Similarly, when λ_b is not real, let $\lambda_{\overline{b}}$, where $\overline{b}\in\{1,\ldots,p\}\setminus\{b\}$, be the eigenvalue which is the complex conjugate of λ_b.

In the case A, consider a unit vector $\widehat{v}\in\mathbb{R}^n$ (i.e. $\left\|\widehat{v}\right\|_2=1$) such that $P_a\widehat{v}=\widehat{v}$ and consider the direction of perturbation

$$\widehat{B}=\widehat{v}\left(\frac{P_b\widehat{y}_0}{\left\|P_b\widehat{y}_0\right\|_2}\right)^H.$$

For $\lambda_k \in \Lambda$, we have

$$\begin{aligned}\widehat{B}P_k\widehat{y}_0 &= \widehat{v}\left(\frac{P_b\widehat{y}_0}{\|P_b\widehat{y}_0\|_2}\right)^H P_k\widehat{y}_0 = \frac{1}{\|P_b\widehat{y}_0\|_2}\left((P_b\widehat{y}_0)^H P_k\widehat{y}_0\right)\widehat{v} \\ &= \begin{cases} 0 \text{ if } k \neq b \\ \|P_b\widehat{y}_0\|_2\,\widehat{v} \text{ if } k = b. \end{cases}\end{aligned}$$

Thus

$$\begin{aligned}\sum_{\lambda_i\in\Lambda}\sum_{\lambda_k\in\Lambda} C(t,\lambda_i,\lambda_k)P_i\widehat{B}P_k\widehat{y}_0 &= \|P_b\widehat{y}_0\|_2 \sum_{\lambda_i\in\Lambda} C(t,\lambda_i,\lambda_b)P_i\widehat{v} \\ &= \|P_b\widehat{y}_0\|_2\, C(t,\lambda_a,\lambda_b)\widehat{v}\end{aligned}$$

and then

$$\begin{aligned}\left\|\sum_{\lambda_i\in\Lambda}\sum_{\lambda_k\in\Lambda} C(t,\lambda_i,\lambda_k)P_i\widehat{B}P_k\widehat{y}_0\right\|_2 &= \|P_b\widehat{y}_0\|_2\,|C(t,\lambda_a,\lambda_b)| \\ &= \|P_b\widehat{y}_0\|_2\,|D(t,\lambda_a,\lambda_b)|\,.\end{aligned}$$

In the case B, consider a unit vector $\widehat{v} \in \mathbb{R}^n$ such that $(P_a + P_{\overline{a}})\,\widehat{v} = \widehat{v}$ and consider the direction of perturbation

$$\widehat{B} = \widehat{v}\left(\frac{P_b\widehat{y}_0}{\|P_b\widehat{y}_0\|_2}\right)^H.$$

We have

$$\begin{aligned}\sum_{\lambda_i\in\Lambda}\sum_{\lambda_k\in\Lambda} C(t,\lambda_i,\lambda_k)P_i\widehat{B}P_k\widehat{y}_0 &= \|P_b\widehat{y}_0\|_2 \sum_{\lambda_i\in\Lambda} C(t,\lambda_i,\lambda_b)P_i\widehat{v} \\ &= \|P_b\widehat{y}_0\|_2\,(C(t,\lambda_a,\lambda_b)P_a\widehat{v} + C(t,\lambda_{\overline{a}},\lambda_b)P_{\overline{a}}\widehat{v})\end{aligned}$$

and then

$$\begin{aligned}&\left\|\sum_{\lambda_i\in\Lambda}\sum_{\lambda_k\in\Lambda} C(t,\lambda_i,\lambda_k)P_i\widehat{B}P_k\widehat{y}_0\right\|_2 \\ &= \|P_b\widehat{y}_0\|_2\,\|C(t,\lambda_a,\lambda_b)P_a\widehat{v} + C(t,\lambda_{\overline{a}},\lambda_b)P_{\overline{a}}\widehat{v}\|_2 \\ &= \|P_b\widehat{y}_0\|_2\,\sqrt{|C(t,\lambda_a,\lambda_b)|^2\,\|P_a\widehat{v}\|_2^2 + |C(t,\lambda_{\overline{a}},\lambda_b)|^2\,\|P_{\overline{a}}\widehat{v}\|_2^2} \\ &= \|P_b\widehat{y}_0\|_2\,|C(t,\lambda_a,\lambda_b)| = \|P_b\widehat{y}_0\|_2\,|D(t,\lambda_a,\lambda_b)|\,,\end{aligned}$$

where the second last = follows since $C(t,\lambda_a,\lambda_b)$ and $C(t,\lambda_{\overline{a}},\lambda_b)$ are complex conjugate.

In the case C, consider a unit vector $\widehat{v} \in \mathbb{R}^n$ such that $P_a\widehat{v} = \widehat{v}$ and consider the direction of perturbation

$$\widehat{B} = \widehat{v}\left(\frac{\left(P_b + P_{\overline{b}}\right)\widehat{y}_0}{\left\|\left(P_b + P_{\overline{b}}\right)\widehat{y}_0\right\|_2}\right)^H.$$

For $\lambda_k \in \Lambda$, we have

$$\widehat{B}P_k\widehat{y}_0 = \begin{cases} 0 \text{ if } k \neq b \text{ and } k \neq \overline{b} \\ \frac{\|P_b\widehat{y}_0\|_2^2}{\left\|\left(P_b+P_{\overline{b}}\right)\widehat{y}_0\right\|_2}\widehat{v} \text{ if } k = b \\ \frac{\left\|P_{\overline{b}}\widehat{y}_0\right\|_2^2}{\left\|\left(P_b+P_{\overline{b}}\right)\widehat{y}_0\right\|_2}\widehat{v} \text{ if } k = \overline{b}. \end{cases}$$

Since

$$\|P_b\widehat{y}_0\|_2^2 + \left\|P_{\overline{b}}\widehat{y}_0\right\|_2^2 = \left\|\left(P_b + P_{\overline{b}}\right)\widehat{y}_0\right\|_2^2 \quad \text{and} \quad \|P_b\widehat{y}_0\|_2 = \left\|P_{\overline{b}}\widehat{y}_0\right\|_2,$$

we get

$$\widehat{B}P_k\widehat{y}_0 = \begin{cases} 0 \text{ if } k \neq b \text{ and } k \neq \overline{b} \\ \frac{\sqrt{2}}{2}\|P_b\widehat{y}_0\|\,\widehat{v} \text{ if } k = b \text{ or } k = \overline{b}. \end{cases}$$

Thus

$$\begin{aligned} &\sum_{\lambda_i\in\Lambda}\sum_{\lambda_k\in\Lambda} C(t,\lambda_i,\lambda_k)P_i\widehat{B}P_k\widehat{y}_0 \\ &= \frac{\sqrt{2}}{2}\|P_b\widehat{y}_0\|\sum_{\lambda_i\in\Lambda}\left(C(t,\lambda_i,\lambda_b)+C(t,\lambda_i,\lambda_{\overline{b}})\right)P_i\widehat{v} \\ &= \frac{\sqrt{2}}{2}\|P_b\widehat{y}_0\|\left(C(t,\lambda_a,\lambda_b)+C(t,\lambda_a,\lambda_{\overline{b}})\right)\widehat{v} \end{aligned}$$

and then

$$\begin{aligned} \left\|\sum_{\lambda_i\in\Lambda}\sum_{\lambda_k\in\Lambda} C(t,\lambda_i,\lambda_k)P_i\widehat{B}P_k\widehat{y}_0\right\|_2 &= \frac{\sqrt{2}}{2}\|P_b\widehat{y}_0\|\left|C(t,\lambda_a,\lambda_b)+C(t,\lambda_a,\lambda_{\overline{b}})\right| \\ &= \|P_b\widehat{y}_0\|_2\left|D(t,\lambda_a,\lambda_b)\right|. \end{aligned}$$

In the case D, consider a unit vector $\widehat{v} \in \mathbb{R}^n$ such that $\left(P_a + P_{\overline{a}}\right)\widehat{v} = \widehat{v}$ and consider the direction of perturbation

$$\widehat{B} = \widehat{v}\left(\frac{\left(P_b + P_{\overline{b}}\right)\widehat{y}_0}{\left\|\left(P_b + P_{\overline{b}}\right)\widehat{y}_0\right\|_2}\right)^H.$$

We have

$$\begin{aligned} &\sum_{\lambda_i\in\Lambda}\sum_{\lambda_k\in\Lambda} C(t,\lambda_i,\lambda_k)P_i\widehat{B}P_k\widehat{y}_0 \\ &= \frac{\sqrt{2}}{2}\|P_b\widehat{y}_0\|_2\sum_{\lambda_i\in\Lambda}\left(C(t,\lambda_i,\lambda_b)+C(t,\lambda_i,\lambda_{\overline{b}})\right)P_i\widehat{v} \\ &= \frac{\sqrt{2}}{2}\|P_b\widehat{y}_0\|_2 \\ &\quad \left(\left(C(t,\lambda_a,\lambda_b)+C(t,\lambda_a,\lambda_{\overline{b}})\right)P_a\widehat{v} + \left(C(t,\lambda_{\overline{a}},\lambda_b)+C(t,\lambda_{\overline{a}},\lambda_{\overline{b}})\right)P_{\overline{a}}\widehat{v}\right) \end{aligned}$$

and then

$$\begin{aligned}
&\left\| \sum_{\lambda_i\in\Lambda}\sum_{\lambda_k\in\Lambda} C(t,\lambda_i,\lambda_k) P_i \widehat{B} P_k \widehat{y}_0 \right\|_2 \\
&= \frac{\sqrt{2}}{2} \left\| P_b \widehat{y}_0 \right\|_2 \\
&\quad \cdot \sqrt{\left| C(t,\lambda_a,\lambda_b) + C(t,\lambda_a,\lambda_{\overline{b}}) \right|^2 \left\| P_a \widehat{v} \right\|_2^2 + \left| C(t,\lambda_{\overline{a}},\lambda_b) + C(t,\lambda_{\overline{a}},\lambda_{\overline{b}}) \right|^2 \left\| P_{\overline{a}} \widehat{v} \right\|_2^2} \\
&= \frac{\sqrt{2}}{2} \left\| P_b \widehat{y}_0 \right\|_2 \left| C(t,\lambda_a,\lambda_b) + C(t,\lambda_a,\lambda_{\overline{b}}) \right| \\
&= \left\| P_b \widehat{y}_0 \right\|_2 \left| D(t,\lambda_a,\lambda_b) \right| .
\end{aligned}$$

where the second last $=$ follows since $C(t,\lambda_a,\lambda_b)+C(t,\lambda_a,\lambda_{\overline{b}})$ and $C(t,\lambda_{\overline{a}},\lambda_b)+C(t,\lambda_{\overline{a}},\lambda_{\overline{b}})$ are complex conjugate.

Now, (3.1.8) and then the lower bound (3.1.6) follow.

The upper bound (3.1.7) follows by observing that

$$\begin{aligned}
&\left\| \sum_{\lambda_i\in\Lambda}\sum_{\lambda_k\in\Lambda} C(t,\lambda_i,\lambda_k) P_i \widehat{B} P_k \widehat{y}_0 \right\|_2 = \left\| \sum_{\lambda_i\in\Lambda} P_i \widehat{B} \left(\sum_{\lambda_k\in\Lambda} C(t,\lambda_i,\lambda_k) P_k \widehat{y}_0 \right) \right\|_2 \\
&= \sqrt{\sum_{\lambda_i\in\Lambda} \left\| P_i \widehat{B} \left(\sum_{\lambda_k\in\Lambda} C(t,\lambda_i,\lambda_k) P_k \widehat{y}_0 \right) \right\|_2^2} \leq \sqrt{\sum_{\lambda_i\in\Lambda} \left\| \sum_{\lambda_k\in\Lambda} C(t,\lambda_i,\lambda_k) P_k \widehat{y}_0 \right\|_2^2} \\
&= \sqrt{\sum_{\lambda_i\in\Lambda}\sum_{\lambda_k\in\Lambda} \left| C(t,\lambda_i,\lambda_k) \right|^2 \left\| P_k \widehat{y}_0 \right\|^2}.
\end{aligned}$$

□

Let j^* be the minimum index $j \in \{1,\dots,q\}$ such that $Q_j y_0 \neq 0$. This index has been introduced at page 19 when we have considered the initial value perturbations. The next theorem gives neater bounds for $\overline{K}_2(t,A,y_0)$.

Theorem 3.1.3. *We have*

$$\overline{K}_2(t,A,y_0) \geq \left(\max_{\substack{\lambda_i\in\Lambda \\ \lambda_k\in \bigcup_{j=j^*}^{q} \Lambda_j}} \left| D\left(t,\lambda_i,\lambda_k\right) \right| \left\| P_k \widehat{y}_0 \right\|_2 \right) \cdot \|A\|_2 \, e^{(r_1 - r_{j^*})t}$$

and

$$\overline{K}_2(t,A,y_0) \leq \frac{\sqrt{|\Lambda|}}{\left\| Q_{j^*} \widehat{y}_0 \right\|_2} \|A\|_2 \, e^{(r_1 - r_{j^*})t} t,$$

where $|\Lambda|$ is the cardinality of the spectrum Λ.
In the generic situation $j^* = 1$ for y_0, we have

$$\|A\|_2 t \leq \overline{K}_2(t, A, y_0) \leq \frac{\sqrt{|\Lambda|}}{\|Q_1 \widehat{y}_0\|_2} \|A\|_2 t.$$

Proof. By the lower bound (3.1.6), we obtain

$$\begin{aligned}
\overline{K}_2(t, A, y_0) &\geq \frac{\max\limits_{\substack{\lambda_i \in \Lambda \\ \lambda_k \in \bigcup\limits_{j=j^*}^{q} \Lambda_j}} |D(t, \lambda_i, \lambda_k)| \, \|P_k \widehat{y}_0\|_2}{\sqrt{\sum\limits_{j=1}^{q} \left(e^{(r_j - r_1)t} \|Q_j \widehat{y}_0\|_2\right)^2}} \|A\|_2 \\
&\geq e^{(r_1 - r_{j^*})t} \max_{\substack{\lambda_i \in \Lambda \\ \lambda_k \in \bigcup\limits_{j=j^*}^{q} \Lambda_j}} |D(t, \lambda_i, \lambda_k)| \, \|P_k \widehat{y}_0\|_2 \, \|A\|_2 .
\end{aligned}$$

By the upper bound (3.1.7) and

$$|C(t, \lambda_i, \lambda_k)| \leq t \text{ for all } \lambda_i, \lambda_k \in \Lambda \tag{3.1.9}$$

(see Proposition 3.1.1), we obtain

$$\begin{aligned}
\overline{K}_2(t, A, y_0) &\leq \frac{\sqrt{\sum\limits_{\lambda_i \in \Lambda} \sum\limits_{\lambda_k \in \Lambda} t^2 \|P_k \widehat{y}_0\|^2}}{\sqrt{\sum\limits_{j=1}^{q} \left(e^{(r_j - r_1)t} \|Q_j \widehat{y}_0\|_2\right)^2}} \|A\|_2 \\
&= \frac{\sqrt{|\Lambda|}}{\sqrt{\sum\limits_{j=1}^{q} \left(e^{(r_j - r_1)t} \|Q_j \widehat{y}_0\|_2\right)^2}} \|A\|_2 t \\
&\leq \frac{\sqrt{|\Lambda|}}{\|Q_{j^*} \widehat{y}_0\|_2} \|A\|_2 e^{(r_1 - r_{j^*})t} t.
\end{aligned}$$

For $j^* = 1$ use the lower bound (3.1.5). □

The previous theorem shows a linear growth in t of $\overline{K}_2(t, A, y_0)$ for $j^* = 1$ and an exponential growth in t of $\overline{K}_2(t, A, y_0)$ for $j^* > 1$ (observe that

$$\max_{\substack{\lambda_i \in \Lambda \\ \lambda_k \in \bigcup\limits_{j=j^*}^{q} \Lambda_j}} |D(t, \lambda_i, \lambda_k)| \, \|P_k \widehat{y}_0\|_2$$

does not vanish as $t \to +\infty$: remind Remark 3.1.1).

Remark 3.1.2. *In the situation $j^* > 1$, $\overline{K}_2(t, A, y_0)$ can be arbitrarily larger than the lower bound (3.1.5), due to the exponential growth in t of the condition number. For $q > 1$, $\overline{K}_2(t, A, y_0)$ can be arbitrarily larger than the lower bound (3.1.5) also in the situation $j^* = 1$.*

In fact, for $q > 1$, the lower bound (3.1.6) gives

$$\overline{K}_2(t, A, y_0) \geq \frac{\max\limits_{\substack{\lambda_i \in \Lambda_1 \\ \lambda_k \in \Lambda \setminus \Lambda_1}} |D(t, \lambda_i, \lambda_k)| \, \|P_k \widehat{y}_0\|_2}{\sqrt{\sum\limits_{j=1}^{q} \left(e^{(r_j - r_1)t} \, \|Q_j \widehat{y}_0\|_2 \right)^2}} \, \|A\|_2$$

and the right-hand side of this inequality is a continuous function of $\|Q_1 \widehat{y}_0\|_2$, whose value for $\|Q_1 \widehat{y}_0\|_2 = 0$ is not smaller than

$$e^{(r_1 - r_2)t} \max_{\substack{\lambda_i \in \Lambda \\ \lambda_k \in \Lambda \setminus \Lambda_1}} |D(t, \lambda_i, \lambda_k)| \, \|P_k \widehat{y}_0\|_2 \, \|A\|_2 \, .$$

Hence, fixed $t \geq 0$ we have, for any $c \in (0, 1)$,

$$\overline{K}_2(t, A, y_0) \geq c e^{(r_1 - r_2)t} \max_{\substack{\lambda_i \in \Lambda \\ \lambda_k \in \Lambda \setminus \Lambda_1}} |D(t, \lambda_i, \lambda_k)| \, \|P_k \widehat{y}_0\|_2 \, \|A\|_2$$

for $\|Q_1 \widehat{y}_0\|_2$ sufficiently small.

This proves the following. Consider y_0 with fixed projections $P_k \widehat{y}_0$, $\lambda_k \in \Lambda \setminus \Lambda_1$. For any $M > 1$, there exists $t \geq 0$ such that,

$$\frac{\overline{K}_2(t, A, y_0)}{\|A\|_2 \, t} \geq M$$

for $\|Q_1 \widehat{y}_0\|_2$ sufficiently small.

The next results concerns the case $q = 1$, namely the case of shifted skew-symmetric matrices.

Theorem 3.1.4. *If $q = 1$, i.e. A is a shifted skew-symmetric matrix, then*

$$\overline{K}_2(t, A, y_0) = \|A\|_2 \, t.$$

Proof. For A shifted skew-symmetric, i.e. $A = \alpha I + S$ for some $\alpha \in \mathbb{R}$ and $S \in \mathbb{R}^{n \times n}$ skew-symmetric, by (3.1.1) we get

$$\overline{K}_2(t, A, y_0, \widehat{B}) = \left\| \int_0^t e^{-sA} \widehat{B} e^{sA} ds \widehat{y}_0 \right\|_2 \|A\|_2 \leq \left\| \int_0^t e^{-sA} \widehat{B} e^{sA} ds \right\|_2 \|A\|_2 \, .$$

Now,

$$\int_0^t e^{-sA} \widehat{B} e^{sA} ds = \int_0^t e^{-s\alpha} e^{-sS} \widehat{B} e^{s\alpha} e^{sS} ds = \int_0^t e^{-sS} \widehat{B} e^{sS} ds$$

and then

$$\left\| \int_0^t e^{-sA} \widehat{B} e^{sA} ds \right\|_2 \leq \int_0^t \left\| e^{-sS} \right\|_2 \left\| \widehat{B} \right\|_2 \left\| e^{sS} \right\|_2 ds = t$$

since e^{-sS} and e^{sS} are orthogonal matrices. Thus

$$\overline{K}_2(t, A, y_0, \widehat{B}) \leq \|A\|_2 t.$$

We conclude that

$$\overline{K}_2(t, A, y_0) \leq \|A\|_2 t.$$

The thesis follows by recalling the lower bound (3.1.5). □

When y_0 stays in the rightmost eigenspace, we have the same situation of the case $q = 1$, namely the condition number is equal to $\|A\|_2 t$.

Theorem 3.1.5. *If $Q_1 y_0 = y_0$, then*

$$\overline{K}_2(t, A, y_0) = \|A\|_2 t.$$

Proof. Assume $Q_1 y_0 = y_0$.

In our discussion we are assuming that A is a normal real matrix, but (3.1.1)-(3.1.2) also holds when A is a normal complex matrix.

So, now, we consider the case where A is a normal complex matrix with a unique complex eigenvalue λ_1 as rightmost eigenvalue. We have, for the numerator (3.1.2) in the right-hand side of (3.1.1),

$$\begin{aligned} &\left\| \sum_{\lambda_i \in \Lambda} \sum_{\lambda_k \in \Lambda} C(t, \lambda_i, \lambda_k) P_i \widehat{B} P_k \widehat{y}_0 \right\|_2 = \left\| \sum_{\lambda_i \in \Lambda} C(t, \lambda_i, \lambda_1) P_i \widehat{B} \widehat{y}_0 \right\|_2 \\ &= \sqrt{\sum_{\lambda_i \in \Lambda} |C(t, \lambda_i, \lambda_1)|^2 \left\| P_i \widehat{B} \widehat{y}_0 \right\|_2^2} \leq t \left\| \widehat{B} \widehat{y}_0 \right\|_2 \leq t \end{aligned}$$

by recalling (3.1.9). Thus, since the denominator in the right-hand side of (3.1.1) is 1, we obtain

$$\overline{K}_2(t, A, y_0, \widehat{B}) \leq \|A\|_2 t.$$

Now, we pass to consider the case where A is a normal real matrix. Fix a direction of perturbation $\widehat{B}$. For any $\varepsilon > 0$, there exists a normal complex matrix A_ε such that A_ε has a unique complex eigenvalue λ_1 as rightmost eigenvalue,

$$\left| \overline{K}_2(t, A, y_0, \widehat{B}) - \overline{K}_2(t, A_\varepsilon, y_0, \widehat{B}) \right| \leq \varepsilon$$

and

$$\|A - A_\varepsilon\|_2 \leq \varepsilon.$$

Thus

$$\begin{aligned}\overline{K}_2(t,A,y_0,\widehat{B}) &= \overline{K}_2(t,A,y_0,\widehat{B}) - \overline{K}_2(t,A_\varepsilon,y_0,\widehat{B}) + \overline{K}_2(t,A_\varepsilon,y_0,\widehat{B}) \\ &\leq \varepsilon + \|A_\varepsilon\|_2 t \\ &\leq \varepsilon + \varepsilon t + \|A\|_2 t.\end{aligned}$$

Since ε is arbitrarily small, we obtain

$$\overline{K}_2(t,A,y_0,\widehat{B}) \leq \|A\|_2 t.$$

By using the lower bound (3.1.5), $\overline{K}_2(t,A,y_0) = \|A\|_2 t$ follows. This is also true when A is a normal complex matrix. □

Observe that now Theorem 3.1.4 becomes a corollary of Theorem 3.1.5.

3.1.3 The condition number $\overline{K}_2(t,A)$ independent of the data

The next theorem gives lower and upper bounds for $\overline{K}_2(t,A)$.

Theorem 3.1.6. *We have the lower bound*

$$\overline{K}_2(t,A) \geq \max_{\substack{\lambda_i\in\Lambda \\ \lambda_k\in\Lambda_q}} |D(t,\lambda_i,\lambda_k)|\,\|A\|_2\, e^{(r_1-r_q)t}. \tag{3.1.10}$$

Moreover, we have the upper bound

$$\overline{K}_2(t,A) \leq \sqrt{\max_{\lambda_k\in\Lambda} \sum_{\lambda_i\in\Lambda} |C(t,\lambda_i,\lambda_k)|^2}\,\|A\|_2\, e^{(r_1-r_q)t}. \tag{3.1.11}$$

Proof. First, we prove the lower bound. For any $\lambda_k \in \Lambda_q$, consider $y_0 \neq 0$ such that $P_k y_0 = y_0$. By (3.1.6) we have

$$\overline{K}_2(t,A) \geq \overline{K}_2(t,A,y_0) \geq \frac{\max_{\lambda_i\in\Lambda} |D\,(t,\lambda_i,\lambda_k)\,|}{e^{(r_q-r_1)t}}\,\|A\|_2\,.$$

Now, we prove the upper bound. For the numerator in (3.1.7), we have

$$\sqrt{\sum_{\lambda_i\in\Lambda}\sum_{\lambda_k\in\Lambda} |C(t,\lambda_i,\lambda_k)|^2\,\|P_k\widehat{y}_0\|^2} \leq \sqrt{\max_{\lambda_k\in\Lambda}\sum_{\lambda_i\in\Lambda} |C(t,\lambda_i,\lambda_k)|^2}$$

and for the denominator we have

$$\sqrt{\sum_{j=1}^{q} \left(e^{(r_j-r_1)t}\,\|Q_j\widehat{y}_0\|_2\right)^2} \geq e^{(r_q-r_1)t}.$$

□

The previous theorem shows that $\overline{K}_2(t,A)$ grows exponentially in t for $q>1$.
Since (3.1.9) holds, the upper bound (3.1.11) gives this other neater upper bound.

Theorem 3.1.7. *We have*

$$\overline{K}_2(t,A) \leq \sqrt{|\Lambda|}\ \|A\|_2\, e^{(r_q-r_1)t}t.$$

Proof. By using (3.1.9) in (3.1.11) we get the result. □

3.2 Asymptotic analysis

In this section, we study the asymptotic behavior of the three condition numbers $\overline{K}_2(t,A,y_0,\widehat{B})$, $\overline{K}_2(t,A,y_0)$ and $\overline{K}_2(t,A)$, as $t\to+\infty$.

We use the following notations.

- Let j^* be the minimum index $j\in\{1,\ldots,q\}$ such that $Q_jy_0\neq 0$. This index has been already introduced in subsection 3.1, just before the Theorem 3.1.3 as well as at page 19.
- Let j^{**} be the minimum index $j\in\{1,\ldots,q\}$ such that $Q_j\widehat{B}P_ky_0\neq 0$ for some $k\in\{1,\ldots,p\}$.
- For $\lambda_i\in\Lambda_j$ and $\lambda_k\in\Lambda_l$, where $j,l\in\{1,\ldots,q\}$ with $j<l$, let

$$C_\infty(\lambda_i,\lambda_k) := \frac{1}{\lambda_i-\lambda_k}.$$

- For $\lambda_i\in\Lambda_j$ and $\lambda_k\in\Lambda_l$, where $j,l\in\{1,\ldots,q\}$ with $j>l$, let

$$C_\infty(t,\lambda_i,\lambda_k) := \frac{e^{\sqrt{-1}(\omega_k-\omega_i)t}}{\lambda_k-\lambda_i},$$

  where ω_i and ω_k are the imaginary parts of λ_i and λ_k, respectively, and $\sqrt{-1}$ is the imaginary unit.

- For $\lambda_i\in\Lambda_j$ and $\lambda_k\in\Lambda_l$, where $j,l\in\{1,\ldots,q\}$ with $j<l$, let

$$D_\infty(\lambda_i,\lambda_k) := \begin{cases} C_\infty(\lambda_i,\lambda_k) \text{ if } \lambda_k \text{ is real} \\ \\ \frac{\sqrt{2}}{2}\left(C_\infty(\lambda_i,\lambda_k)+C_\infty\left(\lambda_i,\overline{\lambda_k}\right)\right) \text{ if } \lambda_k \text{ is not real.} \end{cases}$$

- $f(t)\sim g(t),\ t\to+\infty$, stands for

$$\lim_{t\to+\infty}\frac{f(t)}{g(t)}=1.$$

  This notation has been already introduced at page 19.

- $f(t) \lesssim g(t),\ t \to +\infty$, stands for

$$f(t) \leq h(t), \quad \text{for } t \text{ sufficiently large,}$$

and

$$h(t) \sim g(t),\ t \to +\infty,$$

for some function $h(t)$. Similarly, $f(t) \gtrsim g(t),\ t \to +\infty$, stands for

$$f(t) \geq h(t), \quad \text{for } t \text{ sufficiently large,}$$

and

$$h(t) \sim g(t),\ t \to +\infty,$$

for some function $h(t)$.

The next proposition, which is a trivial consequence of Proposition 3.1.1, describes the asymptotic behavior, as $t \to +\infty$, of the functions $C(t, \lambda_i, \lambda_k)$ defined in (3.1.3).

Proposition 3.2.1. *Let $j, l \in \{1, \ldots, q\}$, let $\lambda_i \in \Lambda_j$ and let $\lambda_k \in \Lambda_l$.*

If $j < l$, then

$$C(t, \lambda_i, \lambda_k) \sim e^{(r_j - r_1)t} C_\infty(\lambda_i, \lambda_k),\ t \to +\infty.$$

If $j > l$, then

$$C(t, \lambda_i, \lambda_k) \sim e^{(r_l - r_1)t} C_\infty(t, \lambda_i, \lambda_k),\ t \to +\infty.$$

If $j = l$ and $\lambda_i \neq \lambda_k$, then

$$C(t, \lambda_i, \lambda_k) = e^{(r_j - r_1)t} \frac{e^{\sqrt{-1}(\omega_k - \omega_i)t} - 1}{\lambda_k - \lambda_i},\ t \geq 0.$$

If $j = l$ and $\lambda_i = \lambda_k$, then

$$C(t, \lambda_i, \lambda_k) = e^{(r_j - r_1)t} t,\ t \geq 0.$$

3.2.1 Asymptotic analysis of the condition number $\overline{K}_2(t, A, y_0, \widehat{B})$ with direction of perturbation

The next theorem describes the asymptotic behavior of $\overline{K}_2(t, A, y_0, \widehat{B})$, as $t \to +\infty$.

Theorem 3.2.1.

If $j^{**} < j^*$, then

$$\overline{K}_2(t, A, y_0, \widehat{B}) = \frac{\left\| \sum\limits_{\lambda_i \in \Lambda_{j^{**}}} \sum\limits_{l=j^*}^{q} \sum\limits_{\lambda_k \in \Lambda_l} C_\infty(\lambda_i, \lambda_k) P_i \widehat{B} P_k \widehat{y}_0 \right\|_2}{\|Q_{j^*} \widehat{y}_0\|_2} \|A\|_2 e^{(r_{j^{**}} - r_{j^*})t} + o(e^{(r_{j^{**}} - r_{j^*})t}),\ t \to +\infty.$$

If $j^{**} = j^*$, then

$$\overline{K}_2(t, A, y_0, \widehat{B}) = \frac{\left\| \sum\limits_{\lambda_i \in \Lambda_{j^*}} P_i \widehat{B} P_i \widehat{y}_0 \right\|_2}{\|Q_{j^*} \widehat{y}_0\|_2} \|A\|_2 \, t + o(t), \ t \to +\infty.$$

If $j^{**} > j^*$, then

$$\overline{K}_2(t, A, y_0, \widehat{B}) = \frac{\left\| \sum\limits_{j=j^{**}}^{q} \sum\limits_{\lambda_i \in \Lambda_j} \sum\limits_{\lambda_k \in \Lambda_{j^*}} C_\infty \left(t, \lambda_i, \lambda_k\right) P_i \widehat{B} P_k \widehat{y}_0 \right\|_2}{\|Q_{j^*} \widehat{y}_0\|_2} \|A\|_2 + o(1), \ t \to +\infty.$$

Proof. We write (3.1.1)-(3.1.2) as

$$\overline{K}_2(t, A, y_0, B) = \frac{\left\| \sum\limits_{j=j^{**}}^{q} \sum\limits_{\lambda_i \in \Lambda_j} \sum\limits_{l=j^*}^{q} \sum\limits_{\lambda_k \in \Lambda_l} C(t, \lambda_i, \lambda_k) P_i \widehat{B} P_k \widehat{y}_0 \right\|_2}{\sqrt{\sum\limits_{j=j^*}^{q} \left(e^{(r_j - r_1)t} \left\| Q_j \widehat{y}_0 \right\|_2 \right)^2}} \|A\|_2 . \tag{3.2.1}$$

Consider the numerator in (3.2.1). If $j^{**} < j^*$, then, by Proposition 3.2.1, the major contributory terms $C(t, \lambda_i, \lambda_k)$ as $t \to +\infty$ are obtained for $j = j^{**}$ and then

$$\begin{aligned} &\left\| \sum_{j=j^{**}}^{q} \sum_{\lambda_i \in \Lambda_j} \sum_{l=j^*}^{q} \sum_{\lambda_k \in \Lambda_l} C(t, \lambda_i, \lambda_k) P_i \widehat{B} P_k \widehat{y}_0 \right\|_2 \\ &= e^{(r_{j^{**}} - r_1)t} \left\| \sum_{\lambda_i \in \Lambda_{j^{**}}} \sum_{l=j^*}^{q} \sum_{\lambda_k \in \Lambda_l} C_\infty \left(\lambda_i, \lambda_k\right) P_i \widehat{B} P_k \widehat{y}_0 \right\|_2 \\ &\quad + o(e^{(r_{j^{**}} - r_{j^*})t}), \ t \to +\infty. \end{aligned}$$

If $j^{**} = j^*$, then the major contributory terms $C(t, \lambda_i, \lambda_k)$ as $t \to +\infty$ are obtained for $j = l = j^*$ and $\lambda_i = \lambda_k$ and then

$$\begin{aligned} &\left\| \sum_{j=j^{**}}^{q} \sum_{\lambda_i \in \Lambda_j} \sum_{l=j^*}^{q} \sum_{\lambda_k \in \Lambda_l} C(t, \lambda_i, \lambda_k) P_i \widehat{B} P_k \widehat{y}_0 \right\|_2 \\ &= e^{(r_{j^*} - r_1)t} t \left\| \sum_{\lambda_i \in \Lambda_{j^*}} P_i \widehat{B} P_i \widehat{y}_0 \right\|_2 + o\left(e^{(r_{j^*} - r_1)t} t \right), \ t \to +\infty. \end{aligned}$$

If $j^{**} > j^*$, the major contributory terms $C(t, \lambda_i, \lambda_k)$ as $t \to +\infty$ are obtained for $l = j^*$ and then

$$\left\| \sum_{j=j^{**}}^{q} \sum_{\lambda_i \in \Lambda_j} \sum_{l=j^*}^{q} \sum_{\lambda_k \in \Lambda_l} C(t, \lambda_i, \lambda_k) P_i \widehat{B} P_k \widehat{y}_0 \right\|_2$$
$$= e^{(r_{j^*} - r_1)t} \left\| \sum_{j=j^{**}}^{q} \sum_{\lambda_i \in \Lambda_j} \sum_{\lambda_k \in \Lambda_{j^*}} C_\infty(t, \lambda_i, \lambda_k) P_i \widehat{B} P_k \widehat{y}_0 \right\|_2 + o\left(e^{(r_{j^*} - r_1)t}\right)$$
$t \to +\infty$.

Consider the denominator in (3.2.1). The major contributory term as $t \to +\infty$ is $e^{(r_{j^*} - r_1)t} \left\| Q_{j^*} \widehat{y}_0 \right\|_2$ and then

$$\sqrt{\sum_{j=j^*}^{q} \left(e^{(r_j - r_1)t} \left\| Q_j \widehat{y}_0 \right\|_2 \right)^2} \sim e^{(r_{j^*} - r_1)t} \left\| Q_{j^*} \widehat{y}_0 \right\|_2, \ t \to +\infty.$$

Now, the theorem follows. □

Remark 3.2.1. *Observe that the generic situation for the initial value y_0 and the direction of the perturbation $\widehat{B}$ is $j^* = 1$, $j^{**} = 1$ and*

$$\sum_{\lambda_i \in \Lambda_1} P_i \widehat{B} P_i \widehat{y}_0 \neq 0,$$

where we have

$$\overline{K}_2(t, A, y_0, \widehat{B}) \sim \frac{\left\| \sum_{\lambda_i \in \Lambda_1} P_i \widehat{B} P_i \widehat{y}_0 \right\|_2}{\left\| Q_1 \widehat{y}_0 \right\|_2} \left\| A \right\|_2 t, \ t \to +\infty.$$

In the non-generic situation $j^* > 1$ or $j^{**} > 1$, the previous theorem shows that:

- *if $j^{**} < j^*$ and*

$$\sum_{\lambda_i \in \Lambda_{j^{**}}} \sum_{l=j^*}^{q} \sum_{\lambda_k \in \Lambda_l} C_\infty(\lambda_i, \lambda_k) P_i \widehat{B} P_k \widehat{y}_0 \neq 0,$$

  then $\overline{K}_2(t, A, y_0, \widehat{B})$ grows exponentially in t as $t \to +\infty$:

$$\overline{K}_2(t, A, y_0, \widehat{B}) \sim \frac{\left\| \sum_{\lambda_i \in \Lambda_{j^{**}}} \sum_{l=j^*}^{q} \sum_{\lambda_k \in \Lambda_l} C_\infty(\lambda_i, \lambda_k) P_i \widehat{B} P_k \widehat{y}_0 \right\|_2}{\left\| Q_{j^*} \widehat{y}_0 \right\|_2} \left\| A \right\|_2 e^{(r_{j^{**}} - r_{j^*})t}$$
$t \to +\infty$;

- *if* $j^{**} = j^*$ *and*

$$\sum_{\lambda_i \in \Lambda_{j^*}} P_i \widehat{B} P_i \widehat{y}_0 \neq 0,$$

then $\overline{K}_2(t, A, y_0, \widehat{B})$ *grows linearly in* t *as* $t \to +\infty$*:*

$$\overline{K}_2(t, A, y_0, \widehat{B}) \sim \frac{\left\| \sum_{\lambda_i \in \Lambda_{j^*}} P_i \widehat{B} P_i \widehat{y}_0 \right\|_2}{\|Q_{j^*} \widehat{y}_0\|_2} \|A\|_2 \, t, \ t \to +\infty;$$

- *if* $j^{**} > j^*$*, then* $\overline{K}_2(t, A, y_0, \widehat{B})$ *oscillates (due to the terms* $C_\infty(t, \lambda_i, \lambda_k)$ *in (3.2.1)), but it remains bounded as* $t \to +\infty$*.*

3.2.2 Asymptotic analysis of the condition number $\overline{K}_2(t, A, y_0)$

The next theorem describes the asymptotic behavior of $\overline{K}_2(t, A, y_0)$, as $t \to +\infty$.

Theorem 3.2.2. *If* $j^* = 1$*, we have*

$$\overline{K}(t, A, y_0) \sim \|A\|_2 \, t, \ t \to +\infty. \tag{3.2.2}$$

If $j^* > 1$*, we have the asymptotic lower bound*

$$\begin{aligned} &\overline{K}(t, A, y_0) \gtrsim \frac{\max\limits_{\lambda_i \in \Lambda_1} \left| D_\infty(\lambda_i, \lambda_k) \right| \|P_k y_0\|_2}{\|Q_{j^*} \widehat{y}_0\|_2} \|A\|_2 \, e^{(r_1 - r_{j^*})t} \\ &t \to +\infty \end{aligned} \tag{3.2.3}$$

and the asymptotic upper bound

$$\begin{aligned} &\overline{K}(t, A, y_0) \lesssim \frac{\sqrt{\sum\limits_{l=j^*}^{q} \sum\limits_{\lambda_k \in \Lambda_l} \left(\sum\limits_{\lambda_i \in \Lambda_1} |C_\infty(\lambda_i, \lambda_k)|^2 \right) \|P_k \widehat{y}_0\|^2}}{\|Q_{j^*} \widehat{y}_0\|_2} \|A\|_2 \, e^{(r_1 - r_{j^*})t} \\ &t \to +\infty. \end{aligned} \tag{3.2.4}$$

Proof. We write the right-hand side of the upper bound (3.1.7) as

$$\frac{\sqrt{\sum\limits_{j=1}^{q} \sum\limits_{\lambda_i \in \Lambda_j} \sum\limits_{l=j^*}^{q} \sum\limits_{\lambda_k \in \Lambda} |C(t, \lambda_i, \lambda_k)|^2 \|P_k \widehat{y}_0\|^2}}{\sqrt{\sum\limits_{j=j^*}^{q} \left(e^{(r_j - r_1)t} \|Q_j \widehat{y}_0\|_2 \right)^2}} \|A\|_2 .$$

Suppose $j^* = 1$. The major contributory terms $C(t, \lambda_i, \lambda_k)$ as $t \to +\infty$ in the numerator are obtained for $j = l = 1$ and $\lambda_i = \lambda_k$ and then

$$\sqrt{\sum_{j=1}^{q} \sum_{\lambda_i \in \Lambda_j} \sum_{l=j^*}^{q} \sum_{\lambda_k \in \Lambda_l} |C(t, \lambda_i, \lambda_k)|^2 \|P_k \widehat{y}_0\|_2^2}$$
$$\sim t \|Q_1 \widehat{y}_0\|_2, \ t \to +\infty.$$

The major contributory term in the numerator is $\|Q_1 \widehat{y}_0\|_2$ and then

$$\sqrt{\sum_{j=1}^{q} \left(e^{(r_j - r_1)t} \|Q_j \widehat{y}_0\|_2\right)^2} \sim \|Q_1 \widehat{y}_0\|_2, \ t \to +\infty.$$

Thus

$$\overline{K}(t, A, y_0) \lesssim \|A\|_2 t, \ t \to +\infty.$$

Now, (3.2.2) follows by the lower bound (3.1.5).

Suppose $j^* > 1$. The major contributory terms $C(t, \lambda_i, \lambda_k)$ as $t \to +\infty$ in the numerator are obtained for $j = 1$ and then

$$\sqrt{\sum_{j=1}^{q} \sum_{\lambda_i \in \Lambda_j} \sum_{l=j^*}^{q} \sum_{\lambda_k \in \Lambda_l} |C(t, \lambda_i, \lambda_k)|^2 \|P_k \widehat{y}_0\|_2^2}$$
$$\sim \sqrt{\sum_{l=j^*}^{q} \sum_{\lambda_k \in \Lambda_l} \left(\sum_{\lambda_i \in \Lambda_1} |C_\infty(\lambda_i, \lambda_k)|^2 \right) \|P_k \widehat{y}_0\|^2}, \ t \to +\infty.$$

The major contributory term in the denumerator is $e^{(r_{j^*} - r_1)t} \|Q_{j^*} \widehat{y}_0\|_2$ and then

$$\sqrt{\sum_{j=1}^{q} \left(e^{(r_j - r_1)t} \|Q_j \widehat{y}_0\|_2\right)^2} \sim e^{(r_{j^*} - r_1)t} \|Q_{j^*} \widehat{y}_0\|_2, \ t \to +\infty.$$

Now, the asymptotic upper bound (3.2.4) follows.

Finally, we prove the asymptotic lower bound (3.2.3). By the lower bound (3.1.6), we have

$$\overline{K}_2(t, A, y_0) \geq \frac{\max\limits_{\substack{\lambda_i \in \Lambda \\ \lambda_k \in \bigcup\limits_{j=j^*}^{q} \Lambda_j}} |D(t, \lambda_i, \lambda_k)| \|P_k y_0\|_2}{\sqrt{\sum\limits_{j=1}^{q} \left(e^{(r_j - r_1)t} \|Q_j \widehat{y}_0\|_2\right)^2}} \|A\|_2.$$

By Proposition 3.2.1, we obtain

$$\max_{\substack{\lambda_i \in \Lambda \\ \lambda_k \in \bigcup\limits_{j=j^*}^{q} \Lambda_j}} |D(t, \lambda_i, \lambda_k)| \|P_k y_0\|_2 \sim \max_{\substack{\lambda_i \in \Lambda_1 \\ \lambda_k \in \bigcup\limits_{j=j^*}^{q} \Lambda_j}} |D_\infty(\lambda_i, \lambda_k)| \|P_k y_0\|_2, \ t \to +\infty.$$

Thus

$$\begin{aligned}
&\frac{\max\limits_{\lambda_i\in\Lambda} \quad |D\left(t,\lambda_i,\lambda_k\right)|\,\|P_k y_0\|_2}{\sqrt{\sum\limits_{j=1}^{q}\left(e^{(r_j-r_1)t}\,\|Q_j\widehat{y}_0\|_2\right)^2}}\,\|A\|_2 \\
&\qquad\qquad \lambda_k\in\bigcup\limits_{j=j^*}^{q}\Lambda_j \\
&\qquad\qquad \max\limits_{\lambda_i\in\Lambda_1} \quad |D_\infty\left(\lambda_i,\lambda_k\right)|\,\|P_k y_0\|_2 \\
\sim\ &\frac{\lambda_k\in\bigcup\limits_{j=j^*}^{q}\Lambda_j}{e^{(r_{j^*}-r_1)t}\,\|Q_{j^*}\widehat{y}_0\|_2}\,\|A\|_2\,,\ t\to+\infty,
\end{aligned}$$

and the asymptotic lower bound follows. □

Remark 3.2.2. *The generic situation for the initial value y_0 is $j^*=1$, where we have the asymptotic behavior (3.2.2) which is independent of y_0.*

It is interesting to observe that, for the problem (3.0.5), the condition number relevant to the spectral norm of matrices is $\|A\|_2\,t$ in case of a normal matrix A (see [54]). So, asymptotically as $t\to+\infty$, the condition numbers of the problems (3.0.2) *and (3.0.5) are equal for a normal matrix in the generic situation $j^*=1$ for y_0.*

Remark 3.2.3. *In the non-generic situation $j^*>1$ for y_0, the previous theorem says that*

$$\begin{aligned}
\overline{K}_2(t,A,y_0) &= O\left(e^{(r_1-r_{j^*})t}\right),\ t\to+\infty \\
\frac{1}{\overline{K}_2(t,A,y_0)} &= O\left(e^{-(r_1-r_{j^*})t}\right),\ t\to+\infty.
\end{aligned}$$

We also have

$$\log\overline{K}_2(t,A,y_0)\sim(r_1-r_{j^*})\,t,\ t\to+\infty.$$

3.2.3 Asymptotic analysis of the condition number $\overline{K}_2(t,A)$ independent of the data

The next theorem describes the asymptotic behavior of $\overline{K}_2(t,A)$, as $t\to+\infty$.

Theorem 3.2.3. *We have the asymptotic lower bound*

$$K(t,A)\gtrsim\max_{\substack{\lambda_i\in\Lambda_1\\ \lambda_k\in\Lambda_q}}|D_\infty(\lambda_i,\lambda_k)|\,\|A\|_2\,e^{(r_1-r_q)t},\ t\to+\infty,$$

and the asymptotic upper bound

$$K(t,A)\lesssim\|A\|_2\,e^{(r_1-r_q)t}t,\ t\to+\infty.$$

Proof. By the lower bound (3.1.10), we have

$$\begin{aligned}\overline{K}_2(t,A) &\geq \max_{\substack{\lambda_i\in\Lambda\\ \lambda_k\in\Lambda_q}} |D(t,\lambda_i,\lambda_k)|\,\|A\|_2\, e^{(r_1-r_q)t}\\ &\sim \max_{\substack{\lambda_i\in\Lambda_1\\ \lambda_k\in\Lambda_q}} |D_\infty(\lambda_i,\lambda_k)|\,\|A\|_2\, e^{(r_1-r_q)t},\ t\to+\infty.\end{aligned}$$

By the upper bound (3.1.11), we have

$$\begin{aligned}\overline{K}_2(t,A) &\leq \sqrt{\max_{\lambda_k\in\Lambda}\sum_{\lambda_i\in\Lambda}|C(t,\lambda_i,\lambda_k)|^2}\,\|A\|_2\, e^{(r_1-r_q)t}\\ &\sim \|A\|_2\, e^{(r_1-r_q)t}t,\ t\to+\infty.\end{aligned}$$

□

Remark 3.2.4. *The previous theorem says that*

$$\begin{aligned}\overline{K}_2(t,A) &= O\left(e^{(r_1-r_q)t}t\right),\ t\to+\infty\\ \frac{1}{\overline{K}_2(t,A)} &= O\left(e^{-(r_1-r_q)t}\right),\ t\to+\infty.\end{aligned}$$

We also have

$$\log\overline{K}_2(t,A)\sim(r_1-r_q)\,t,\ t\to+\infty.$$

3.3 Numerical tests

The numerical tests involve the condition number $\overline{K}_2(t,A,y_0)$. We consider skew symmetric matrices in the Example 3.3.1, with the aim to confirming Theorem 3.1.4, and symmetric matrices in the Example 3.3.2, with the aim of confirming Theorem 3.2.2.

Example 3.3.1. *Consider the following two cases of a skew symmetric matrix A in (3.0.1):*

- *the 2×2 matrix*
$$A=\begin{bmatrix}0 & 3\\ -3 & 0\end{bmatrix},$$
which has the pair of pure imaginary eigenvalues $\pm 3\sqrt{-1}$;

- *the 4×4 matrix*
$$A=\begin{bmatrix}0 & 2 & -1 & 3\\ -2 & 0 & -4 & 1\\ 1 & 4 & 0 & 2\\ -3 & -1 & -2 & 0\end{bmatrix}$$
which has the two pairs of pure imaginary eigenvalues $\pm 5.7913\sqrt{-1}$ and $\pm 1.2087\sqrt{-1}$.

In Figure 3.1 , for both the skew symmetric matrices and for t in a uniform mesh over the interval $[0, 50]$, we plot the maximum of the values

$$\frac{\frac{\xi(t)}{\varepsilon}}{\|A\|_2\, t} = \frac{\overline{K}_2(t, A, y_0, \widehat{B})}{\|A\|_2\, t} + o(1),\ \epsilon \to 0, \tag{3.3.1}$$

over 10000 random selections of the unit matrix $\widehat{B}$. We consider the initial values $y_0 = (1, 2)$ for the 2×2 matrix and $y_0 = (1, 2, 3, 4)$ for the 4×4 matrix. We take $\epsilon = 10^{-4}$.

For both matrices, as t varies, the maximum of the values (3.3.1) is always close to 1, confirming Theorem 3.1.4.

For the matrix 2×2, we observe a slight deviation from 1 as t increases. This is due to the error $o(1)$, as $\epsilon \to 0$, in (3.3.1).

The maximum values for the 2×2 matrix are closer to 1 than the maximum values for the matrix 4×4. This is due to the fact that much more than 10000 random selections of the matrix $\widehat{B}$ are necessary for having maximum values very close to 1, in case of the matrix 4×4.

Example 3.3.2. *Consider the following two cases of a symmetric matrix A in (3.0.1):*

- *the 2×2 matrix*
$$A = \begin{bmatrix} -2 & 1 \\ 1 & -2 \end{bmatrix},$$
which has the eigenvalues -1 and -3;

- *the 4×4 matrix*
$$A = 1/2 \begin{bmatrix} -1 & 2 & 1 & 0 \\ 2 & -1 & 0 & -1 \\ 1 & 0 & -1 & -2 \\ 0 & -1 & -2 & -1 \end{bmatrix},$$
which has the eigenvalues 1, 0, -1 and -2.

In Figure 3.2, for both the symmetric matrices and for t in a uniform mesh over the interval $[0, 15]$, we plot the maximum of the values (3.3.1) over 10000 random selections of the unit matrix $\widehat{B}$. We consider the initial values $y_0 = (1, 2)$ for the 2×2 matrix and $y_0 = (1, 2, 3, 4)$ for the 4×4 matrix. For such initial values we have $j^* = 1$ (the index j^* is defined at the beginning of Section 3.2). We take $\epsilon = 10^{-4}$.

For both matrices, as t varies, the maximum of the values (3.3.1) tends asymptotically to 1, after an initial hump. This confirms Theorem 3.2.2, case $j^* = 1$. About the initial hump, see Remark 3.1.2.

In Figure 3.3, for the 2×2 matrix and for t in a uniform mesh over the interval $[0, 15]$, we plot the maximum of the values (3.3.1) in 10000 random selections of matrix $\widehat{B}$, when the initial values are $y_0 = (1, 1)$, which is eigenvector of the rightmost eigenvalue -1, and $y_0 = (1, -1)$, which is eigenvector of the other eigenvalue -3. We take $\epsilon = 10^{-4}$.

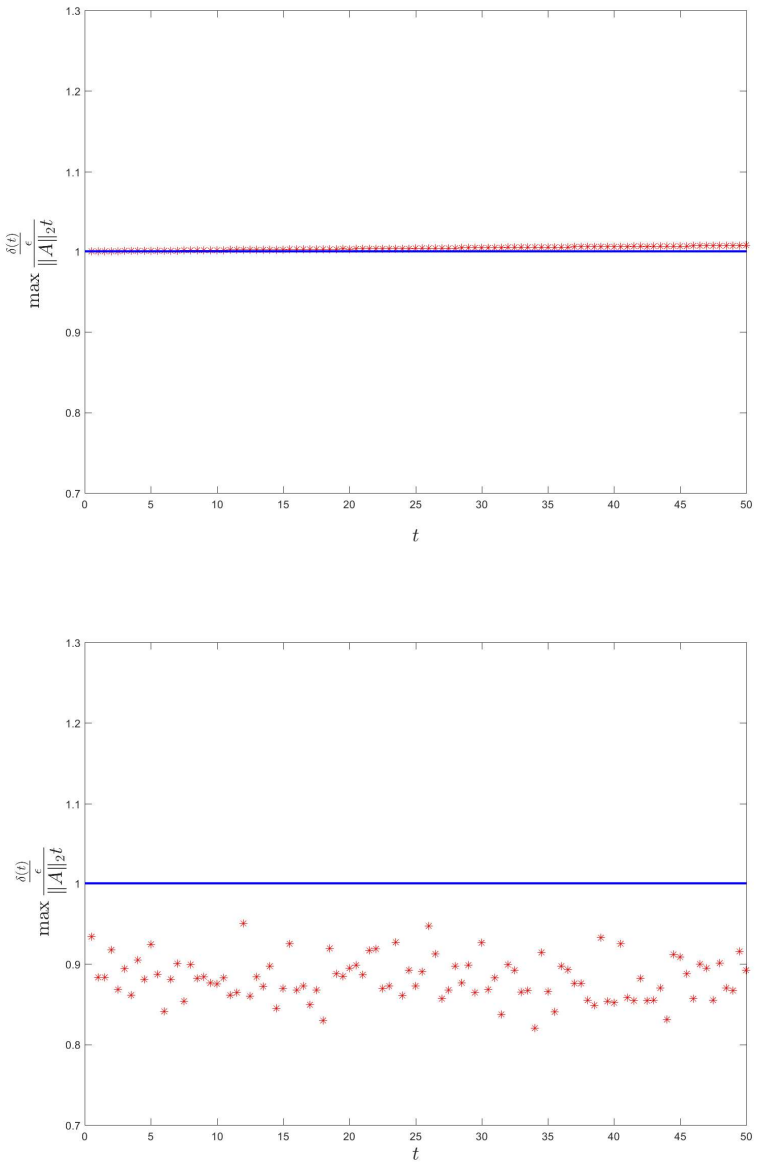

Figure 3.1: For the skew symmetric matrices of Example 3.3.1, maximum value of $\frac{\frac{\delta(t)}{\epsilon}}{\|A\|_2 t}$ in 10000 random selections the matrix $\widehat{B}$, for t varying from 0 to 50 with step 0.5. The maximum values are the red points. The blue line is the constant value 1. Upper part: 2×2 matrix. Lower part: 4×4 matrix.

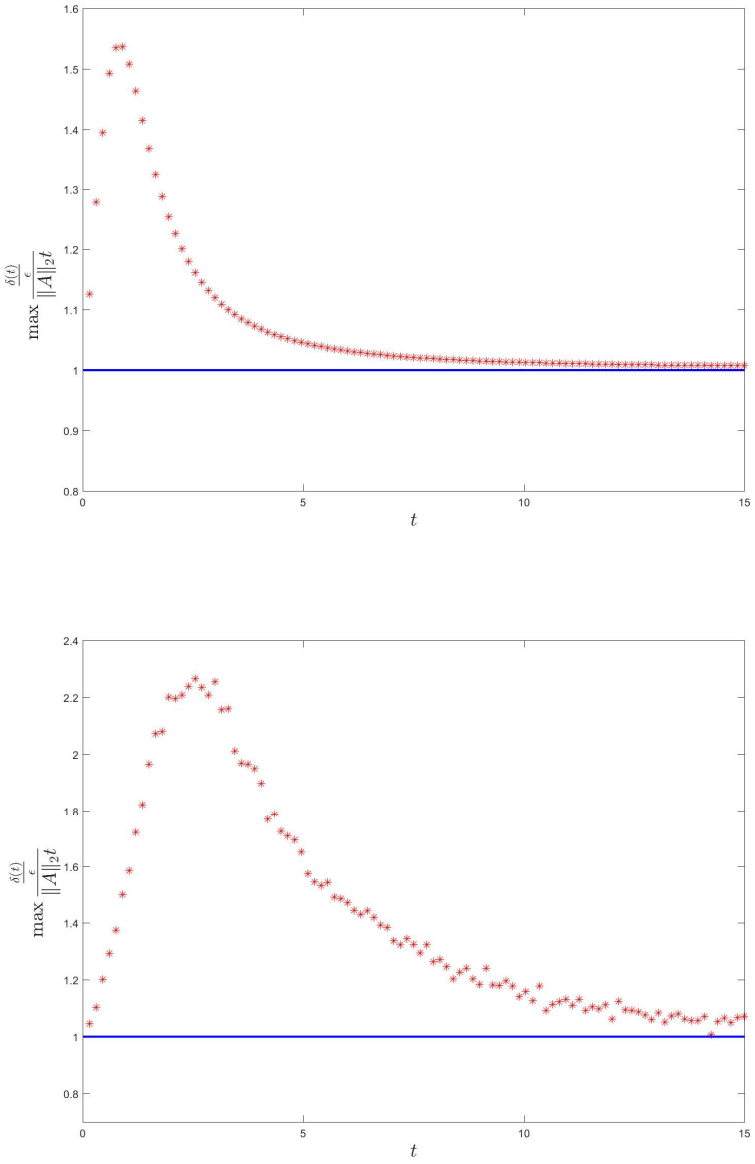

Figure 3.2: For the symmetric matrices of Example 3.3.2, maximum value of $\frac{\frac{\delta(t)}{\epsilon}}{\|A\|_2 t}$ in 10000 random selections the matrix $\widehat{B}$, for t varying from 0 to 15 with step 0.15. The maximum values are the red points. The blue line is the constant value 1. Upper part: 2×2 matrix. Lower part: 4×4 matrix.

For the initial value $y_0 = (1, 1)$, as t varies, the maximum of the values (3.3.1) is always close to 1. *Since y_0 stays in the rightmost eigenspace, we have the same situation of the case $q = 1$, namely the condition number is equal to $\|A\|_2 t$ (see Theorem 3.1.5).*

For the initial value $y_0 = (1, -1)$, as t varies, the maximum of the values (3.3.1) does not tend asymptotically to 1, *but it grows indefinitely, by confirming Theorem 3.2.2, case $j^* > 1$.*

In Figure 3.4, for the 4×4 matrix and for t in a uniform mesh over the interval $[0, 15]$, we plot the maximum of the values

$$\frac{\log \frac{\delta(t)}{\varepsilon}}{(r_1 - r_{j^*})t} = \frac{\log \overline{K}_2(t, A, y_0, \widehat{B})}{(r_1 - r_{j^*})t} + o(1), \ \epsilon \to 0, \tag{3.3.2}$$

over 10000 *random selections of matrix $\widehat{B}$. We consider the initial values $y_0 = (1, 1, -1, 1)$, which is eigenvector of the eigenvalue* 0, *$y_0 = (-1, 1, -1, -1)$, which is eigenvector of the eigenvalue -1, and $y_0 = (1, -1, -1, -1)$, which is eigenvector of the eigenvalue -2. For these three initial values, we have $j^* = 2, 3, 4$, respectively.*

For all initial values, as t varies, the maximum of the values (3.3.2) tends asymptotically to 1, *by confirming Remark 3.2.3.*

3.3.1 Behavior of the condition number for a non-normal matrix

Example 3.3.3. *This example shows the behavior of the ration $\frac{\delta(t)}{\epsilon}$ for two non-normal matrices A. The matrices are taken from MATLAB gallery test in the following manner*

- *$A = gallery('lesp', n)$ with dimension $n = 10$.*
- *$A = -gallery('parter', n)$ with dimension $n = 10$.*

We take $\epsilon = 10^{-4}$. In figure 3.5 *for both these matrices and for t in a uniform mesh over the interval $[0, 15]$, we plot the maximum of the values given by* (3.3.1) *in 10000 random selection of the unit matrix $\hat{B}$. We consider the initial value y_0=(1 2 4 5 6 7 2 4 6 3) for the matrix 'lesp' and y_0=(9 10 2 10 7 1 3 6 10 10) for the matrix 'parter'. In both cases we observe the maximum of the values* (3.3.1) *remain close to 1 as t varies.*

We conclude this section by illustrating the procedure of the random selection of the unit matrix $\widehat{B}$, namely the random selection of the direction of perturbation.

Fixed the order n of the matrix, we construct the Singular Value Decomposition

$$\widehat{B} = UTV$$

of the matrix $\widehat{B}$, where U and V are $n \times n$ randomly selected orthonormal matrices and T is a $n \times n$ diagonal matrix with diagonal $(\sigma_1, \sigma_2, \ldots, \sigma_n)$, where $\sigma_1 = 1$ and $\sigma_2, \ldots, \sigma_n \in [0, 1]$ are randomly selected. Our computations are implemented in MATLAB and, for the random selections of U, V and T, we use:

$$\begin{aligned} U &= orth(rand(n)) \\ V &= orth(rand(n)) \\ T &= diag([1, rand(1, n-1)]), \end{aligned}$$

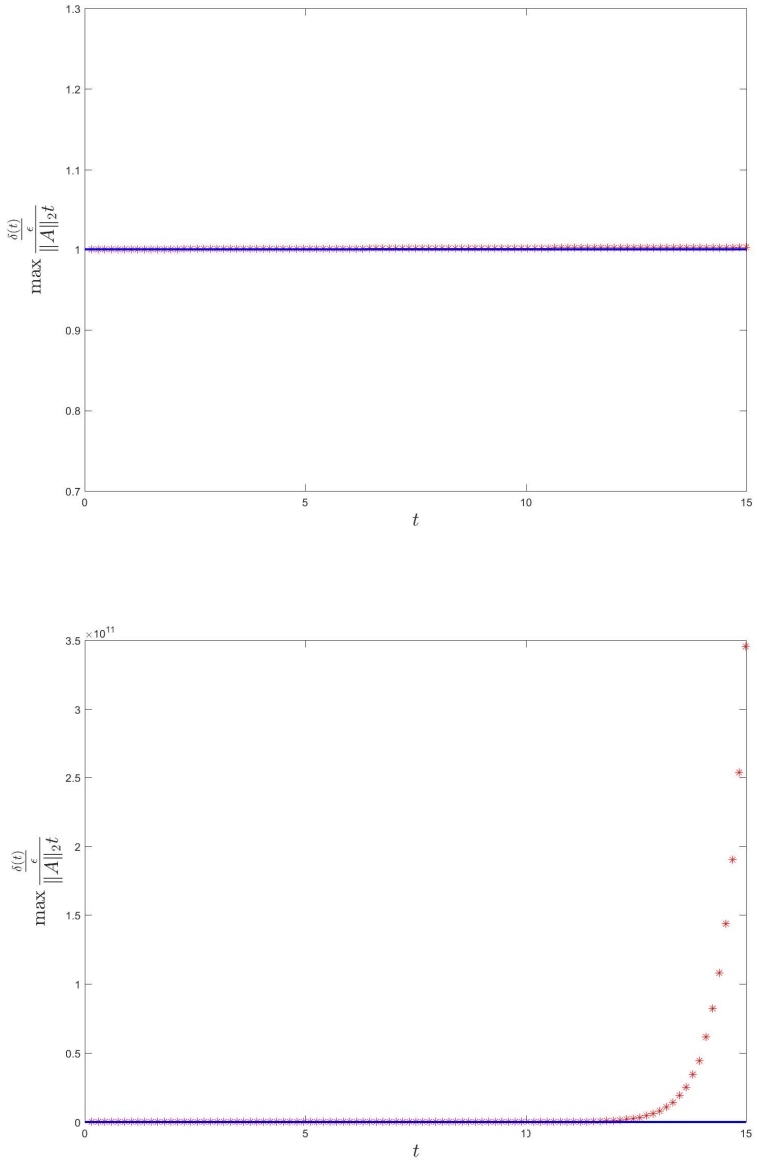

Figure 3.3: For the 2×2 symmetric matrix of Example 3.3.2, maximum value of $\frac{\frac{\delta(t)}{\epsilon}}{\|A\|_2 t}$ in 10000 random selections the matrix $\widehat{B}$, for t varying from 0 to 15 with step 0.15. The maximum values are the red points. The blue line is the constant value 1. Upper part: $y_0 = (1, 1)$. Lower part: $y_0 = (1, -1)$.

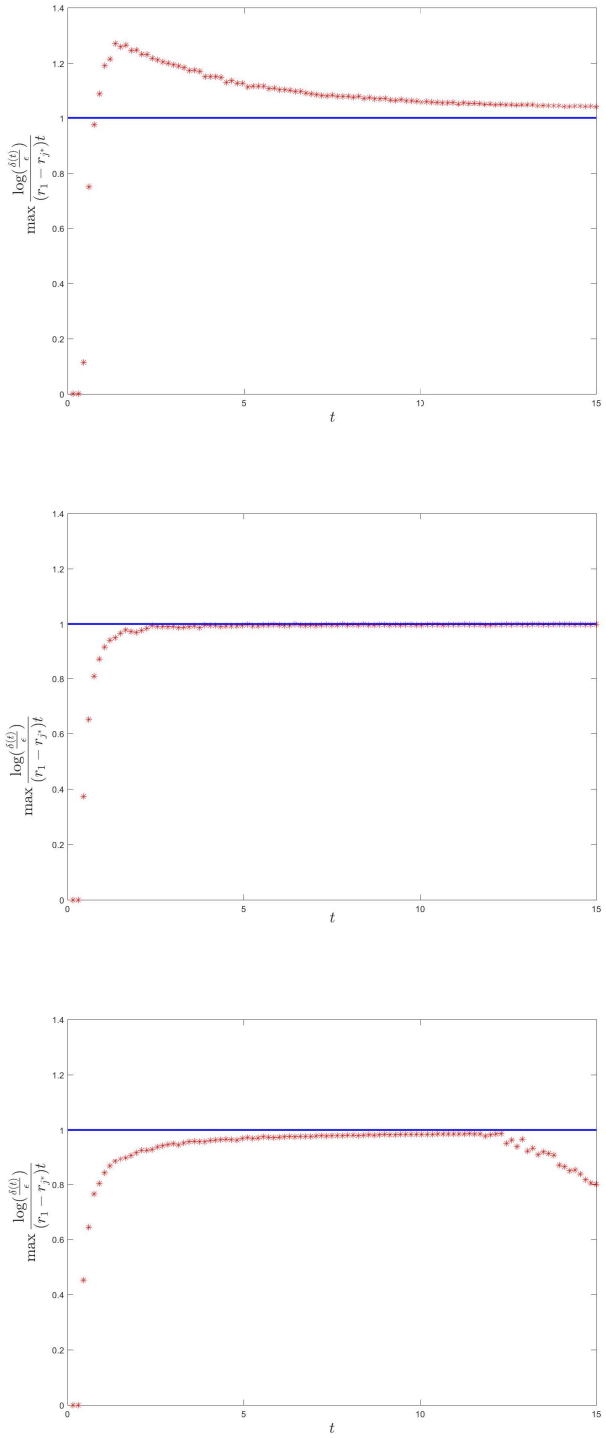

Figure 3.4: For the 4×4 matrix of Example 3.3.2, maximum value of $\frac{\log \frac{\delta(t)}{\epsilon}}{(r_j - r_{j^*})t}$ in 10000 random selections the matrix $\widehat{B}$, for t varying from 0 to 15 with step 0.15. The maximum values are the red points. The blue line is the constant value 1. Upper part: $y_0 = (1, 1, -1, 1)$ and $j^* = 2$. Middle part: $y_0 = (-1, 1, -1, -1)$ and $j^* = 3$. Lower part: $y_0 = (1, -1, -1, -1)$ and $j^* = 4$.

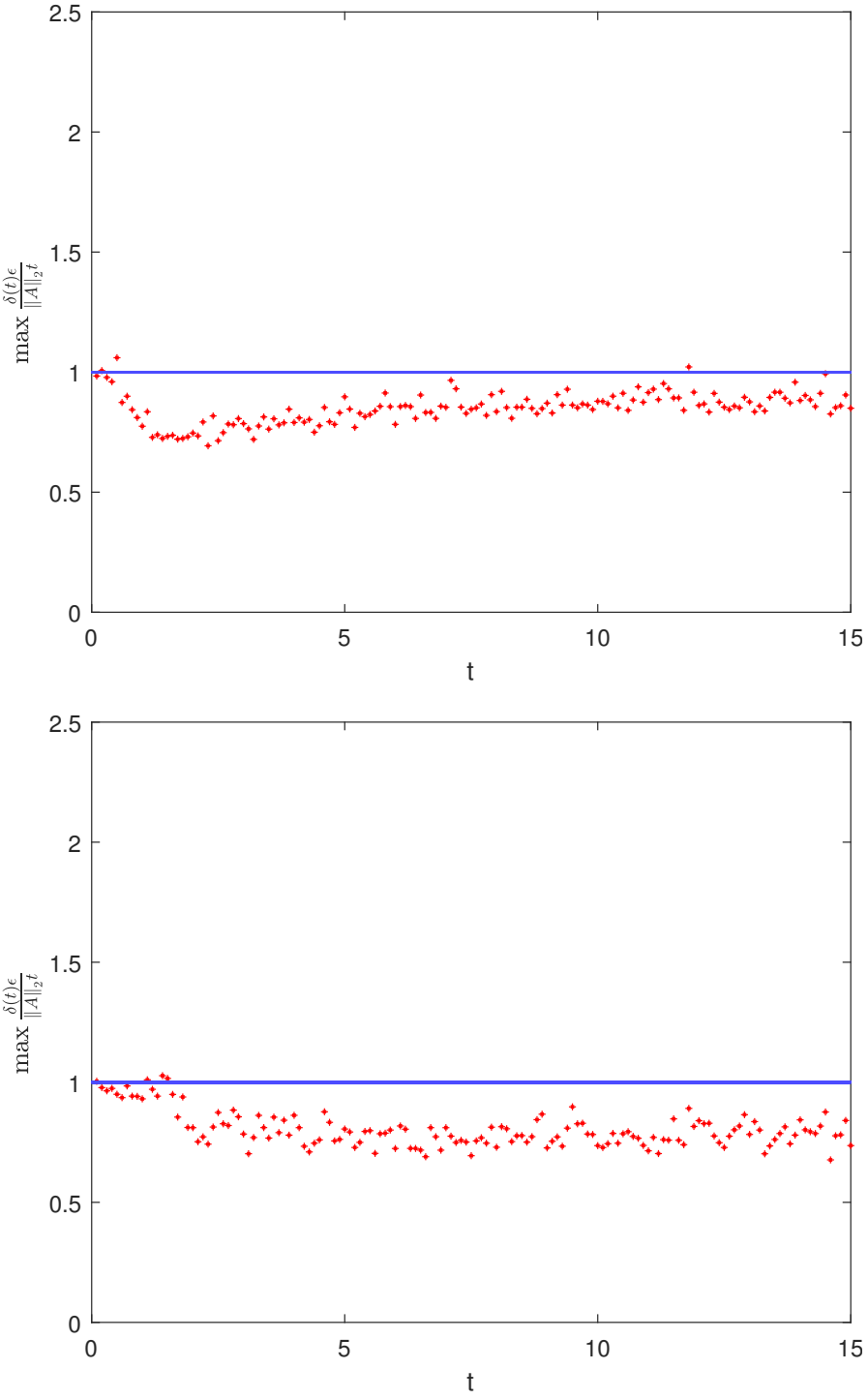

Figure 3.5: For the non normal matrices of Example 3.3.3, maximum value of $\frac{\frac{\delta(t)}{\epsilon}}{\|A\|_2 t}$ in 10000 random selections the matrix $\widehat{B}$, for t varying from 0 to 15 with step 0.1. The maximum values are the red points. The blue line is the constant value 1. Upper part: 10×10 matrix constructed by using $A = gallery('lesp', 10)$. Lower part: 10×10 matrix constructed by using $A = -gallery('parter', 10)$.

where the MATLAB function orth(C) computes a matrix whose columns are an orthonormal basis of the range of C, and the MATLAB function rand(p, q) computes a $p \times q$ matrix of uniformly distributed elements in $[0, 1]$ (rand(p) is rand(p,p)).

By constructing the matrix $\widehat{B}$ as

$$\widehat{B} = \frac{B}{\|B\|_2},$$

where B is obtained in MATLAB by

$$B = rand(n),$$

does not give good results, since this procedure misses some directions of perturbation.

3.4 Conclusion

In this chapter, we have studied the conditioning of the problem

$$A \mapsto e^{tA} y_0,$$

namely how a perturbation in the matrix $A \in \mathbb{R}^{n \times n}$ propagates to $e^{tA} y_0$. We considered the case of a normal matrix A, perturbed to a possibly non-normal matrix, and three condition numbers have been analyzed:

- the condition number $\overline{K}_2(t, A, y_0, \widehat{B})$ with direction of perturbation defined in (3.0.10);
- the condition number $\overline{K}_2(t, A, y_0)$ defined in (3.0.12);
- the condition number $\overline{K}_2(t, A)$ independent of the data defined in (3.0.15).

The euclidean norm is used as vector norm $\|\cdot\|$ and the spectral norm $\|\cdot\|_2$ is used as matrix norm. The spectrum of the normal matrix A has been partitioned by decreasing real parts in the subsets $\Lambda_1, \ldots, \Lambda_q$, where the eigenvalues in Λ_j, $j = 1, \ldots, q$, have real part r_j, and $r_1 > \cdots > r_q$ holds. We denoted by j^* the minimum index in $\{1, \ldots, q\}$ such that y_0 has a non-zero component on the sum of the eigenspaces relevant to the eigenvalues in Λ_j. The generic situation for y_0 is $j^* = 1$.

Regarding the condition number $\overline{K}_2(t, A, y_0)$, we have obtained the following results:

- if A is shifted skew-symmetric, then $\overline{K}_2(t, A, y_0)$ is equal to $\|A\|_2 t$.
- If A is not shifted skew-symmetric and $j^* = 1$, then $\overline{K}_2(t, A, y_0)$ is asymptotically, as $t \to +\infty$, equal to $\|A\|_2 t$.
- If A is not shifted skew symmetric and $j^* > 1$, then $\overline{K}_2(t, A, y_0)$ grows exponentially in t and $\log \overline{K}_2(t, A, y_0)$ is asymptotically, as $t \to +\infty$, equal to $(r_1 - r_{j^*})t$.

Regarding the condition number $\overline{K}_2(t, A)$ independent of the data, we obtained the following result:

- $\overline{K}_2(t, A)$ grows exponentially in t and $\log \overline{K}_2(t, A)$ is asymptotically, as $t \to +\infty$, equal to $(r_1 - r_q)t$.

Chapter 4

Perturbations in the initial value: a componentwise relative error analysis

In chapter 2, we have studied the conditioning of the problem

$$y_0 \mapsto e^{tA} y_0,$$

i.e how the error

$$\varepsilon = \frac{\|\widetilde{y}_0 - y_0\|}{\|y_0\|}$$

is magnified in the error

$$\delta(t) = \frac{\|\tilde{y}(t) - y(t)\|}{\|y(t)\|},$$

where $\widetilde{y}_0$ is the perturbation of y_0 and $\widetilde{y}(t)$ is the perturbed solution.
In some cases, we can be interested in the relative errors

$$\delta_l(t) = \frac{|\widetilde{y}_l(t) - y_l(t)|}{|y_l(t)|},\ l = 1, \ldots, n \tag{4.0.1}$$

of the perturbed solution components.
These componentwise relative errors can be very different from the normwise relative error $\delta(t)$. Indeed, we can have a small $\delta(t)$, but some large componentwise relative error $\delta_l(t)$. Viceversa, if all the componentwise relative errors $\delta_l(t)$ are small, $\delta(t)$ is also small (see Remark 1.3.1).
Measuring relative errors by means of the norm has a few drawbacks. One of its drawbacks is that normwise errors cannot exactly convey the effect of perturbation in small parts of the data, (either input or output). Since the norm measures the global size of perturbation. It does not take into account the structure of the data in sense of scaling and/or sparsity.

To handle the problem arising by the use of a norm in the relative error, two types of condition numbers, namely mixed condition numbers and componentwise condition numbers, are emerging in the field of numerical linear algebra. For the mixed condition numbers, we use the normwise relative error $\delta(t)$ for the output data and the following componentwise relative error

$$RelErr(y, \tilde{y}) = \max_{i=1,\ldots,n} \frac{|\tilde{y}_i - y_i|}{|y_i|}$$

given in (1.0.1) for the intput data. At first, a few authors such as Skeel [68], Rohn [67] has given explicit expression for such condition numbers. Later the subject has been studied by many other authors for instance see [16, 21, 52, 71, 80]. For the componentwise condition numbers, we use componentwise relative errors both for the input and output data.
In this chapter, we consider yet an other approach where we use the normwise relative error ε for the input data and the componentwise relative errors given by (4.0.1) rather than the relative error given by (1.0.1) for the output. We comment that a componentwise relative errors $\delta_l(t)$ can give more information than the normwise relative error $\delta(t)$. However the error $\delta_l(t)$ has the drawback to become infinite when the component $y_l(t)$ becomes zero. Note that the normwise relative error $\delta(t)$ remains finite if some component (but not all) of the solution becomes zero.

In this chapter, we assume that the matrix A is diagonalizable, that is a generic situation for the matrix A. We study the (relative) conditioning of the problem

$$y_0 \mapsto y_l(t) = e_l^T \mathrm{e}^{tA} y_0, \tag{4.0.2}$$

where e_l^T is the l-th vector of the canonical basis of $\mathbb{R}^n$, for a given arbitrary component index $l = 1, \ldots, n$.
The content of this chapter is substantially the paper [26].

The chapter is organized in the following manner. Section 4.1 is devoted to define two condition numbers. Section 4.2 of the chapter gives the condition numbers for the diagonalizable matrices. The asymptotic behavior of these condition numbers is given by section 4.4. To testify our analysis we make a few numerical tests that are given in section 4.5 of the chapter. Finally, the conclusion is given in section 4.6.

Remark 4.0.1. *Since $y_0 \neq 0$, the error ε is well-defined. On the other hand, it could happen to have $y_l(t_1) = 0$ for some $t_1 \geq 0$. We could consider the error $\delta_l(t_1)$ not defined, or equal to $+\infty$ and indeterminate for $\widetilde{y}_l(t_1) \neq 0$ and $\widetilde{y}_l(t_1) = 0$, respectively. However, we are not interested in studying $\delta_l(t)$ for t close to t_1, because the relative error loses its importance in favor of the absolute error when $y_l(t)$ is close to be zero.*

4.1 Condition numbers

By substituting the expression

$$y_l(t) = e_l^T e^{tA} y_0$$

in the equation (4.0.1), we get

$$\delta_l(t) = \frac{|e_l^T e^{tA}\widetilde{y}_0 - e_l^T e^{tA} y_0|}{|e_l^T e^{tA} y_0|},$$

and by writing the perturbation in y_0 as $\widetilde{y}_0 = y_0 + \varepsilon \|y_0\| \widehat{z}_0$, where $\widehat{z}_0 \in \mathbb{R}^n$ is a unit vector, i.e. $\|\widehat{z}_0\| = 1$, we obtain

$$\delta_l(t) = K_l(t, A, y_0, \widehat{z}_0)\varepsilon, \ t \geq 0, \tag{4.1.1}$$

where

$$K_l(t, A, y_0, \widehat{z}_0) = \frac{|e_l^T \mathrm{e}^{tA}\widehat{z}_0|}{|e_l^T \mathrm{e}^{tA}\widehat{y}_0|} \tag{4.1.2}$$

with $\widehat{y}_0 = \frac{y_0}{\|y_0\|}$. We define $K_l(t, A, y_0, \widehat{z}_0)$ as the *condition number with direction of perturbation* of the problem (4.0.2). The formula (4.1.1) is of theoretical interest and, from the practical point of view when there is no information about the direction of perturbation, we can write

$$\delta_l(t) \leq K_l(t, A, y_0)\varepsilon, \ t \geq 0,$$

where

$$K_l\left(t, A, y_0\right) := \sup_{\substack{\widehat{z}\in\mathbb{R}^n \\ \|\widehat{z}\|=1}} K_l\left(t, A, y_0, \widehat{z}_0\right) = \frac{\left\|e_l^T \mathrm{e}^{tA}\right\|}{|e_l^T \mathrm{e}^{tA}\widehat{y}_0|} \tag{4.1.3}$$

with $\left\|e_l^T \mathrm{e}^{tA}\right\|$ the matrix norm of the row vector $e_l^T \mathrm{e}^{tA}$ relevant to the vector norm $\|\cdot\|$. We define $K_l\left(t, A, y_0\right)$ as the *condition number* of the problem (4.0.2).

In the next section, we analyze the conditions numbers (4.1.2) and (4.1.3) by assuming that A is diagonalizable. This is a generic situation for the matrix A. A generic situation for A or y_0 or $\widehat{z}_0$, means that A satisfy a property which is generic according to measure theory definition (the complementary property corresponds to a zero measure subset) or the topological definition (the property corresponds to a dense open subset) given for example in [10, ch.2] see also [5, 48, 50, 82]. Roughly speaking, a generic situation considers " typical" not "exceptional" cases.

4.2 Condition numbers for a diagonalizable matrix

Let A be a diagonalizable matrix. We partition the spectrum Λ of A in the same manner as in Chapters 2 and 3.
For $i = 1, \ldots, p$, let $v^{(i,1)}, \ldots, v^{(i,\nu_i)}$ be a basis for the eigenspace of the eigenvalue λ_i,

where ν_i is the multiplicity of λ_i. Let $V \in \mathbb{C}^{n\times n}$ be the matrix of columns $v^{(i,1)}, \ldots, v^{(i,\nu_i)}$, $i = 1, \ldots, p$, and let $W = V^{-1}$. We denote the rows of W by $w^{(i,1)}, \ldots, w^{(i,\nu_i)}$, $i = 1, \ldots, p$, correspondly to the columns of V. Of course, we have $A = VDV^{-1}$, where D is the diagonal matrix with the eigenvalues of A on the diagonal.
Observe that the transposes of rows of W are eigenvectors of A^T. In fact, we have

$$\begin{aligned} A &= VDV^{-1} = VDW \\ A^T &= W^T D(W^{-1})^T \\ A^T &= W^T D(W^T)^{-1}. \end{aligned}$$

For $i = 1, \ldots, p$, the projection $P_i \in \mathbb{C}^{n\times n}$ on the eigenspace of λ_i is given by

$$P_i = V^{(i)} W^{(i)}, \tag{4.2.1}$$

where $V^{(i)} \in \mathbb{C}^{n\times \nu_i}$ is the matrix of columns $v^{(i,1)}, \ldots, v^{(i,\nu_i)}$ and $W^{(i)} \in \mathbb{C}^{\nu_i \times n}$ is the matrix of rows $w^{(i,1)}, \ldots, w^{(i,\nu_i)}$.
Finally, for $i = 1, \ldots, p$, let ω_i be the imaginary part of the eigenvalue λ_i.

The next theorem gives expressions for the condition numbers $K_l(t, A, y_0, \widehat{z}_0)$ and $K_l(t, A, y_0)$, $l = 1, \ldots, p$. Here and in the following, $\sqrt{-1}$ denotes the imaginary unit.

Theorem 4.2.1. *We have*

$$K_l(t, A, y_0, \widehat{z}_0) = \frac{\left| \sum_{j=1}^{q} \mathrm{e}^{(r_j - r_1)t} \sum_{\lambda_i \in \Lambda_j} \mathrm{e}^{\sqrt{-1}\omega_i t} e_l^T P_i \widehat{z}_0 \right|}{\left| \sum_{j=1}^{q} \mathrm{e}^{(r_j - r_1)t} \sum_{\lambda_i \in \Lambda_j} \mathrm{e}^{\sqrt{-1}\omega_i t} e_l^T P_i \widehat{y}_0 \right|} \tag{4.2.2}$$

and

$$K_l(t, A, y_0) = \frac{\left\| \sum_{j=1}^{q} \mathrm{e}^{(r_j - r_1)t} \sum_{\lambda_i \in \Lambda_j} \mathrm{e}^{\sqrt{-1}\omega_i t} e_l^T P_i \right\|}{\left| \sum_{j=1}^{q} \mathrm{e}^{(r_j - r_1)t} \sum_{\lambda_i \in \Lambda_j} \mathrm{e}^{\sqrt{-1}\omega_i t} e_l^T P_i \widehat{y}_0 \right|}. \tag{4.2.3}$$

Proof. Since A is diagonalizable, we have

$$\mathrm{e}^{tA} = \sum_{\lambda_i \in \Lambda} \mathrm{e}^{\lambda_i t} P_i = \sum_{j=1}^{q} \mathrm{e}^{r_j t} \sum_{\lambda_i \in \Lambda_j} \mathrm{e}^{\sqrt{-1}\omega_i t} P_i \tag{4.2.4}$$

Substituting (4.2.4), in equation (4.1.2), we obtain

$$K_l(t, A, y_0, \widehat{z}_0) = \frac{\left| \sum_{j=1}^{q} \mathrm{e}^{r_j t} \sum_{\lambda_i \in \Lambda_j} \mathrm{e}^{\sqrt{-1}\omega_i t} e_l^T P_i \widehat{z}_0 \right|}{\left| \sum_{j=1}^{q} \mathrm{e}^{r_j t} \sum_{\lambda_i \in \Lambda_j} \mathrm{e}^{\sqrt{-1}\omega_i t} e_l^T P_i \widehat{y}_0 \right|} = \frac{\left| \sum_{j=1}^{q} \mathrm{e}^{(r_j - r_1)t} \sum_{\lambda_i \in \Lambda_j} \mathrm{e}^{\sqrt{-1}\omega_i t} e_l^T P_i \widehat{z}_0 \right|}{\left| \sum_{j=1}^{q} \mathrm{e}^{(r_j - r_1)t} \sum_{\lambda_i \in \Lambda_j} \mathrm{e}^{\sqrt{-1}\omega_i t} e_l^T P_i \widehat{y}_0 \right|}.$$

Now, to obtain the condition number independent of direction of perturbation, we substitute expression given by (4.2.4) in (4.1.3) and we get

$$K_l(t,A,y_0)=\frac{\left\|\sum\limits_{j=1}^{q}\mathrm{e}^{r_jt}\sum\limits_{\lambda_i\in\Lambda_j}\mathrm{e}^{\sqrt{-1}\omega_it}e_l^TP_i\right\|}{\left|\sum\limits_{j=1}^{q}\mathrm{e}^{r_jt}\sum\limits_{\lambda_i\in\Lambda_j}\mathrm{e}^{\sqrt{-1}\omega_it}e_l^TP_i\widehat{y}_0\right|}=\frac{\left\|\sum\limits_{j=1}^{q}\mathrm{e}^{(r_j-r_1)t}\sum\limits_{\lambda_i\in\Lambda_j}\mathrm{e}^{\sqrt{-1}\omega_it}e_l^TP_i\right\|}{\left|\sum\limits_{j=1}^{q}\mathrm{e}^{(r_j-r_1)t}\sum\limits_{\lambda_i\in\Lambda_j}\mathrm{e}^{\sqrt{-1}\omega_it}e_l^TP_i\widehat{y}_0\right|}.$$

□

Remark 4.2.1.

1. *In (4.2.2) and (4.2.3) all the exponentials* $\mathrm{e}^{(r_j-r_1)t}$, $j=2,\dots,q$, *have* $r_j-r_1<0$ *so they are vanishing functions of* t.
2. *For a pair of complex conjugate eigenvalues* λ_i *and* $\lambda_k=\overline{\lambda_i}$, *we obtain (since* $P_k=\overline{P_i}$, *where* $\overline{P_i}$ *is the matrix whose elements are the complex conjugates of the elements of* P_i*)*

$$\mathrm{e}^{\sqrt{-1}\omega_it}e_l^TP_i+\mathrm{e}^{\sqrt{-1}\omega_kt}e_l^TP_k=\mathrm{e}^{\sqrt{-1}\omega_it}e_l^TP_i+\overline{\mathrm{e}^{\sqrt{-1}\omega_it}e_l^TP_i}=2\mathrm{Re}\left(\mathrm{e}^{\sqrt{-1}\omega_it}e_l^TP_i\right).$$

Then, in (4.2.2) and (4.2.3) we have, for $j=1,\dots,q$ *and* $\lambda_i\in\Lambda_j$,

$$\sum_{\lambda_i\in\Lambda_j}\mathrm{e}^{\sqrt{-1}\omega_it}e_l^TP_i=\sum_{\substack{\lambda_i\in\Lambda_j\\ \lambda_i\in\mathbb{R}}}e_l^TP_i+2\sum_{\substack{\lambda_i\in\Lambda_j\\ \omega_i>0}}\mathrm{Re}\left(\mathrm{e}^{\sqrt{-1}\omega_it}e_l^TP_i\right).$$

4.3 Asymptotic analysis

The next theorem describes the asymptotic behavior, as $t\to+\infty$, of the condition numbers $K_l(t,A,y_0,\widehat{z}_0)$ and $K_l(t,A,y_0)$, $l=1,\dots,p$. We use the notation

$$f(t)\sim g(t),\ t\to+\infty,\ \text{for}\ \lim_{t\to+\infty}\frac{f(t)}{g(t)}=1,$$

already used in the previous chapters.

Theorem 4.3.1. *We have*

$$K_l(t,A,y_0,\widehat{z}_0)\sim\mathrm{e}^{\left(r_{j^{**}}-r_{j^*}\right)t}\frac{\left|\sum\limits_{\lambda_i\in\Lambda_{j^{**}}}\mathrm{e}^{\sqrt{-1}\omega_it}e_l^TP_i\widehat{z}_0\right|}{\left|\sum\limits_{\lambda_i\in\Lambda_{j^*}}\mathrm{e}^{\sqrt{-1}\omega_it}e_l^TP_i\widehat{y}_0\right|},\ t\to+\infty,\qquad(4.3.1)$$

$$K_l(t,A,y_0)\sim\mathrm{e}^{\left(r_{\overline{j}^*}-r_{j^*}\right)t}\frac{\left\|\sum\limits_{\lambda_i\in\Lambda_{\overline{j}^*}}\mathrm{e}^{\sqrt{-1}\omega_it}e_l^TP_i\right\|}{\left|\sum\limits_{\lambda_i\in\Lambda_{j^*}}\mathrm{e}^{\sqrt{-1}\omega_it}e_l^TP_i\widehat{y}_0\right|},\ t\to+\infty,\qquad(4.3.2)$$

where

$$j^* := \min\left\{j \in \{1,\dots,q\} : e_l^T P_i \widehat{y}_0 \neq 0 \text{ for some } \lambda_i \in \Lambda_j\right\}$$
$$j^{**} := \min\left\{j \in \{1,\dots,q\} : e_l^T P_i \widehat{z}_0 \neq 0 \text{ for some } \lambda_i \in \Lambda_j\right\}$$
$$\overline{j}^* := \min\left\{j \in \{1,\dots,q\} : e_l^T P_i \neq 0 \text{ for some } \lambda_i \in \Lambda_j\right\}.$$

Proof. For the numerator or denominator in (4.2.2) we have, with $u = \widehat{z}_0$ and $j(u) = j^{**}$ or $u = \widehat{y}_0$ and $j(u) = j^*$,

$$\left|\sum_{j=1}^{q} \mathrm{e}^{(r_j - r_1)t} \sum_{\lambda_i \in \Lambda_j} \mathrm{e}^{\sqrt{-1}\omega_i t} e_l^T P_i u\right| = \left|\sum_{\lambda_i \in \Lambda_{j(u)}} \mathrm{e}^{(r_{j(u)} - r_1)t} \mathrm{e}^{\sqrt{-1}\omega_i t} e_l^T P_i u\right| (1+E)$$

with

$$|E| \leq \frac{\left|\sum\limits_{j=j(u)+1}^{q} \mathrm{e}^{(r_j - r_1)t} \sum\limits_{\lambda_i \in \Lambda_j} \mathrm{e}^{\sqrt{-1}\omega_i t} e_l^T P_i u\right|}{\left|\mathrm{e}^{(r_{j(u)} - r_1)t} \sum\limits_{\lambda_i \in \Lambda_{j(u)}} \mathrm{e}^{\sqrt{-1}\omega_i t} e_l^T P_i u\right|} = \frac{\left|\sum\limits_{j=j(u)+1}^{q} \mathrm{e}^{(r_j - r_{j(u)})t} \sum\limits_{\lambda_i \in \Lambda_j} \mathrm{e}^{\sqrt{-1}\omega_i t} e_l^T P_i u\right|}{\left|\sum\limits_{\lambda_i \in \Lambda_{j(u)}} \mathrm{e}^{\sqrt{-1}\omega_i t} e_l^T P_i u\right|}.$$

Now, by letting $t \to +\infty$ (see point 1 in Remark 4.3.1 below), we obtain (4.3.1). Similarly, we obtain (4.3.2). □

Remark 4.3.1.

1. *In (4.3.1) we assume there exists $\sigma > 0$ such that*

$$\mathcal{A}_\sigma = \left\{t \geq 0 : \left|\sum_{\lambda_i \in \Lambda_{j^{**}}} \mathrm{e}^{\sqrt{-1}\omega_i t} e_l^T P_i \widehat{z}_0\right| \geq \sigma \text{ and } \left|\sum_{\lambda_i \in \Lambda_{j^*}} \mathrm{e}^{\sqrt{-1}\omega_i t} e_l^T P_i \widehat{y}_0\right| \geq \sigma\right\} \quad (4.3.3)$$

   has $+\infty$ as an accumulation point. In (4.3.1), we consider $t \to +\infty$ with $t \in \mathcal{A}_\sigma$. Analogously, in (4.3.2) we assume there exists $\sigma > 0$ such that

$$\mathcal{B}_\sigma = \left\{t \geq 0 : \left|\sum_{\lambda_i \in \Lambda_{j^*}} \mathrm{e}^{\sqrt{-1}\omega_i t} e_l^T P_i \widehat{y}_0\right| \geq \sigma\right\} \quad (4.3.4)$$

   has $+\infty$ as an accumulation point. In (4.3.2), we consider $t \to +\infty$ with $t \in \mathcal{B}_\sigma$.

2. *We have $\overline{j}^* \leq j^*$ and then $r_{\overline{j}^*} - r_{j^*} \geq 0$ in the exponential $\mathrm{e}^{(r_{\overline{j}^*} - r_{j^*})t}$ in (4.3.2) and so it is an increasing function of t .*

3. *A generic situation for A, y_0 and $\widehat{z}_0$ is $j^* = j^{**} = \overline{j}^* = 1$, where we have, as $t \to +\infty$,*

$$K_l(t, A, y_0, z_0) \sim \frac{\left|\sum\limits_{\lambda_i \in \Lambda_1} \mathrm{e}^{\sqrt{-1}\omega_i t} e_l^T P_i \widehat{z}_0\right|}{\left|\sum\limits_{\lambda_i \in \Lambda_1} \mathrm{e}^{\sqrt{-1}\omega_i t} e_l^T P_i \widehat{y}_0\right|}$$

$$K_l(t,A,y_0) \sim \frac{\left\| \sum\limits_{\lambda_i\in\Lambda_1} \mathrm{e}^{\sqrt{-1}\omega_i t} e_l^T P_i \right\|}{\left| \sum\limits_{\lambda_i\in\Lambda_1} \mathrm{e}^{\sqrt{-1}\omega_i t} e_l^T P_i \widehat{y}_0 \right|}.$$

The next theorem considers the generic situation for A, y_0 and $\widehat{z}_0$ described in point 3 in the previous remark, namely $j^* = j^{**} = \overline{j}^* = 1$.

Theorem 4.3.2. *Suppose that A is diagonalizable and it has a unique real eigenvalue λ_1 of multiplicity one, or a unique pair λ_1 and $\lambda_2 = \overline{\lambda_1}$ of complex conjugate eigenvalues of multiplicity one, as rightmost eigenvalues. Let v be an eigenvector of λ_1 and let w be the first row of $W = V^{-1}$, V being the matrix of the eigenvectors with v as first column. Let $l = 1,\dots,n$ such that $v_l \neq 0$. If $w\widehat{y}_0 \neq 0$ and $w\widehat{z}_0 \neq 0$, then, as $t \to +\infty$,*

$$K_l(t,A,y_0,\widehat{z}_0) \to \frac{|w\widehat{z}_0|}{|w\widehat{y}_0|} \text{ and } K_l(t,A,y_0) \to \frac{\|w\|}{|w\widehat{y}_0|} \tag{4.3.5}$$

when the rightmost eigenvalue is the real eigenvalue and

$$K_l(t,A,y_0,\widehat{z}_0) \sim \frac{\left|\mathrm{Re}\left(\mathrm{e}^{\sqrt{-1}\omega_1 t} v_l w\right)\widehat{z}_0\right|}{\left|\mathrm{Re}\left(\mathrm{e}^{\sqrt{-1}\omega_1 t} v_l w\right)\widehat{y}_0\right|} \quad \text{and} \quad K_l(t,A,y_0) \sim \frac{\left\|\mathrm{Re}\left(\mathrm{e}^{\sqrt{-1}\omega_1 t} v_l w\right)\right\|}{\left|\mathrm{Re}\left(\mathrm{e}^{\sqrt{-1}\omega_1 t} v_l w\right)\widehat{y}_0\right|} \tag{4.3.6}$$

when the rightmost eigenvalues are the complex conjugate pair.

Proof. We have (see (4.2.1)) $P_1 = vw$ and then $e_l^T P_1 = v_l w \neq 0$, $e_l^T P_1 \widehat{y}_0 = v_l w \widehat{y}_0 \neq 0$ and $e_l^T P_1 \widehat{z}_0 = v_l w \widehat{z}_0 \neq 0$. So $j^* = j^{**} = \overline{j}^* = 1$ and then (see point 3 in Remark 4.3.1), as $t \to +\infty$,

$$K_l(t,A,y_0,z_0) \sim \frac{\left|e_l^T P_1 \widehat{z}_0\right|}{\left|e_l^T P_1 \widehat{y}_0\right|} = \frac{|w\widehat{z}_0|}{|w\widehat{y}_0|}$$

and

$$K_l(t,A,y_0) \sim \frac{\left\|e_l^T P_1\right\|}{\left|e_l^T P_1 \widehat{y}_0\right|} = \frac{\|w\|}{|w\widehat{y}_0|}$$

when the rightmost eigenvalue is the real eigenvalue.
When the rightmost eigenvalues is the complex conjugate pair

$$K_l(t,A,y_0,z_0) \sim \frac{\left|\left(\mathrm{e}^{\sqrt{-1}\omega_1 t} e_l^T P_1 + \mathrm{e}^{\sqrt{-1}\omega_2 t} e_l^T P_2\right)\widehat{z}_0\right|}{\left|\left(\mathrm{e}^{\sqrt{-1}\omega_1 t} e_l^T P_1 + \mathrm{e}^{\sqrt{-1}\omega_2 t} e_l^T P_2\right)\widehat{y}_0\right|}$$

Since $\lambda_1 = \overline{\lambda_2}$, and $\mathrm{e}^{\sqrt{-1}\omega_2 t} e_l^T P_2 = \overline{\mathrm{e}^{\sqrt{-1}\omega_1 t} e_l^T P_1}$, we obtain

$$K_l(t,A,y_0,z_0) \sim \frac{\left|\mathrm{Re}\left(\mathrm{e}^{\sqrt{-1}\omega_1 t} v_l w\right)\widehat{z}_0\right|}{\left|\mathrm{Re}\left(\mathrm{e}^{\sqrt{-1}\omega_1 t} v_l w\right)\widehat{y}_0\right|}$$

Similarly, we get

$$K_l(t,A,y_0) \sim \frac{\left\| \mathrm{e}^{\sqrt{-1}\omega_1 t} e_l^T P_1 + \mathrm{e}^{\sqrt{-1}\omega_2 t} e_l^T P_2 \right\|}{\left| \left(\mathrm{e}^{\sqrt{-1}\omega_1 t} e_l^T P_1 + \mathrm{e}^{\sqrt{-1}\omega_2 t} e_l^T P_2 \right) \widehat{y}_0 \right|} = \frac{\left\| \mathrm{Re}\left(\mathrm{e}^{\sqrt{-1}\omega_1 t} v_l w \right) \right\|}{\left| \mathrm{Re}\left(\mathrm{e}^{\sqrt{-1}\omega_1 t} v_l w \right) \widehat{y}_0 \right|}$$

□

Remark 4.3.2.

1. *In the theorem we suppose that A is diagonalizable and it has a unique real eigenvalue of multiplicity one, or a unique pair of complex conjugate eigenvalues of multiplicity one, as rightmost eigenvalues. This is a generic situation for A. Moreover, also $v_l \neq 0$ for any $l = 1, \ldots, n$ is a generic situation for A. Finally, $w\widehat{y}_0 \neq 0$ and $w\widehat{z}_0 \neq 0$ are generic situations for y_0 and $\widehat{z}_0$.*

2. *When the rightmost eigenvalue is the real eigenvalue, there exists $\sigma > 0$ such that $\mathcal{A}_\sigma = \mathbb{R}^+$ and there exists $\sigma > 0$ such that $\mathcal{B}_\sigma = \mathbb{R}^+$ (remind (4.3.3) and (4.3.4)). So in (4.3.5) we can consider $t \to +\infty$ without restrictions on t. Moreover, observe that the limits in (4.3.5) are independent of l (independent of the particular component).*

3. *In (4.3.6), by setting*

$$v_l = |v_l| \mathrm{e}^{\sqrt{-1}\alpha_l}, \ \widehat{w} = \frac{w}{\|w\|} = \left(|\widehat{w}_k| \, \mathrm{e}^{\sqrt{-1}\beta_k} \right)_{k=1,\ldots,n}$$

(observe that $\widehat{w}$ is a unit vector and, if $\| \cdot \|$ is a p-norm, $|\widehat{w}_k| \leq 1$, $k = 1, \ldots, n$) and

$$\widehat{w}\widehat{y}_0 = |\widehat{w}\widehat{y}_0| \, \mathrm{e}^{\sqrt{-1}\gamma(\widehat{y}_0)}, \ \widehat{w}\widehat{z}_0 = |\widehat{w}\widehat{z}_0| \mathrm{e}^{\sqrt{-1}\gamma(\widehat{z}_0)},$$

we can write

$$\frac{\left| \mathrm{Re}\left(\mathrm{e}^{\sqrt{-1}\omega_1 t} v_l w \right) \widehat{z}_0 \right|}{\left| \mathrm{Re}\left(\mathrm{e}^{\sqrt{-1}\omega_1 t} v_l w \right) \widehat{y}_0 \right|} = \frac{\left| \cos\left(\omega_1 t + \alpha_l + \gamma\left(\widehat{z}_0 \right) \right) \right|}{\left| \cos\left(\omega_1 t + \alpha_l + \gamma\left(\widehat{y}_0 \right) \right) \right|} \cdot \frac{|\widehat{w}\widehat{z}_0|}{|\widehat{w}\widehat{y}_0|},$$

and

$$\frac{\left\| \mathrm{Re}\left(\mathrm{e}^{\sqrt{-1}\omega_1 t} v_l w \right) \right\|}{\left| \mathrm{Re}\left(\mathrm{e}^{\sqrt{-1}\omega_1 t} v_l w \right) \widehat{y}_0 \right|} = \frac{\left\| \left(|\widehat{w}_k| \cos\left(\omega_1 t + \alpha_l + \beta_k \right) \right)_{k=1,\ldots,n} \right\|}{\left| \cos\left(\omega_1 t + \alpha_l + \gamma\left(\widehat{z}_0 \right) \right) \right|} \cdot \frac{1}{|\widehat{w}\widehat{y}_0|},$$

So, the long-time oscillations of $K_l(t, A, y_0, \widehat{z}_0)$ and $K_l(t, A, y_0)$ are scaled by the factors $\frac{|\widehat{w}\widehat{z}_0|}{|\widehat{w}\widehat{y}_0|} = \frac{|w\widehat{z}_0|}{|w\widehat{y}_0|}$ and $\frac{1}{|\widehat{w}\widehat{y}_0|} = \frac{\|w\|}{|w\widehat{y}_0|}$, respectively, independent of l (independent of the particular component). Moreover, observe that

$$\begin{aligned} \mathcal{A}_\sigma \ = \ & \{ t \geq 0 : |\cos\left(\omega_1 t + \alpha_l + \gamma\left(\widehat{z}_0 \right) \right)| \cdot |v_l| \cdot |w\widehat{z}_0| \geq \sigma \\ & \qquad \textit{and} \ |\cos\left(\omega_1 t + \alpha_l + \gamma\left(\widehat{y}_0 \right) \right)| \cdot |v_l| \cdot |w\widehat{y}_0| \geq \sigma \} \end{aligned}$$

and

$$\mathcal{B}_\sigma = \{ t \geq 0 : |\cos\left(\omega_1 t + \alpha_l + \gamma\left(\widehat{y}_0 \right) \right)| \cdot |v_l| \cdot |w\widehat{y}_0| \geq \sigma \}$$

Thus, there exists $\sigma > 0$ such that $\mathcal{A}_\sigma$ and $\mathcal{B}_\sigma$ are countable unions of intervals (whose lengths are uniformly away from zero) with $+\infty$ as an accumulation point.

4.4 Numerical test

We show two examples with the matrix A taken from the MATLAB gallery test. We use the euclidean norm $\|\cdot\| = \|\cdot\|_2$ for measuring the relative error of the perturbation of y_0.

In the first example, we consider $A = gallery('lesp', n)$ with dimension $n = 10$. The matrix has ten real eigenvalues: the rightmost is -4.5491. In Figures 4.1 and 4.2, for two different initial values y_0, the graphs of $t \mapsto K_l(t, A, y_0) = \frac{\|e_l^T \mathrm{e}^{tA}\|}{|e_l^T \mathrm{e}^{tA}\widehat{y}_0|}$ (blue line), $l = 1, 2, 3, 4$, are plotted along with the constant value $\frac{\|w\|}{|w\widehat{y}_0|}$ (red line). Just for a comparison, in Figure 4.3 we see the graph of $t \mapsto \|e_l^T \mathrm{e}^{tA}\|$, where $\|e_l^T \mathrm{e}^{tA}\|$ is the worst magnification factor of the absolute error at the time t.

In the second example, we consider as $A = -gallery('parter', n)$ with dimension $n = 10$. The matrix has five complex conjugate pairs of eigenvalues: the rightmost pair is $-0.9066 \pm \sqrt{-1} \cdot 2.7709$. In Figures 4.4 and 4.5, for two different initial values y_0, we see the graphs of $t \mapsto K_l(t, A, y_0)$ (blue line), $l = 1, 2, 3, 4$, along with the graph of $t \mapsto \frac{\|\mathrm{Re}(\mathrm{e}^{\sqrt{-1}\omega_1 t} v_l w)\|}{|\mathrm{Re}(\mathrm{e}^{\sqrt{-1}\omega_1 t} v_l w)\widehat{y}_0|}$ (red line) and the constant value $\frac{\|w\|}{|w\widehat{y}_0|}$ (yellow line). In Figure 4.6, we see the graph of $t \mapsto \|e_l^T \mathrm{e}^{tA}\|$ for the absolute error.
In both examples, the asymptotic behavior described in Theorem 4.3.2 is confirmed. Observe that the peaks of $K_l(t, A, y_0)$ not shown in Figures 4.2 and 4.4 are not of interest, because the components of the solution become zero at the peaks and the important error becomes the absolute error (recall point 1 in Remark 4.0.1). As remarked in point 3 of Remark 4.3.2, the quantity of interest in Figures 4.2 and 4.4 is the scale factor $\frac{\|w\|}{|w\widehat{y}_0|}$ of the oscillations (the constant yellow line).

4.5 Conclusions

In this chapter we have studied the propagation of a perturbation in the initial value along the solution of a linear ODE. A normwise relative error is used for the perturbed initial value and componentwise relative errors are used for the perturbed solution.

- The main result of the paper says that in the generic situation of a linear ODE with a diagonalizable matrix having a real eigenvalue of multiplicity one, or a complex conjugate pair of eigenvalues of multiplicity one, as rightmost eigenvalues, the error in the initial value is magnified in the components of the solution, in the worst case, by the factor $\frac{\|w\|\ \|y_0\|}{|wy_0|}$ over a long-time, where y_0 is the initial value and w is the first row of the inverse of the eigenvectors matrix (i.e. w^T is an eigenvector of A^T relevant to the rightmost real eigenvalue or to the rightmost complex conjugate pair). The magnification factor is the same for all the components. In case of a complex conjugate pair of eigenvalues, as rightmost eigenvalue, oscillations are present in the long-time relative error of the perturbed solution.

- In non-generic situation, the magnification in the components of the solution of

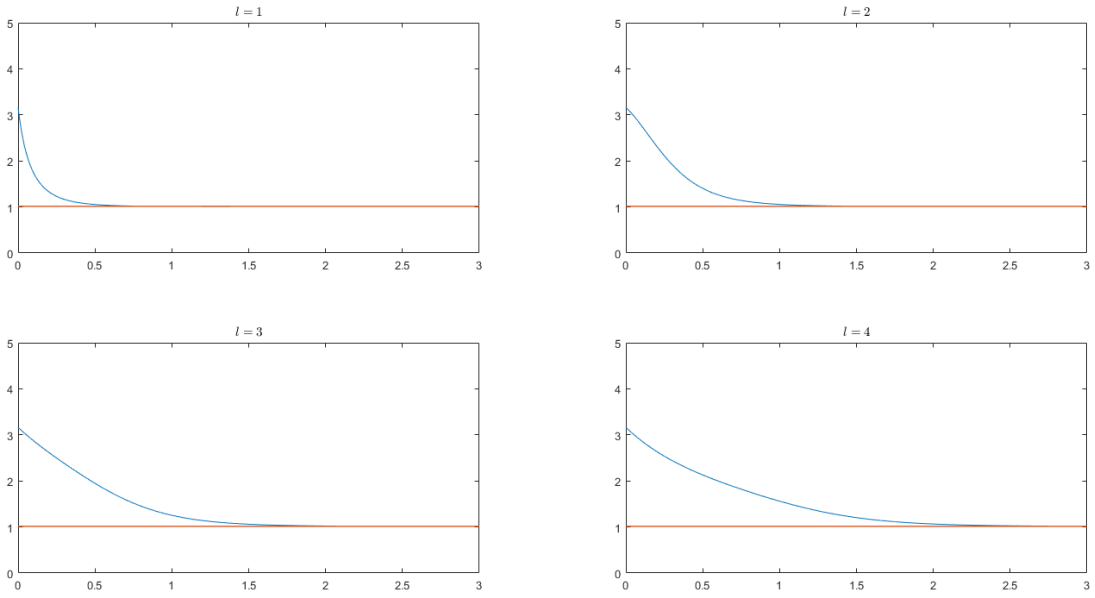

Figure 4.1: $K_l(t, A, y_0)$ (blue line), $l = 1, 2, 3, 4$, along with the constant value $\frac{\|w\|}{|w\widehat{y}_0|} = 1.0143$ (red line) for $y_0 = (1, \ldots, 1)$. The abscissas are the times $t \in [0, 3]$.

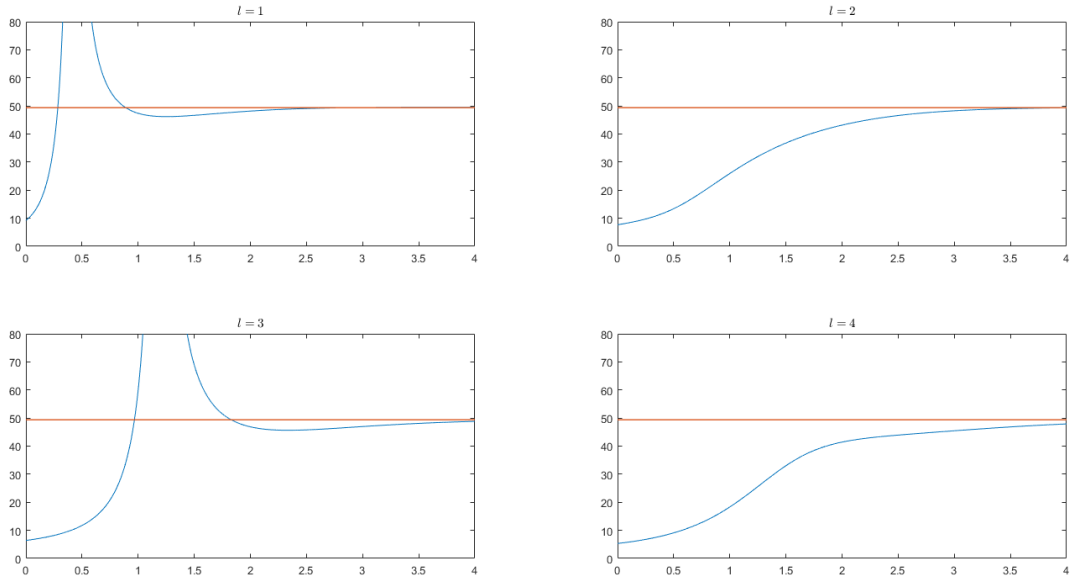

Figure 4.2: $K_l(t, A, y_0)$ (blue line), for $l = 1, 2, 3, 4$, along with the constant value $\frac{\|w\|}{|w\widehat{y}_0|} = 49.3891$ (red line) for $y_0 = ((-1.2)^l)_{l=1,\ldots,10}$. The abscissas are the times $t \in [0, 4]$.

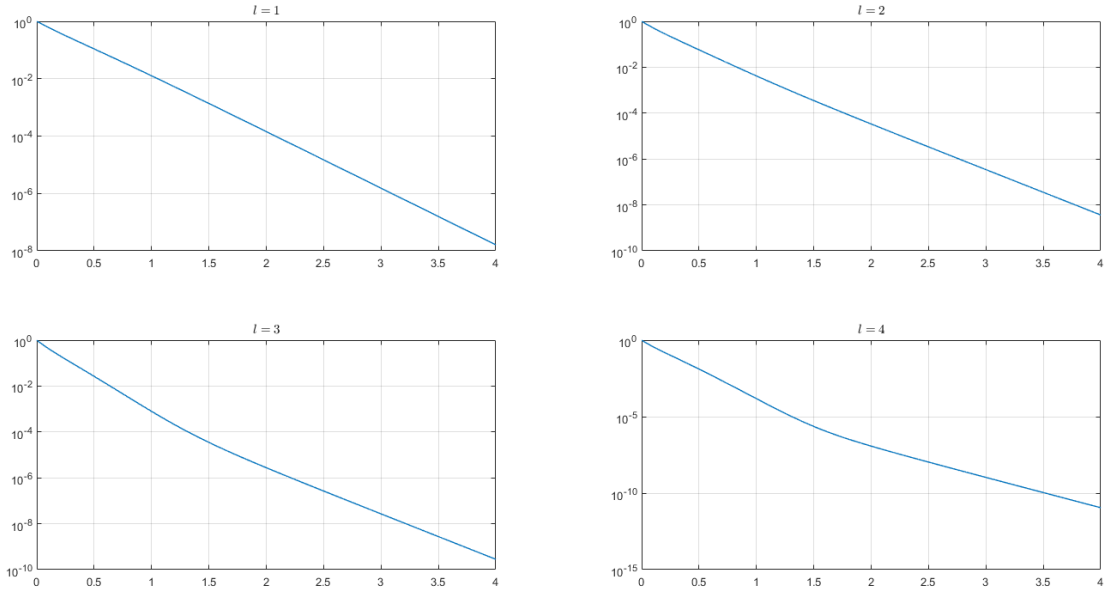

Figure 4.3: $\left\|e_l^T \mathrm{e}^{tA}\right\|$ (blue line) for $l = 1, 2, 3, 4$. The abscissas are the times $t \in [0, 4]$.

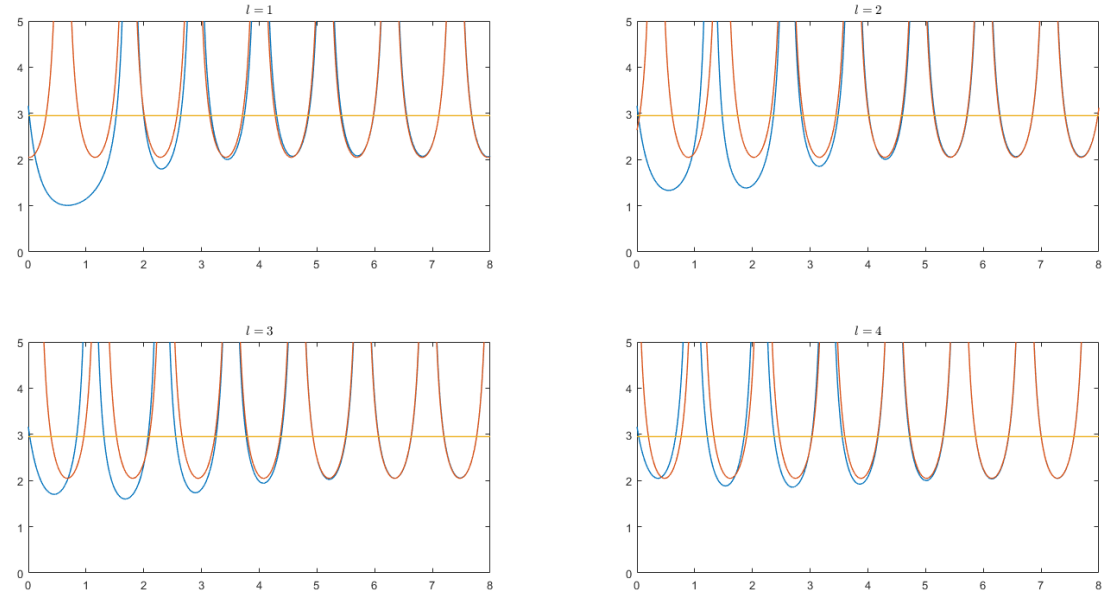

Figure 4.4: $K_l(t, A, y_0)$ (blue line) for $l = 1, 2, 3, 4$, along with $\frac{\left\|\mathrm{Re}\left(\mathrm{e}^{\sqrt{-1}\omega_1 t} v_l w\right)\right\|}{\left|\mathrm{Re}\left(\mathrm{e}^{\sqrt{-1}\omega_1 t} v_l w\right)\widehat{y_0}\right|}$ (red line) and the constant value $\frac{\|w\|}{|w\widehat{y_0}|} = 2.9509$ (yellow line) for $y_0 = (1, \ldots, 1)$. The abscissas are the times $t \in [0, 8]$.

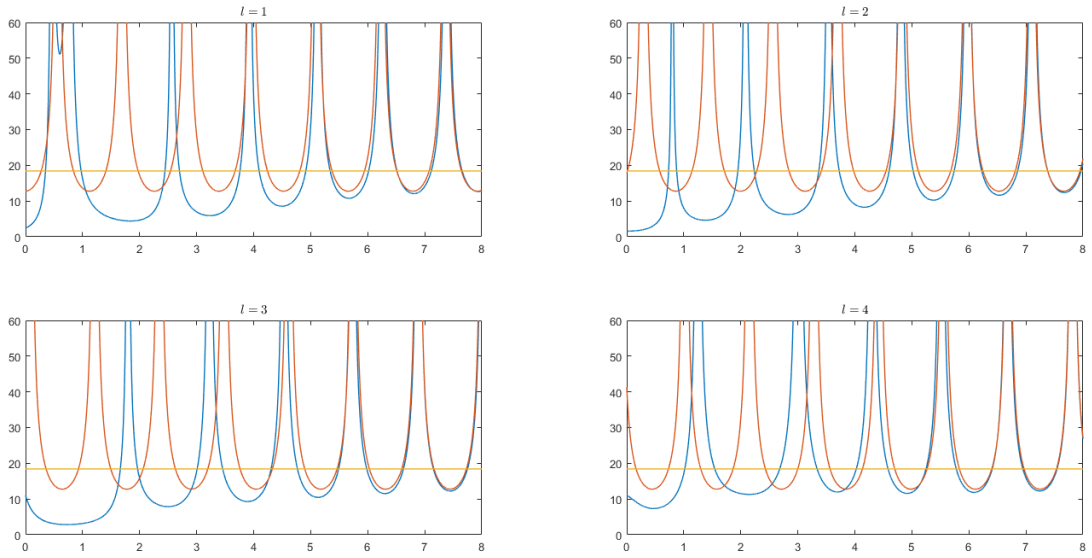

Figure 4.5: $K_l(t, A, y_0)$ (blue line), for $l = 1, 2, 3, 4$, along with $\frac{\left\|\mathrm{Re}\left(\mathrm{e}^{\sqrt{-1}\omega_1 t} v_l w\right)\right\|}{\left|\mathrm{Re}\left(\mathrm{e}^{\sqrt{-1}\omega_1 t} v_l w\right)\widehat{y}_0\right|}$ (red line) and the constant value $\frac{\|w\|}{|w\widehat{y}_0|} = 18.4079$ (yellow line) for $y_0 = (0.9, -1.4, 0.2, 0.2, -0.2, 0.9, -0.4, -0.8, 0.3, 0.5)$. The abscissas are the times $t \in [0, 8]$.

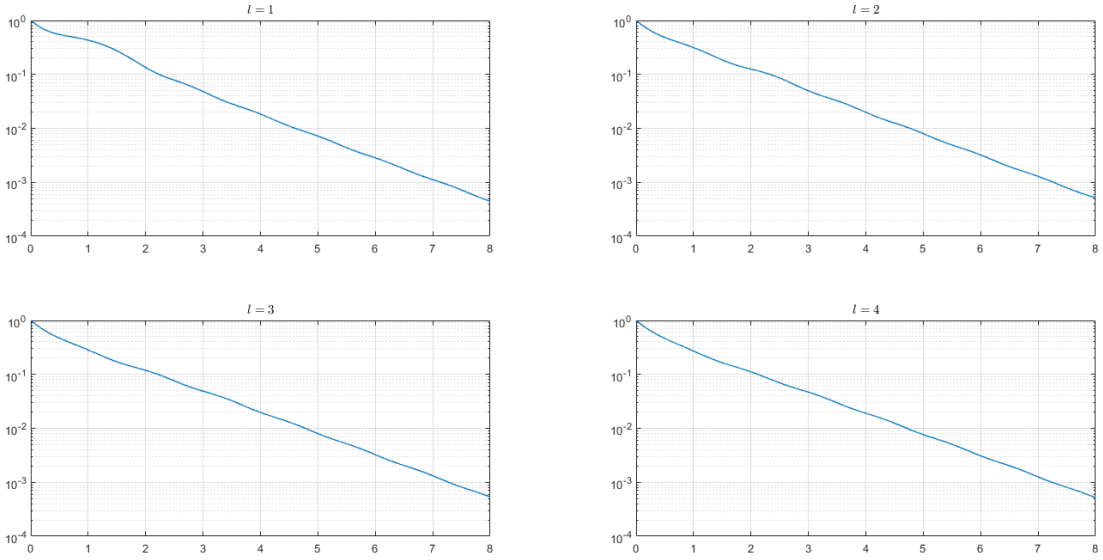

Figure 4.6: $\left\|e_l^T \mathrm{e}^{tA}\right\|$ (blue line) for $l = 1, 2, 3, 4$. The abscissas are the times $t \in [0, 8]$.

the initial value error is, in worst case, scaled by an exponential factor increasing in time.

Conclusion

This thesis studies how perturbations either in the initial value y_0 or in the matrix A propagate along the solution of the following linear ODE

$$\begin{cases} y'(t) = Ay(t),\ t \geq 0, \\ y(0) = y_0, \end{cases} \tag{4.5.1}$$

where $A \in \mathbb{R}^{n\times n}$ and $y_0 \in \mathbb{R}^n$ and $y(t) = e^{tA}y_0$ is the solution of the equation.
The error in input and output data is measured in a relative sense. To quantify errors we use the euclidean norm as a vector norm and the spectral norm as a matrix norm. We also give the componentwise relative error analysis, i.e. by considering how perturbations in the initial value propagate in each component of the solution.

We describe the relation between relative error in the input data and output data by defining three condition numbers. The three condition numbers are: a condition number with the direction of perturbation, a condition number independent of the direction of perturbation (this is the classical definition of condition number for a problem) and a condition number independent of specific data. We study these condition numbers in great depth in the case the when the matrix A is a normal matrix for normwise error analysis and in the case when A is a diagonalizable for the componentwise analysis. We give very useful upper and lower bounds on these condition numbers. A prominent aspect of the thesis is the long term behavior of these condition numbers.
Of course, beyond this thesis further work needs to be done. In particular, the general case of an arbitrary matrix, not necessarily normal, has to be developed.

We comment that the condition numbers developed in this thesis are different from the condition number one can find in many papers in literature. Perturbation analyses found in literature do not take into account the role of the initial value which is part of our study, as we are dealing with a vector quantity $e^{tA}y_0$ rather than a matrix quantity e^{tA}. Another difference is that the present study considers the analysis of dependence of condition numbers on time t which is missing in previous studies.
We address a few observations below.

- Specific structural conditions on the matrix A. For example, consider

$$A = \begin{bmatrix} a & 0 & b \\ 0 & c & 0 \\ b & 0 & d \end{bmatrix},$$

and the perturbation matrix $\tilde{A}$ is given by $\tilde{A} = A + \epsilon \|A\| \hat{B}$, where $\hat{B}$ is the direction of perturbation. It makes sense to consider only perturbations in the non-zero elements of the matrix A, and so consider $\hat{B}$ possessing the same structure as of the matrix A i.e.

$$\hat{B} = \begin{bmatrix} x & 0 & y \\ 0 & z & 0 \\ y & 0 & w \end{bmatrix}.$$

So, we have condition numbers: $K(t, A, y_0, \hat{B})$ with $\hat{B}$ with this particular structure;

$$K(t, A, y_0) = \sup_{\substack{\hat{B} \in \mathbb{R}^{n \times n} \\ |||\hat{B}||| = 1}} K(t, A, y_0, \hat{B})$$

where $\hat{B}$ is taken over all unit matrices $\hat{B}$ of this particular structure and

$$K(t, A) := \sup_{\substack{y_0 \in \mathbb{R}^n \\ y_0 \neq 0}} K(t, A, y_0)$$

- Suppose we have nonhomogeneous linear ODE i.e.

$$\begin{cases} y'(t) = Ay(t) + f(t), \ t \geq 0, \\ y(0) = y_0, \end{cases}$$

where

$$y(t) = e^{tA} y_0 + \int_0^t e^{(t-s)A} f(s) ds \tag{4.5.2}$$

is the solution of the equation. By considering perturbations in the initial value y_0, we observe that the absolute errors of the solution (4.5.2) and of the solution of the homogeneous problem remains the same. Since

$$\tilde{y}(t) = e^{tA} \tilde{y}_0 + \int_0^t e^{(t-s)A} f(s) ds,$$

we have

$$\|\tilde{y}(t) - y(t)\| = \left\| e^{tA} \tilde{y}_0 - e^{tA} y_0 \right\|.$$

On the other hand, the relative errors of homogeneous and nonhomogeneous equations are different. The relative error for the nonhomogeneous linear ODE is given by

$$\zeta(t) = K(t, A, y_0, \hat{z}_0) \cdot \frac{\left\| e^{tA} y_0 \right\|}{\|y(t)\|} \epsilon,$$

where ϵ is the relative error in the initial value and $K(t, A, y_0, \hat{z}_0)$ is the condition number with the direction of perturbation of homogeneous problem. The relation between $\zeta(t)$ and ϵ can be interesting and useful to know. We have the same observation for the case when the matrix perturbs.

- The perturbation analysis of this thesis could be used for inferring something about the expected behavior of a numerical integration of the ODE (4.5.1) by interpreting the numerical errors as perturbations in the matrix A.

- In the case of a slowly verying normal matrix $A(t)$, we can make use of results in this thesis by replacing $A(t)$ by its average over a long period of time

$$\overline{A} = \frac{L}{T} \int_0^T A(s) ds$$

which is a matrix constant in time. Now suppose $A(t)$ is perturbed to

$$\tilde{A}(t) = A(t) + \epsilon(t) \left\| A(t) \right\| \hat{B}$$

where $\hat{B}(t)$ is slowly verying. In the context where we replace $A(t)$ by its average $\overline{A}$, we can consider the time constant perturbation of $\overline{A}$.

$$\tilde{A} = \overline{A} + \bar{\epsilon} \left\| \overline{A} \right\| \overline{\hat{B}}$$

where $\bar{\epsilon}$ and $\overline{\hat{B}}$ are the averages over the time T of $\epsilon(t)$ and $\hat{B}(t)$.

Bibliography

[1] A.Al-Mohy and N. Higham. Computing the action of the matrix exponential, with an application to exponential integrators. *SIAM Journal on Scientific Computing*, 33(2):488–511, 2011.

[2] A.Al-Mohy and N.Higham. Computing the Frechét derivative of the matrix exponential, with an application to condition number estimation. *SIAM Journal on Matrix Analysis and Applications*, 30:1639–1657, 2008/2009.

[3] A. Al-Mohy and N. Higham. A New Scaling and Squaring Algorithm for the Matrix Exponential. *SIAM Journal on Matrix Analysis and Applications*, 31, 01 2009.

[4] A. Al-Mohy, N. Higham, and S. Relton. Computing the Fréchet Derivative of the Matrix Logarithm and Estimating the Condition Number. *SIAM Journal on Scientific Computing*, 35:C394–C410, 07 2013.

[5] F. S. Alfio Quarteroni, Riccardo Sacco. *Numerical mathematics*. Springer Science Business Media, 2010.

[6] J. Awrejcewicz. *Ordinary differential equations and mechanical systems*. Springer, 2014.

[7] D. Behmardi and E. D. Nayeri. Introduction of Fréchet and Gâteaux Derivative. *Applied Mathematical Sciences.*, 2(20):975 – 980, 2008.

[8] R. Bellman. *Introduction to Matrix Analysis*. McGraw-Hill, New York, 1960.

[9] Å. Björck. Component-wise perturbation analysis and error bounds for linear least squares solutions. *BIT Numerical Mathematics*, 31(2):237–244, 1991.

[10] H. Broer, F. Takens, and B. Hasselblatt. *Handbook of dynamical systems*. Elsevier, 2010.

[11] P. Bürgisser and F. Cucker. *Condition: The Geometry of Numerical Algorithms*, volume 349. 01 2013.

[12] H.-W. Cheng and S. S.-T. Yau. More explicit formulas for the matrix exponential. *Linear algebra and its applications*, 262:131–163, 1997.

[13] C. Chicone. *Ordinary differential equations with applications*, volume 34. Springer Science & Business Media, 2006.

[14] L. Cobb. Stochastic differential equations for the social sciences. *Mathematical frontiers of the social and policy sciences*, pages 37–68, 1981.

[15] F. Cucker. Probabilistic analyses of condition numbers. *Acta Numerica*, 25, 2016.

[16] F. Cucker and H. Diao. Mixed and componentwise condition numbers for rectangular structured matrices. *Calcolo*, 44:89–115, 2007.

[17] F. Cucker, H. Diao, and Y. Wei. On mixed and componentwise condition numbers for Moore–Penrose inverse and linear least squares problems. *Mathematics of Computation*, 76(258):947–963, 2007.

[18] C. Davis. Explicit functional calculus. *Linear Algebra and its Applications*, 6:193–199, 1973.

[19] E. Deadman. Estimating the condition number of $f(a)b$. *Numerical Algorithms*, 70, 2015.

[20] J. W. Demmel. On condition numbers and the distance to the nearest ill-posed problem. *Numerische Mathematik*, 10, 1987.

[21] H. Diao, H. Xiang, and Y. Wei. Mixed, componentwise condition numbers and small sample statistical condition estimation of sylvester equations. *Numerical Linear Algebra with Applications*, 19, 08 2012.

[22] M. DRAZIN. ON DIAGONABLE AND NORMAL MATRICES. *The Quarterly Journal of Mathematics*, 2:189–198, 1951.

[23] F.-C. S. M. Ellahi, Rahmat. Recent Advances in the Application of Differential Equations in Mechanical Engineering Problems. *Mathematical Problems in Engineering*, 2018.

[24] L. Elsner and K. Ikramov. Normal matrices: an update. *Linear Algebra and its Applications*, 285(1):291–303, 1998.

[25] R. . G. F. *The Theory of Matrices,*. Interscience Publishers,, New York, 1959.

[26] A. Farooq and S. Maset. Propagation of perturbations in the initial value along solutions of linear odes: a componentwise relative error analysis. *Rendiconti dell'Istituto di Matematica dell'Università di Trieste: an International Journal of Mathematics*, 53:1–14, 2021.

[27] A. Farooq and S. Maset. How perturbations in the matrix of linear systems of ordinary differential equations propagate along solutions. *Journal of Computational and Applied Mathematics*, 407:114046, 2022.

[28] H. Fassbender and K. Ikramov. Conjugate-normal matrices: A survey. *Linear Algebra and its Applications*, 429:1425–1441, 10 2008.

[29] A. Frommer and V. Simoncini. *Matrix Functions*, volume 13, pages 275–303. Model Order Reduction: Theory, Research Aspects and Applications, 01 2008.

[30] I. Gohberg and I. Koltracht. Mixed Componentwise and Structured Condition Numbers. *SIAM J. Matrix Anal. Appl.*, 14(3), 1993.

[31] G. H. Golub and C. F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, USA, 1996.

[32] G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3. JHU press, 2013.

[33] O. Gonzalez-Gaxiola. A note on the derivation of Fréchet and Gâteaux. *Applied Mathematical Sciences (Ruse)*, 3, 01 2009.

[34] R. Grone, C. R. Johnson, E. M. Sa, and H. Wolkowicz. Normal matrices. *Linear Algebra and its Applications*, 87:213–225, 1987.

[35] B. Hall. *Lie groups, Lie algebras, and representations: an elementary introduction*, volume 222. Springer, 2015.

[36] D. J. Higham. Condition numbers and their condition numbers. *Linear Algebra and its Applications*, 214, 1995.

[37] D. J. Higham and N. J. Highamt. Componentwise perturbation theory for linear systems with multiple right-hand sides. *Linear Algebra and its Applications*, 174:111–129, 1992.

[38] N. Higham and L. Lin. An Improved Schur-Padé Algorithm for Fractional Powers of a Matrix and Their Fréchet Derivatives. *SIAM J. Matrix Anal. Appl.*, 34:1341–1360, 2013.

[39] N. Higham and S. Relton. Higher Order Fréchet Derivatives of Matrix Functions and the Level-2 Condition Number. *SIAM Journal on Matrix Analysis and Applications*, 35:1019–1037, 07 2014.

[40] N. J. Higham. *A survey of componentwise perturbation theory*, volume 48. American Mathematical Society, 1994.

[41] N. J. Higham. *Functions of Matrices*. Society for Industrial and Applied Mathematics, 2008.

[42] N. J. Higham and A. H. Al-Mohy. Computing matrix functions. *Acta Numerica*, 19:159–208, 2010.

[43] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012.

[44] R. M. Ilea M, Turnea M. Ordinary differential equations with applications in molecular biology. *Revista medico-chirurgicală̆a Societăţii de Medici şi Naturalişti din Iaşi*, 116:347–52, 10 2012.

[45] B. Kågström. Bounds and perturbation bounds for the matrix exponential. *BIT Numerical Mathematics*, 17, 1977.

[46] C. Kenney and A. J. Laub. Condition Estimates for Matrix Functions. *SIAM Journal on Matrix Analysis and Applications*, 10(2):191–209, 1989.

[47] R. Kirchner. An explicit formula for e^{At}. *The American Mathematical Monthly*, 74(10):1200–1204, 1967.

[48] J. J. Koliha. BLOCK DIAGONALIZATION. *MATHEMATICA BOHEMICA*, 126(1):237–246, 2001.

[49] N. Kushida. Condition Number Estimation of Preconditioned Matrices. *PloS one*, 10, 2015.

[50] D. Laksov. Diagonalization of matrices over rings. *Journal of Algebra*, 376:123–138, 2013.

[51] A. Levis. Some computational aspects of the matrix exponential. *Automatic Control, IEEE Transactions on*, AC14:410 – 411, 09 1969.

[52] Y. Lin and Y. Wei. Normwise, mixed and componentwise condition numbers of nonsymmetric algebraic riccati equations. *Journal of Applied Mathematics and Computing*, 27:137–147, 2008.

[53] Q. Liu and Z. Jia. On condition numbers of the total least squares problem with linear equality constraint, 2020.

[54] C. V. Loan. The sensitivity of the matrix exponential. *SIAM Journal on Numerical Analysis*, 14:971–981, 1977.

[55] K. Long. Gateaux Differentials and Frechet Derivativs. *AMS*, pages 12–26, 2009.

[56] C. C. MacDuffee. *The theory of matrices.* Chelsea, 1956.

[57] W. Magnus. On the exponential solution of differential equations for a linear operator. *Communications on Pure and Applied Mathematics*, 7(4), 1954.

[58] J. Marques. An application of ordinary differential equations in Economics: modelling consumer's preferences using marginal rates of substitution. *Mathematical Methods in Science and Mechanics: Mathematics and Computers in Science and Engineering Series*, 33, 2014.

[59] S. Maset. Conditioning and relative error propagation in linear autonomous ordinary differential equations. *Discrete & Continuous Dynamical Systems - B*, 23, 2018.

[60] I. Mihai, M. Turnea, and M. Rotariu. Differential equations with applications in cancer diseases. *Revista medico-chirurgicală a Societăţii de Medici şi Naturalişti din Iaşi*, 117:572–7, 12 2013.

[61] C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM review*, 45(1):3–49, 2003.

[62] I. Najfeld and T. F. Havel. Derivatives of the matrix exponential and their computation. *Advances in applied mathematics*, 16(3):321–375, 1995.

[63] M. Z. Nashed. Some Remarks on Variations and Differentials. *The American Mathematical Monthly*, 73(4P2):63–76, 1966.

[64] A. Pyzara, B. Bylina, and J. Bylina. The influence of a matrix condition number on iterative methods' convergence. 2011 Federated Conference on Computer Science and Information Systems, FedCSIS 2011, 2011.

[65] J. R. Rice. A Theory of Condition. *SIAM Journal on Numerical Analysis*, 3(2):287–310, 1966.

[66] R. F. Rinehart. The Equivalence of Definitions of a Matric Function. *The American Mathematical Monthly*, 62(6):395–414, 1955.

[67] J. Rohn. New condition numbers for matrices and linear systems. *Computing*, 41(1-2):167–169, 1989.

[68] R. D. Skeel. Scaling for Numerical Stability in Gaussian Elimination. *J. ACM*, 26(3):494–526, July 1979.

[69] S. J. Szarek. Condition numbers of random matrices. *Journal of Complexity*, 7(2):131–149, 1991.

[70] R. B. Taher and M. Rachidi. Some explicit formulas for the polynomial decomposition of the matrix exponential and applications. *Linear Algebra and its Applications*, 350(1-3):171–184, 2002.

[71] Y. Tang, L. Bao, and Y. Lin. Perturbation analysis of the generalized Sylvester equation and the generalized Lyapunov equation. *International Journal of Computer Mathematics*, 88(2):408–420, 2011.

[72] R. C. Thompson. A Note on Normal Matrices. *Canadian Journal of Mathematics*, 15:220–225, 1963.

[73] R. C. Thompson. Special cases of a matrix exponential formula. *Linear Algebra and its Applications*, 107:283–292, 1988.

[74] T.Kato. *Perturbation Theory for Linear Operators*. Spinger-Verlay New York, 1966.

[75] H. F. Trotter. On the product of semi-groups of operators. *Proceedings of the American Mathematical Society*, 10(4):545–551, 1959.

[76] A. Turing. Rounding-off Errors in Matrix Processes. *The Quarterly Journal of Mechanics and Applied Mathematics, 1, 287-308.*, 1948.

[77] C. F. Van Loan. A study of the matrix exponential. *Manchester Institute for Mathematical Sciences, University of Manchester*, 2006.

[78] V. S. Varadarajan. *Lie Groups, Lie Algebras, and Their Representations.* Prentice-Hall, Englewood Cliffs, NJ, USA, 1974.

[79] J. von Neumann and H. H. Goldstine. Numerical inverting of matrices of high order. *Bull. Amer. Math. Soc.*, 53, 1947.

[80] W.-g. Wang and N. Hao. On mixed and component wise condition numbers for hyperbolic QR factorization. *Filomat*, 1, 01 2008.

[81] J. X. . W. G. Weifang Zhu. The Sensitivity of the Exponential of an Essentially Nonnegative Matrix. *Journal of Computational Mathematics*, 26(2):250–258, 2008.

[82] S. H. Weintraub. *Jordan Canonical Form: Theory and Practice.* 2009.

[83] R. M. Wilcox. Exponential Operators and Parameter Differentiation in Quantum Physics. *J. Math. Phys.*, 8, 1967.

[84] H. Xiang and Y. Wei. Structured mixed and componentwise condition numbers of some structured matrices. *Journal of Computational and Applied Mathematics*, 202(2):217–229, 2007.

[85] H. Zhang, H. Xiang, and Y. Wei. Condition numbers for linear systems and Kronecker product linear systems with multiple right-hand sides. *International Journal of Computer Mathematics*, 84(12):1805–1817, 2007.

[86] L. Zhou, Y. Lin, Y. Wei, and S. Qiao. Perturbation analysis and condition numbers of symmetric algebraic Riccati equations. *Automatica*, 45:1005–1011, 04 2009.

[87] A. D. Ziebur. On Determining the Structure of A by Analyzing e^{At}. *SIAM Review*, 12:98–102, 1970.