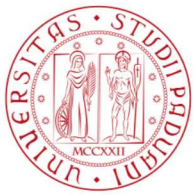


Università degli Studi di Padova

Department of Information Engineering

Master Thesis in Control Systems Engineering

A linear system approach to explain neuronal network excitability

Supervisor
Sandro Zampieri
University of Padova

Co-supervisor
Giacomo Baggio
University of Padova

Master Candidate
Umberto Casti

September 5, 2022

Academic Year 2021-2022

Abstract

This thesis has the scope to investigate neuron behaviour from a dynamic system point of view. Starting from the neuron models existing in the literature, that are typically complex and *non-linear*, we present different techniques for their analysis. We present various methodologies that permit the study of the transient behaviour of these systems beyond a pure stability analysis, in such a way to understand why excitable behaviours and oscillator phenomena can occur. Finally, we also present a global linearisation method based on the *Koopman operator*. This technique permits the analysis of complex *non-linear* neuron models from a *linear* perspective. In this domain it seems to exist a relation between the *non-normality*, a *mathematical* property of the *linear* system matrix, and the characterization of the excitability, a biological *qualitative* property, that is fundamental in the definition of the neuron from a functional point of view. This result helps in giving a more precise characterization of this fleeing concept.

Sommario

Questa tesi ha come obbiettivo quello di studiare il comportamento del neurone da un punto di vista della teoria dei sistemi dinamici. In letteratura è possibile trovare numerosi modelli che descrivono il funzionamento del neurone, però sono complessi e *non-lineari*. In questo elaborato presenteremo diversi metodi di analisi applicabili a questa tipologia di modelli. Questi metodi sono strumenti utili per lo studio, non solo della stabilità, ma anche di fenomeni più complessi come evoluzioni transitorie e fenomeni oscillatori. Successivamente presenteremo una tecnica di linearizzazione globale basata sul *Koopman operator*, tramite questo strumento, un complesso modello *non-lineare* può essere analizzato con le tecniche della teoria dei sistemi *lineari*. Da questo studio si vedrà come sia naturale pensare che esista una relazione tra la *non-normalità*, una proprietà matematica di una matrice associata ad un sistema lineare, e l'eccitabilità, una proprietà biologica qualitativa, che è fondamentale per la definizione del neurone da un punto di vista del funzionamento. In questo modo cercheremo di dare una caratterizzazione più precisa a questo sfuggente concetto.

Contents

Listing of figures

1 Introduction

The present thesis has the purpose to examine different techniques for the analysis of the neuronal and brain behaviour. This type of study is fundamental in neuroscience but, at the same time, it is surprising how many common elements exist between engineering and neuroscience. For example, engineering-related subjects such as artificial intelligence and machine learning have been developed under the inspiration of computational neuroscience, similarly tools from dynamical systems and control theory have been used for decades to shed light on the dynamic phenomena of the brain. So, the study of the brain and of the neurons from a dynamical systems and control perspective seems to be natural [1].

In neurophysiology the neuron is typically modelled as a *non-linear* electronic circuit in which an external stimulus produces a signal represented by the variation of an electric quantity that is the membrane potential. Indeed, the neurons have a membrane that can maintain a voltage difference between the extracellular and intracellular medium, but the voltage across the membrane can vary because ions can flow through the membrane via specific transmembrane proteins called ion channels [2]. The key idea of this type of analysis is modelling the ions pumps and the membrane potential as circuits composed by resistors, voltage sources and electrical conductances, and the final result is typically a complex set of *non-linear* differential equations. These types of models are called conductance-based models and seem qualitatively and quantitatively adequate to represent the nature of the neural behaviours. One of the most studied and representative classes of neuronal models are those based on the Hodgkin-Huxley model [3]. But these models, that in many cases seem to work well, are

typically difficult to analyse; indeed, it is often needed to resort to simplified versions such as the FitzHugh-Nagumo model [4]. This simplification makes the problem more tractable with some advanced techniques, such as *Singular Perturbation Theory* [5], but anyway makes the model less accurate.

These models are proposed to try to reproduce the properties of the neurons. In particular one of the most important characteristic that a neuron must have is excitability. Excitability is the capability of a neuron to respond to an external stimulus with a fast and large variation of its membrane potential, which after few instants finishes and brings again the neuron to a rest situation, see figure 1.1. The focus of this thesis is the study of the single neuron as an abstract

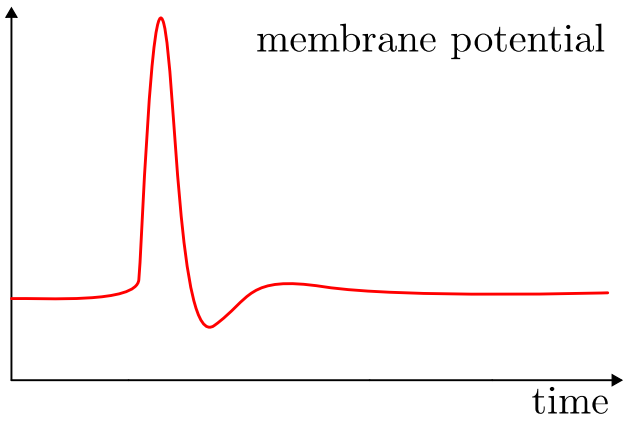


Figure 1.1: In this plot we illustrate the typical shape of the action potential passing through the neuron.

element which is able to encode excitability. In particular in this thesis we review some of the *non-linear* models proposed in the literature and we present *non-linear* techniques for their analysis. After that, we present a global linearisation method that permits to study the neuron dynamics from a linear point of view.

By doing so, it is possible to simplify the analysis and try to answer to more fundamental questions. The various *non-linear* models already proposed depend strongly on our biological knowledge of the neuron, which is something that changes drastically and rapidly with the experimental evidence. Typically, the way to update these models is to add new terms and complexity, but in this way it is possible that the models become more and more specific and less general.

So the aim of this thesis is to focus only on the excitability as a general concept that could be captured by some kind of linear systems property that is a high *Non-Normality* degree. *Non-Normality* is a property of a dynamical system that is strongly linked to excitability. Someone could think that linear systems are too simple to describe so complex phenomena, as neural

activity, but this is not necessarily true. Indeed, *large-scale linear network systems* can exhibit very complex behaviours. In this case, the abundance of properties does not depend on the complexity of the singles elements of the system but by the complex interaction between the state variables. In addition, the neuron can exhibit more complex behaviour such the generation of *limit cycles*. It is well known that these behaviours appear only in *non-linear* systems and can not be reproduced by linear systems. So the translation *non-linear/linear* system can fail to reproduce these behaviours. For completeness in this thesis we present also more recent techniques, based on the *dissipativity theory* [6] and the *p*-dominance [7]. These instruments seem to be powerful tools to study these behaviours and also their generalization to interconnected systems. The thesis is structured in the following way. In the first part we present a section of math preliminaries 2 where the context of study is present and the fundamental instruments necessary to this analysis are introduced. After that, we present a linear model for the neuron behaviour. To do that it is possible to start from a *non-linear* model and to use some advanced global linearisation techniques to derive a linear model with particular structures that encode excitability. In this thesis it is used the *Koopman Operator* [8], a very powerful tool that can give a global linearisation of a *non-linear* system translating a *non-linear*, finite dimension problem into a linear but infinite dimensional system (see section 2.4). To apply this instrument in that context it is necessary to find a first *non-linear* model that can be a starting point for that analysis. This model is the one presented in section 3.2. In the end of this thesis, in chapter 4, we will present some simulations results and we try to study and evaluate the *Non-normality* degree of the *linear* neuron system so derived.

2

Maths Preliminaries

Some neural models proposed in the literature are presented in chapter 3. These biological models are complex and difficult to analyse. To deal with this problem, a lot of mathematical tools have been derived and applied to this context of study. These tools are very diverse and touch elements of *non-linear* and *linear dynamical system analysis*.

We mentioned that the neuronal behaviour is typically an excitatory behaviour, characterized by rapidly changing phenomena and *limit-cycles*. So stability analysis is not sufficient and it is necessary to use more accurate instruments to study the trajectories of the system. In this section, some of these tools are presented.

Most of the comprehension of the behaviour of the existing neural models is based on the use of the *Singular Perturbation Theory* (in this thesis presented from a geometrical point of view in 2.1.2). This theory can be very useful, but at the same time has some weaknesses:

- the analysis is typically *qualitative*, and often it is not possible to derive *quantitative-information* about the system. For example, for systems with *limit-cycles* it is easy to understand the support of the orbits, but it is more problematic to try to estimate the *period*.
- The analysis is typically tractable for low dimensional systems, but can become very problematic when increasing the dimension of the system.
- The study is *case-by-case*: there does not exist simple and general ways to attack the problem.

Another interesting approach is to resort to the *dissipativity theory* (section 2.2). This is a very elegant theory that is founded on the idea of studying a *dynamical system* by observing its energetic exchange with the environment. One of the advantages of this approach consists in its capability to easily understand the behaviour of interconnected *systems*. Moreover, the analysis is typically systemic and able to deal with high dimensional systems. Although these are classical and well-known approaches, in section 2.3 we present an extension of the *dissipativity* analysis that can be used to derive a theory for *multistability* and *limit-cycles*. In contrast to these, in section 2.4, we present the *Koopman operator*, a new and interesting instrument that allows to characterize the *non-linear* behaviour of a system from a *linear* point of view. We do this by resorting to a translation of the problem in an *infinite-dimensional* framework. The advantage of this approach is in the fact that this linearisation is *global* and *exact*. Unfortunately, an *attracting* or *repelling limit-cycle* can not be reproduced by a linear system and so by the *Koopman operator* is not possible to study this behaviour. But it is possible to analyse other simpler excitatory dynamics. The *non-normal* property (introduced in section 2.5) of the linearised system can be used as a quantitative characterization of excitability.

2.1 Singular Perturbation Theory

Exact closed-form analytic solutions of *non-linear* differential equations can be obtained only for limited special classes of differential equations. In general it is necessary resort to approximate solutions. There are two distinct categories of approximation methods used to analyse *non-linear* systems:

1. numerical solution methods;
2. asymptotic methods.

In this section we introduce an asymptotic method for the analysis of the *non-linear* differential equation

$$\dot{x}(t) = f(x(t), \epsilon)$$

where ϵ is a "small" scalar parameter. Under certain conditions, the equation has an exact solution $x_\epsilon(t)$. The goal of an asymptotic method is to obtain an approximate solution $\tilde{x}_\epsilon(t)$ such that the approximation error, in some norm, $x_\epsilon(t) - \tilde{x}_\epsilon(t)$ is small for small $|\epsilon|$. The approximate solution $\tilde{x}_\epsilon(t)$ is expressed in terms of equations that are simpler than the original ones. This method is used for revealing underlying structural properties possessed by the original state equation for small $|\epsilon|$. In the following we derived the *perturbation* method, where we exploit the small size of the perturbation parameter ϵ to construct approximate solutions.

2.1.1 Regularly and singularly perturbed systems

Typically the perturbation problems are identified by the presence of a small parameter $\epsilon > 0$. These problems are divided in two types: *regular* and *singular*. The scope of this section is to introduce some fundamental concepts and properties of the regularly and singularly perturbed problems. Consider a system of the form

$$\dot{x}(t) = f(x(t), \epsilon), \tag{2.1}$$

where $x(t) \in \mathbb{R}^n$, $\epsilon > 0$ and f sufficiently regular. By setting $\epsilon = 0$ in (2.1), we define the so-called unperturbed problem

$$\dot{x}(t) = f(x(t), 0). \tag{2.2}$$

Let $x_\epsilon(t)$ and $x_0(t)$ the solutions of (2.1) and (2.2) on some finite interval $0 \leq t \leq T$, and with same initial condition x_0. If f is continuously differentiable with respect to x and continuous in ϵ, then for a sufficiently small ϵ, we can represent the solution $x_\epsilon(t)$ through $x_0(t)$ as

$$x_\epsilon(t) = x_0(t) + R_0(t, \epsilon), \tag{2.3}$$

where $R_0(t, \epsilon) \to 0$ as $\epsilon \to 0$. If now we suppose that the right hand side of (2.1) is continuously differentiable $m \geq 1$ times with respect to x and ϵ, then, if ϵ is small, the solution $x_\epsilon(t)$ can be expanded as

$$x_\epsilon(t) = x_0(t) + \epsilon x_1(t) + \cdots + \epsilon^{m-1} x_{m-1}(t) + R_m(t, \epsilon), \tag{2.4}$$

where, if $\epsilon \to 0$, $R_m(t, \epsilon)$ goes to zero as a quantity of order ϵ^m for all t on the interval $0 \leq t \leq T$. If the right hand side of (2.1) is an analytic function of x and ϵ, then, again for ϵ small, $x_\epsilon(t)$ can be rewritten as a converging series

$$x_\epsilon(t) = x_0(t) + \sum_{m=1}^{\infty} \epsilon^m x_m(t) \tag{2.5}$$

uniformly on the interval $0 \leq t \leq T$. In the cases when this holds, we call system (2.1) a *regular perturbation* of (2.2). Instead, when the asymptotic expansions is much more complex and unpredictable, and in the series appear terms such as $\epsilon^{\frac{m}{k}} \log^l \epsilon$ or $\epsilon^{\frac{m}{k}}$ with $m,\ k,\ l \in \mathbb{N}$ we call system (2.1) a *singular perturbation problem*. Informally *singular perturbation prob-*

lems are problems in which there is a breakdown of the limit $x_\epsilon(t) \to x_0(t)$. In this thesis we consider *singular perturbation problems* applied to the FitzHugh-Nagumo model (3.5), which describes processes evolving on time scales with different orders of magnitude.

2.1.2 Geometric Singular Perturbation Theory

In this section we present a brief introduction to the *Geometric Singular Perturbation Theory* [5], a classic instrument used to study *slow-fast systems*.

Consider a basic system of the form:

$$\begin{cases} \dfrac{d}{dt}x(t) = f(x(t), y(t), \epsilon) & \text{(2.6a)} \\ \dfrac{d}{dt}y(t) = \epsilon g\left(x(t), y(t), \epsilon\right) & \text{(2.6b)} \end{cases}$$

with $x\left(t\right), y\left(t\right), \epsilon \in \mathbb{R}$. We suppose that the functions f and g are sufficiently regular. The system (2.6) is characterized by the presence of the parameter ϵ, which is typically small, that introduces a time evolution separation between the two dynamics. One is described by system (2.6) the other by system (2.7).

$$\begin{cases} \epsilon\dfrac{d}{d\tau}x(\tau) = f\left(x(\tau), y(\tau), \epsilon\right) & \text{(2.7a)} \\ \dfrac{d}{d\tau}y(\tau) = g\left(x(\tau), y(\tau), \epsilon\right) & \text{(2.7b)} \end{cases}$$

To derive the equivalence of the two systems, we make a change of time-scale in the equations (2.6). This change of time, if ϵ is different from zero, consists in the reformulation of the same problem (2.6). The orbits remain the same, what we change is the time scale. Indeed consider two sets of all solutions:

$$S = \left\{x\left(t\right), y\left(t\right)\right\} \qquad \text{solutions to (2.6)} \qquad \text{(2.8)}$$

$$\bar{S} = \left\{\bar{x}\left(\tau\right), \bar{y}\left(\tau\right)\right\} \qquad \text{solutions to (2.7).} \qquad \text{(2.9)}$$

So

$$\left(x\left(t\right), y\left(t\right)\right) \in S \Rightarrow \left(\bar{x}\left(\tau\right), \bar{y}\left(\tau\right)\right) \in \bar{S} \text{ where } \bar{x}\left(\tau\right) \coloneqq x\left(\frac{\tau}{\epsilon}\right), \ \bar{y}\left(\tau\right) \coloneqq y\left(\frac{\tau}{\epsilon}\right). \tag{2.10}$$

Similarly

$$\left(\bar{x}\left(\tau\right), \bar{y}\left(\tau\right)\right) \in \bar{S} \Rightarrow \left(x\left(t\right), y\left(t\right)\right) \in S \text{ where } x\left(t\right) \coloneqq \bar{x}\left(\epsilon t\right), \ y\left(t\right) \coloneqq \bar{y}\left(\epsilon t\right). \tag{2.11}$$

Indeed consider problem (2.10)

$$\frac{d}{d\tau}\bar{x}\left(\tau\right)=\frac{d}{d\tau}\bar{x}\left(\frac{\tau}{\epsilon}\right)=\frac{d}{dt}x\left(t\right)\Big|_{t=\frac{\tau}{\epsilon}}\frac{d}{d\tau}\frac{\tau}{\epsilon}=\frac{1}{\epsilon}f\left(x\left(\frac{\tau}{\epsilon}\right),y\left(\frac{\tau}{\epsilon}\right),\epsilon\right)=$$
$$=\frac{1}{\epsilon}f\left(\bar{x}\left(\tau\right),\bar{y}\left(\tau\right),\epsilon\right),$$

applying the same reasoning to (2.11)

$$\frac{d}{dt}x\left(t\right)=\frac{d}{dt}x\left(\epsilon t\right)=\frac{d}{d\tau}\bar{x}\left(\tau\right)\Big|_{\tau=\epsilon t}\frac{d}{dt}\epsilon t=\frac{1}{\not{\epsilon}}f\left(\bar{x}\left(\epsilon t\right),\bar{y}\left(\epsilon t\right),\epsilon\right)\not{\epsilon}=$$
$$=f\left(x\left(t\right),y\left(t\right),\epsilon\right),$$

and applying the same reasoning to coordinates $y\left(t\right)$ and $\bar{y}\left(\tau\right)$ we prove (2.10) and (2.11). Usually, system (2.6) is called the *fast system* while system (2.7) is called the *slow system*. The intuition behind the *geometric singular perturbation theory* is that, in the case $\epsilon=0$, the *fast system* and the *slow system* can be studied separately. The two different problems for $\epsilon=0$ are

$$\begin{cases}\dfrac{d}{dt}x(t)=f(x(t),y(t),0) & \text{(2.12a)}\\ \dfrac{d}{dt}y(t)=0 & \text{(2.12b)}\end{cases}$$

and

$$\begin{cases}0=f(x(\tau),y(\tau),0) & \text{(2.13a)}\\ \dfrac{d}{d\tau}y(\tau)=g(x(\tau),y(\tau),0) & \text{(2.13b)}\end{cases}$$

that are simpler to study then the original problem. But, under certain hypotheses, after this separate analysis, it is possible to combine the results to have a description of the whole original system (2.6). More in detail (2.12) defines the so-called *layer problem* instead (2.13) the so-called *reduced problem*. The *layer problem* can be re-conducted to a one-dimensional system in x where y represents a constant in time. In this case, the problem is simpler than the original because we have only a one-dimensional dynamics. Instead, in the *reduced problem*, the dynamics obtained by setting $\epsilon=0$ induces a constraint on the system. Also in this case we can think that the dynamics of the system is one dimensional but now it is constrained to move in the set $M_0=\{(x,y)\in\mathbb{R}^2:f(x,y,0)=0\}$ called the critical manifold . Observe that, the equilibrium points of the dynamic of the *layer problem* are exactly the points of the

critical manifold M_0.

So, in a certain sense, if we combine the *layer* and the *reduced problem*, we can think that there is a *fast system* that tends to bring the trajectories to their equilibrium points, identified by M_0 and after reaching that, the trajectories proceed inside M_0 driven by the *slow dynamic*. The fundamental result of the theory, due to Fenichel [5], consists in proving that this qualitative dynamic is present also in case if ϵ is small. Now it is not possible to drastically distinguish the *reduced* and the *layer problem*, but under appropriate conditions, the trajectories of the original system are attracted or repelled by an invariant manifold M_ϵ that lives in a neighbourhood of the critical manifold M_0. As in the case $\epsilon = 0$, inside the manifold M_ϵ the dynamic is almost completely governed by the *slow system* instead outside by the *fast system*. The fundamental hypothesis that must hold is that the *fast system* acts before the *slow system*. In other words, we require that there exist a well-defined hierarchy between the action of the *fast* and the *slow dynamics*. From a mathematical point of view, the condition is translated into requiring that the critical manifold is *normally hyperbolic* so that the attraction and/or the repulsion from the manifold is stronger than the dynamics inside the manifold. For a rigorous definition of *normally hyperbolicity*, we refer to [9].

However, points where the *normal hyperbolicity* hypothesis breaks are common in applications and Fenichel's result is not more applicable in these cases. For us a *hyperbolic* point is a equilibrium of the *layer dynamic* where the partial derivative with respect to x is zero.

$$\frac{\partial f}{\partial x}(x, y, 0) = 0 \quad \text{with } (x, y) \in M_0 \tag{2.14}$$

In this situation, thinking of the *layer problem* as an approximation of the *fast system*, the equilibrium with respect to the *fast dynamic* is no more *hyperbolic* and so the trajectories diverge or converge slower than in the *hyperbolic* case where the convergence or divergence has an exponential rate. In other words, the *fast system* is not so fast and now its time evolution is comparable with respect to the *slow dynamic*. Hence the problem can not be separated into two different dynamics, but we must study the whole original problem where the one dynamic influences the other. So, the analysis of these points must be done case-by-case.

In the models considered in this thesis, we will encounter a type of *non-hyperbolic* point called a *fold* point. In the following, we give a heuristic derivation of the behaviour of the trajectories near this point. A *fold* point $(x_0, y_0) \in \mathbb{R}^2$ is determined by the following con-

ditions on the structure of the dynamical system.

$$f(x_0, y_0, 0) = 0, \quad \frac{\partial f}{\partial x}(x_0, y_0, 0) = 0 \tag{2.15}$$

$$\frac{\partial^2 f}{\partial x^2}(x_0, y_0, 0) \neq 0 \quad \frac{\partial f}{\partial y}(x_0, y_0, 0) \neq 0 \quad g(x_0, y_0, 0) \neq 0. \tag{2.16}$$

Without loss of generality, we can assume:

$$(x_0, y_0) = (0, 0) \quad \frac{\partial^2 f}{\partial x^2}(0, 0, 0) > 0 \quad \frac{\partial f}{\partial y}(0, 0, 0) < 0$$

Knowing that outside the fold point *Fenichel's theorem* is valid our goal is to understand the behaviour of the trajectories in a neighbourhood of this point, a *non-hyperbolic* point for the *fast dynamic*. By the previous assumptions, we are sure that there exists a neighbourhood U where the origin is the only point in the critical manifold for which $\frac{\partial f}{\partial x}$ vanishes. Indeed, if we choose U such that for all $x, y \in U$ we have $\frac{\partial f}{\partial y} < 0$ and $\frac{\partial^2 f}{\partial x^2} > 0$, then $\frac{\partial f}{\partial x}$ cannot vanish along x thanks to the strictly positivity of the second partial derivative and if we fix $x = 0$ and we move along y thanks to $\frac{\partial f}{\partial y} < 0$ we exit from the critical manifold. Consider the following expansion around $(0, 0)$

$$f(x, y, 0) = \cancel{f(0,0,0)} + \cancel{\frac{\partial f}{\partial x}((0,0,0))x} + \frac{\partial f}{\partial y}((0,0,0))y + \\ + \frac{\partial^2 f}{\partial x^2}((0,0,0))x^2 + O\left(xy, y^2, x^3\right) \Rightarrow$$

$$y = -\frac{\frac{\partial^2 f}{\partial x^2}((0,0,0))}{\frac{\partial f}{\partial y}((0,0,0))}x^2 + O\left(xy, y^2, x^3\right) \text{ for } (x, y) \in M_0.$$

where $O(xy, y^2, x^3)$ denotes the high order terms. Thus, we can approximate the critical manifold $M_0 = M_{a,0} \cup M_{r,0} = \{(x, y) \in \mathbb{R}^2 : f(x, y, 0) = 0\}$ inside the neighbourhood U as a parabola, where with $M_{a,0}$ and $M_{r,0}$ we denote the left and the right branches of the parabola, see figure 2.1. The condition $\frac{\partial^2 f}{\partial x^2}(0, 0, 0) > 0$ also implies that for $y > 0$, the left branch $M_{a,0}$ of the critical manifold is attractive for the *layer dynamic*, while the right branch $M_{r,0}$ is instead repelling. This derives directly from the qualitative analysis of the *layer problem*. Indeed fixing $y \in U$, the *layer problem* is composed of two equilibrium points inside U. They are points of the two branches of the parabola and, by condition $\frac{\partial f}{\partial y}(0, 0, 0) < 0$,

we argue that in the points (x, y) between the two branches we have that $f(x, y, 0)$ is negative, while it is zero in the equilibrium points, and it is positive outside the parabola, because $\frac{\partial^2 f}{\partial x^2}(0, 0, 0) > 0$. So, the left branch $M_{a,0}$ of the parabola is attracting instead the right branch $M_{r,0}$ is repelling for the *layer problem*. The origin is instead *non-hyperbolic*, a weakly attracting and a weakly repelling point. Finally, inside the critical manifold, the dynamic is governed by the *reduced dynamic*.

Thanks to the implicit function theorem the constrain $f(x, y, 0) = 0$ can be translated in a function of x, i.e. $y = h(x)$. Again by the implicit function theorem and the conditions of a *fold point* we have that $h'(x) = \frac{\partial h}{\partial x}(x) < 0$ for $x < 0$ and $h'(x) > 0$ for $x > 0$. The *reduced problem* becomes

$$h'(x(\tau))\,\dot{x} = g(x(\tau), h(x(\tau)), 0)$$

which is singular in $x = 0$. But in the first analysis, x is increasing before the *fold* point and x is decreasing after the *fold* point if $g(0, 0, 0)$ is negative vice versa if $g(0, 0, 0)$ is positive. At this point, it is possible to give the intuitive behaviour of the system trajectories at the *fold* point as depicted in figure 2.1. We expected that the trajectories are attracted driven by the *fast dynamic* to the left branch $M_{a,0}$ of the critical manifold, and they proceed along the

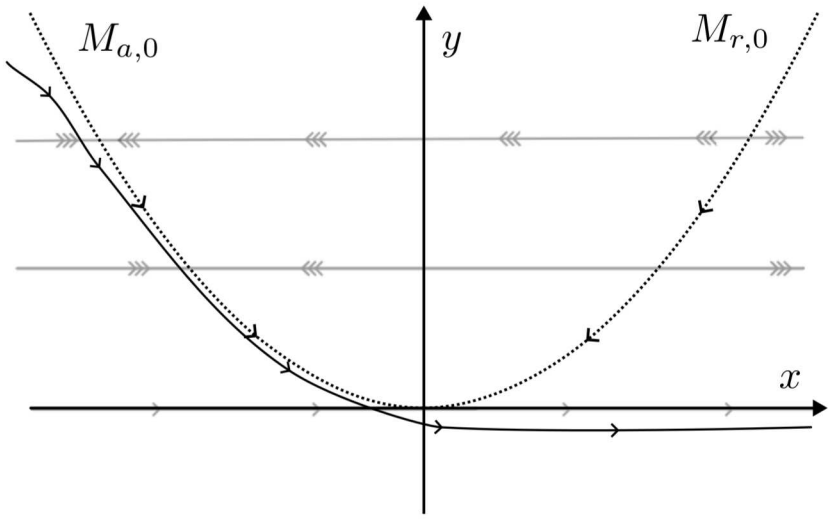


Figure 2.1: In the figure are plotted the critical manifold $M_0 = M_{a,0} \cap M_{r,0}$ where the *slow dynamics* is depicted as a dashed line; the grey line with the triple arrows represents the *fast dynamic*; the black line with a single arrow represents an hypothetical trajectory. In the x-axis the arrows denote that the origin is a weakly attracting and a weakly repelling point.

critical manifold to the *fold* point driven by the *reduced dynamic*, reached the *fold* point. It is possible to prove [10] rigorously that the trajectories follow the *fast dynamic* and tend to escape from the neighbourhood U. This situation typically generates an oscillation as we will see in the application of the theory to the FitzHugh-Nagumo model in the last part of this thesis.

Blow-up method

We conclude this section presenting a method typically used to extend the *geometric singular perturbation theory* for *non-hyperbolic* points.

Example 1: Consider the following planar system

$$\frac{d}{dt}x\left(t\right)=f\left(x\left(t\right)\right)=\begin{cases}\frac{d}{dt}x_1\left(t\right) & =x_2\left(t\right)\\ \frac{d}{dt}x_2\left(t\right) & =x_1\left(t\right)^3+x_1\left(t\right)x_2\left(t\right)\end{cases} \tag{2.17}$$

with $x\left(t\right)\in\mathbb{R}^2$. The origin $x=0$ is the unique equilibrium point and the behaviour of the linearised system at the origin is given by:

$$J\left(x\right)\big|_{x=0}=\frac{\partial f}{\partial x}\left(x\right)\big|_{x=0}=\begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}. \tag{2.18}$$

Thus the origin is a *non-hyperbolic* point and, moreover, it is nilpotent. Now, we want to qualitatively describe the orbits of (2.17) near the equilibrium point. The problem is that neither the linearisation around the origin nor centre manifold reduction are useful, since in this case the centre manifold corresponds to the whole phase-space. So the fundamental idea of the *blow-up* method is to induce a new system, with an appropriate change of coordinates, such that it has only *hyperbolic* equilibrium points and to study the new system with the standard techniques. Let consider a *weighted polar change of coordinates*

$$\Phi:\theta\times r\to\mathbb{R}^2\qquad\Phi\left(\theta,\,r\right)=\left(r\cos\theta,r^2\sin\theta\right)=\left(x_1,x_2\right) \tag{2.19}$$

with $r\in I\subseteq\mathbb{R}$, where I is an interval containing the origin, and $\theta\in[0,2\pi)$. It can be seen that this change of coordinates brings to the following form:

$$\begin{aligned}\frac{d}{dt}\theta\left(t\right)&=\frac{r\left(t\right)}{\sin^2\theta\left(t\right)+1}\left(1+\sin\theta\left(t\right)-4\sin^2\theta\left(t\right)-\sin^3\theta\left(t\right)+\sin^4\theta\left(t\right)\right)\\ \frac{d}{dt}r\left(t\right)&=\frac{r\left(t\right)^2}{\sin^2\theta\left(t\right)+1}\cos\theta\left(t\right)\sin\theta\left(t\right)\left(\sin\theta\left(t\right)-\sin^2\theta\left(t\right)+2\right)\end{aligned}. \tag{2.20}$$

The change of coordinate Φ maps the origin of the system (2.17) into the circle with $r\neq0$ and $\theta\in[0,2\pi)$. Note, however that (2.20) vanishes in the circle $r=0$, so the circle is a set of equilibrium point for (2.20). But it is possible to desingularize the vector field (2.20)

by a division by r. This operation does not change the qualitative dynamics on the set with $r = 0$ up to a time rescaling $\tau = \frac{t}{r}$ [11]. Nevertheless the rescaling does drastically change the dynamics on the circle identified for $r = 0$. So defining

$$\bar{\theta}(\tau) := \theta(t)\big|_{t=\frac{\tau}{r}} = \theta\left(\frac{\tau}{r}\right)$$
$$\bar{r}(\tau) := r(t)\big|_{t=\frac{\tau}{r}} = r\left(\frac{\tau}{r}\right)$$

The *desingularized system* becomes

$$\begin{aligned}\frac{d}{d\tau}\bar{\theta}(\tau) &= \frac{1}{\sin^2\bar{\theta}(\tau)+1}\left(1+\sin\bar{\theta}(\tau)-4\sin^2\bar{\theta}(\tau)-\sin^3\bar{\theta}(\tau)+\sin^4\bar{\theta}(\tau)\right)\\ \frac{d}{d\tau}\bar{r}(\tau) &= \frac{\bar{r}(\tau)}{\sin^2\bar{\theta}(\tau)+1}\cos\bar{\theta}(\tau)\sin\bar{\theta}(\tau)\left(\sin\bar{\theta}(\tau)-\sin^2\bar{\theta}(\tau)+2\right)\end{aligned} \tag{2.21}$$

It is now straightforward to show that (2.21) has four *hyperbolic saddle equilibrium* points,

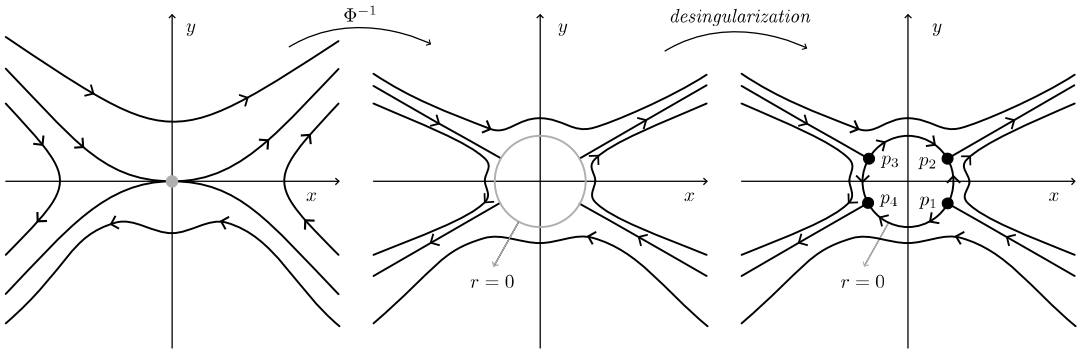


Figure 2.2: Sketch of the main steps of the *blow-up* method for example 1. Starting from the right figure: original vector field (2.17) with *non-hyperbolic* fixed points at the origin; *blow-up* vector field (2.20) with a full circle of fixed points; desingularized *blow-up* vector field (2.21) with precisely four *hyperbolic* saddle fixed points.

namely:

$$
\begin{aligned}
p_1 &= \left(-\arcsin\left(\sqrt{2}-1\right), 0\right) \\
p_2 &= \left(\arcsin\left(\frac{\sqrt{5}}{2}-\frac{1}{2}\right), 0\right) \\
p_3 &= \left(\pi-\arcsin\left(\frac{\sqrt{5}}{2}-\frac{1}{2}\right), 0\right) \\
p_4 &= \left(\pi+\arcsin\left(\sqrt{2}-1\right), 0\right)
\end{aligned}
$$

Since these equilibrium points are *hyperbolic*, it follows from linear analysis that the phase portrait of (2.21) in a small neighbourhood of the circle for $r = 0$ can be derived by the linearised system evaluated in the equilibrium points, and is as shown in figure 2.2.

The procedure exemplified above exploits the transformation Φ^{-1} to "blows the origin up to a circle". The main advantage is that with *blow-up*, the resulting system is simpler to analyse. Indeed, in the previous example we transform a nilpotent equilibrium point into four *hyperbolic* equilibrium points. Once the *blown-up* system is understood, it is possible to *blow-down* the phase-portrait of (2.21) resulting in a qualitative description of the original system (2.17).

2.2 Dissipativity

The theory of dissipative systems plays a fundamental role in the context of modern control and dynamical system analysis. This theory is presented in this thesis for its links with the problems of

- theory of electric networks;
- control and analysis of non linear systems;

Here, we present, following the derivation presented in[6], the formal concepts of dissipative system and we introduce the notions of *supply rate* and *energy* function.

The family of systems under consideration are :

$$\Sigma : \begin{cases} \dot{x}(t) &= f(x(t), u(t)) \\ y(t) &= g(x(t), u(t)) \end{cases} \quad t \in \mathbb{R} \tag{2.22}$$

where

$$\begin{aligned} x(t) &\in \mathcal{X} \coloneqq \mathbb{R}^n \\ u(t) &\in \mathcal{U} \coloneqq \mathbb{R}^m \\ y(t) &\in \mathcal{Y} \coloneqq \mathbb{R}^p \end{aligned}$$

In the following the functions $f : \mathcal{X} \times \mathcal{U} \to \mathcal{X}$ and $g : \mathcal{X} \times \mathcal{U} \to \mathcal{Y}$, with f sufficiently regular such that, for any initial condition $x_0 \in \mathcal{X}$ and for any input $u(t)$ piecewise-continuous defined in $[t_0, +\infty)$, there exists a unique $x(t)$ defined in $[t_0, +\infty)$ such that

$$\begin{cases} \dot{x}(t) &= f(x(t), u(t)) \\ x(t_0) &= x_0 \end{cases} \quad t \geq t_0$$

So given a continuous dynamical system Σ, it is possible define the following function, called supply rate.

Definition 1 (Supply rate)**.** A function

$$\begin{array}{rccc} w & : & \mathcal{U} \times \mathcal{Y} & \longrightarrow & \mathbb{R} \\ & & (u, y) & \mapsto & w(u, y) \end{array}$$

is called *supply rate* for Σ if, for any interval $[0, T]$, for any input $u(t)$ defined in $[0, T]$ and for any initial state $x_0 \in \mathcal{X}$, we have that:

$$\int_0^T |w(u(t), y(t))| \, dt < +\infty.$$

Where $y(t)$ it is the output generated by Σ from initial condition $x(0) = x_0$ and input $u(t)$.

To understand the concept of *supply rate*, it is interesting consider the following examples:

- Mechanical system
  Consider a body, depicted in figure 2.3, under the action of a set of forces such that the resulting force applied to the centre of gravity is $F(t) \in \mathbb{R}^3$, function of the time. From the laws of mechanics it is possible to derive a dynamical system of the form:

$$\begin{cases} \dot{x}(t) &= f(x(t), F(t)) \\ v(t) &= g(x(t)) \end{cases} \qquad t \in \mathbb{R}$$

  where $v(t)$ it is the velocity of the centre of mass of the body. A natural *supply rate* for this system is the mechanical power:

$$w(F, v) := F^T v;$$

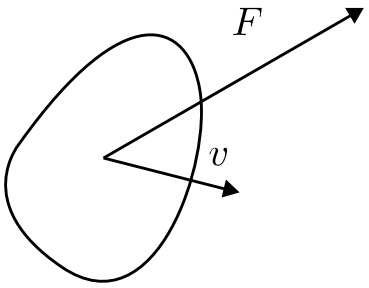


Figure 2.3: Simple mechanical example in which the *supply rate* is $w(F, v) := F^T v$.

- Electrical system
  Consider the simple RLC circuit in figure 2.4. The *supply rate*:

$$w(V, I) = VI$$

  is the electrical power absorbed by the network.

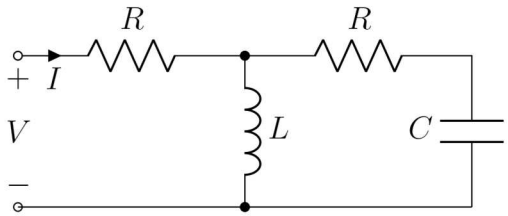


Figure 2.4: Simple RLC circuit. The *supply rate* in this case is $w(V, I) = VI$, the electrical power absorbed by the network.

In other words, the *supply rate* encodes the rate at which a dynamical system exchanges *energy* with the environment. Typically the *energy* has the usual physical meaning, but can also be generalized and formalized with the following definition.

Definition 2. A dynamical system Σ it is called *dissipative* with respect to the *supply rate* w if there is a function

$$S : \mathcal{X} \to \mathbb{R}$$

such that:

1. $S(x) \geq 0$ for any $x \in \mathcal{X}$;
2. $\exists\, x^* \in \mathcal{X}$ such that $S(x^*) = 0$;
3. for any $T \geq 0$, for any input u defined in $[0, T]$ and for any initial state $x(0) \in \mathcal{X}$ one has that:

$$\int_0^T w(u(t), y(t))\, dt \geq S(x(T)) - S(x(0)), \tag{2.23}$$

where $x(t)$ and $y(t)$ are respectively the state and the output at the time t of the system Σ with input $u(t)$.

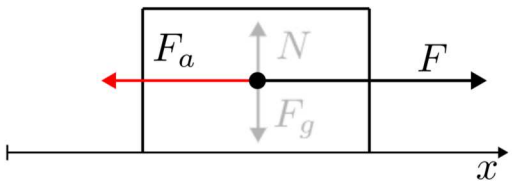


Figure 2.5: Simple mechanical example of a *dissipative* system.

Example 2: Consider the physical system illustrated in figure 2.5. From the laws of mechanics it is possible to derive the following equations of motion

$$\begin{aligned} m\ddot{x}(t) &= F + F_a(t) \\ -mg\mu_k sign(\dot{x}(t)) &= F_a(t) \end{aligned} \tag{2.24}$$

where $x(t)$ represents the horizontal position of the block, $\dot{x}(t)$ its velocity, $\ddot{x}(t)$ its acceleration, μ_k the static constant friction equal to the dynamic one and $sign(\cdot)$ the sign function. Then setting

$$\begin{aligned} x_1(t) &= x(t) \\ x_2(t) &= \dot{x}(t) \\ u(t) &= F(t). \end{aligned}$$

It $\mathcal{U} = \{u(t) = F \,\forall\, t \in \mathbb{R} : F \in \mathbb{R} \setminus [-mg\mu_k, mg\mu_k]\}$, it is possible to rewrite equation (2.24) the following dynamical system

$$\begin{cases} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= -sign(x_2(t))\, g\mu_k + \frac{u}{m} \\ y(t) &= x_2(t) \end{cases}.$$

In this case the natural *supply rate* is the mechanical work made by the force F that is:

$$w(u, y(t)) = uy(t) = Fx_2(t).$$

And the natural *energy* function $S(x(t))$ is the kinetic energy of the system

$$S(x(t)) = S(x_2(t)) = \frac{m}{2} x_2(t)^2.$$

Indeed, for any $T \geq 0$, input $u = F$ defined in $[0, T]$ and for any initial state $x(0) \in \mathcal{X}$ we

have:

$$\begin{aligned}\int_0^T w\left(u, y\left(t\right)\right)\, dt &= \int_0^T F x_2\left(t\right)\, dt = \int_0^T m x_2\left(t\right) \dot{x}_2\left(t\right)\, dt + \\ &\qquad + \int_0^T x_2\left(t\right) sign\left(x_2\left(t\right)\right) mg\mu_k\, dt \\ &= \int_{x_2(0)}^{x_2(T)} m x_2\, dx_2 + mg\mu_k \int_0^T \left|x_2\left(t\right)\right|\, dt \\ &= \underbrace{m\, \frac{x_2^2}{2}\Big|_{x_2(0)}^{x_2(T)}}_{S(x_2(T)) - S(x_2(0))} + \underbrace{mg\mu_k \int_0^T \left|x_2\left(t\right)\right|\, dt}_{\geq 0} \\ &\geq S\left(x_2\left(T\right)\right) - S\left(x_2\left(0\right)\right)\end{aligned} \tag{2.25}$$

The final inequality of (2.25) proves (2.23) and so ensures the *dissipativity* of the considered system.

Dissipativity is a useful instrument to study many physical and mathematical systems. In definition 2 it is required that the *energy* function must be *non-negative*. This is strictly linked to the fact that the *energy function* defined in 2 represents a physical energy, as the kinetic energy, and so typically it is always *non-negative*. But, also from physics, it is useful dealing with *energy* functions that can take negative values. For this reason it is possible , in addition to the notion of *dissipativity*, to introduce the concept of *generalized dissipativity*.

Definition 3. A dynamical system Σ it is called *generalized dissipative* with respect to the *supply rate* w if there exists a function

$$S : \mathcal{X} \to \mathbb{R}$$

such that:

1. for any $T \geq 0$, for any input u defined in $[0, T]$ and for any initial state $x\left(0\right) \in \mathcal{X}$ one has that:

$$\int_0^T w\left(u\left(t\right), y\left(t\right)\right) dt \geq S\left(x\left(T\right)\right) - S\left(x\left(0\right)\right), \tag{2.26}$$

where $x\left(t\right)$ and $y\left(t\right)$ are respectively the state and the output at the time t of the system Σ driven by the input $u\left(t\right)$;

such function is called *signed energy*.

It is immediate to see that a *dissipative* system is *generalized dissipative* but not the vice-versa. The difference between *dissipativity* and *generalized dissipativity* is that in this second case the *energy* function can assume negative values. As a consequence, it can happen that from a *generalized dissipative* system infinite energy can be extracted in some states.

The following example represent the example of a *generalized dissipative* system.

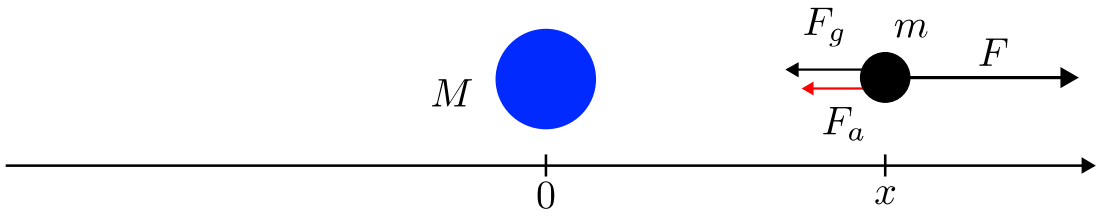


Figure 2.6: Simple mechanical example of a *generalized dissipative* system. The mass m is attracted by the fixed mass M by the gravitational force F_g. F_a is the viscous friction and F is the actuating force.

Example 3: Consider the system presented in figure 2.6. It consist in a 1-dimensional system with two masses m and M, with the body of mass M fixed in the origin of the reference frame. The forces acting on the body m are:

- F_g the gravitational force;
- F_a a viscous friction force;
- F a input force acting on m.

From the physics of the system it is possible to derive the following equation of motion:

$$m\ddot{x}(t) = F_g(t) + F_a(t) + F(t) \tag{2.27}$$

with

$$\begin{aligned} F_g &= -\gamma sign(x(t)) \frac{Mm}{x(t)^2} \\ F_a &= -\nu \dot{x}(t) \\ F &= u(t) \end{aligned}$$

and ν and γ two appropriate constants.

Now, as before, by setting

$$\begin{aligned} x_1(t) &= x(t) \\ x_2(t) &= \dot{x}(t) \\ u(t) &= F(t) \end{aligned}$$

it is possible to derive the following two-dimensional system:

$$\begin{cases} \dot{x}_1(t) & = x_2(t) \\ \dot{x}_2(t) & = -\gamma sign(x_1(t)) \frac{M}{x_1(t)^2} - \nu \frac{x_2(t)}{m} + \frac{u(t)}{m} \\ y(t) & = x_2(t) \end{cases} . \tag{2.28}$$

As in the previous example the natural *supply rate* is

$$w(u(t), y(t)) = u(t) y(t) = F(t) x_2(t) .$$

In this case the total mechanical energy, that is the sum of the *potential gravitation energy* $U(x_1(t))$ and the *kinetic energy* $K(x_2(t))$, is a *signed energy* function $S(x(t)) = K(x(t)) + U(x(t))$. Where

$$\begin{aligned} K(x_2) &= \frac{m x_2^2}{2} \\ U(x_1) &= Mm\gamma \int_{x_1(0)}^{x_1(T)} \frac{1}{|x_1| x_1} dx_1 \end{aligned}$$

proof of inequality (2.26), to verify the *generalized dissipativity* of system (2.28), is presented in the following. Indeed for any $T \geq 0$, input $u(t) = F(t)$ defined in $[0, T]$ and for any

initial state $x(0) \in \mathcal{X}$:

$$\begin{aligned}
\int_0^T w(u(t), y(t))\, dt &= \int_0^T F(t)\, x_2(t)\, dt = \int_0^T m x_2(t)\, \dot{x}_2(t)\, dt + \\
&\quad + \int_0^T \gamma sign(x_1(t)) \frac{Mm}{x_1(t)^2} x_2(t)\, dt + \\
&\quad + \int_0^T \nu x_2(t)^2\, dt \\
&= \int_{x_2(0)}^{x_2(T)} m x_2\, dx_2 + \int_0^T \gamma sign(x_1(t)) \frac{Mm}{x_1(t)^2} \dot{x}_1(t)\, dt + \\
&\quad + \int_0^T \nu x_2(t)^2\, dt \\
&= \underbrace{\frac{m x_2^2}{2}\Big|_{x_2(0)}^{x_2(T)}}_{K(x_2(T)) - K(x_2(0))} + \underbrace{\int_{x_1(0)}^{x_1(T)} \gamma sign(x_1(t)) \frac{Mm}{{x_1}^2}\, dx_1}_{U(x_1(T)) - U(x_1(0))} + \\
&\quad + \int_0^T \nu x_2(t)^2\, dt \\
&= S(x(T)) - S(x(0)) + \underbrace{\int_0^T \nu x_2(t)^2\, dt}_{\geq 0} \\
&\geq S(x(T)) - S(x(0))
\end{aligned} \tag{2.29}$$

Condition (2.29) proves *generalized dissipativity*, but not the previous definition of *dissipativity*. This is because the *energy* function $S(x(t))$ in this case can be negative due to the presence of the potential energy $U(x_1(t))$ that can be arbitrarily negative.

Dissipativity and *generalized dissipativity* are the fundamental notions on which dissipativity theory is founded. In this section these notions are introduced in their general version, valid also for non linear systems. In the next paragraph the theory is particularized for reachable linear systems.

2.2.1 Linear Dissipative Systems

In this section the dynamical systems considered are of the form:

$$\Sigma : \begin{cases} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{cases} \quad t \in \mathbb{R} \tag{2.30}$$

and the *supply rate* $w(u(t), y(t))$ used are of the form:

$$w(u(t), y(t)) = \begin{bmatrix} y(t)^T & u(t)^T \end{bmatrix} \begin{bmatrix} \bar{Q} & \bar{S} \\ \bar{S}^T & \bar{R} \end{bmatrix} \begin{bmatrix} y(t) \\ u(t) \end{bmatrix} \tag{2.31}$$

Since $y(t) = Cx(t) + Du(t)$, it is possible consider $w(u(t), y(t))$ as a function of the input $u(t)$ and the state $x(t)$ so $w(u(t), y(t)) = w(u(t), x(t))$ instead of function of the input $u(t)$ and the output $y(t)$. Indeed

$$w(u(t), y(t)) = w(u(t), x(t)) = \begin{bmatrix} x(t)^T & u(t)^T \end{bmatrix} \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}$$

with

$$\begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} = \begin{bmatrix} C^T\bar{Q}C & C^T\bar{Q}D + C^T\bar{S} \\ D^T\bar{Q}C + \bar{S}^TC & D^T\bar{Q}D + \bar{S}^TD + D^T\bar{S} + \bar{R} \end{bmatrix}.$$

Usually, as done in the examples before, the verification of *dissipativity* is not simple and requires to verify an inequality along any possible trajectory of the system. In the linear, case instead, the problem is typically much more tractable. Indeed, with the following theorem it is possible to prove the *generalized dissipativity* by solving only a Linear Matrix Inequality (LMI).

Before presenting this important theorem, it is necessary to introduce some fundamental preliminary concepts and results.

Definition 4 (Required energy function (starting from x^*))**.** Given a dynamical system Σ, a *supply rate* w and a state $x^* \in \mathcal{X}$, the *required energy function* (starting from x^*) is the map

$$S_{r,x^*} : \mathcal{X} \to \mathbb{R} \cup \{+\infty, -\infty\}$$

defined for any $x^* = x(0) \in \mathcal{X}$ as

$$S_{r,x^*}(x) := \begin{cases} +\infty & \text{if } x \text{ is not reachable from } x^* \\ \inf_{T\geq 0,\, u} \left\{ \int_0^T w(u(t), y(t)) \right\} & \text{if } x \text{ is reachable from } x^* \end{cases}$$

where the inf is computed from $x(0) = x^*$ and $u(t)$ that drives $x(t)$ from $x(0) = x^*$ to $x(T) = x$.

Intuitively S_{r,x^*} consists in the minimum energy necessary to drive Σ from x^* to x.

Definition 5 (Available energy function (ending in x^*))**.** Given a dynamical system Σ, a *supply rate* w and a state $x^* \in \mathcal{X}$. Then the *available energy function* (ending in x^*) is the map

$$S_{d,x^*} : \mathcal{X} \to \mathbb{R} \cup \{+\infty, -\infty\}$$

defined for any $x^* = x(0) \in \mathcal{X}$ as

$$S_{d,x^*}(x) := \begin{cases} -\infty & \text{if } x^* \text{ is not reachable from } x \\ \sup_{T\geq 0,\, u} \left\{ -\int_0^T w(u(t), y(t)) \right\} & \text{if } x^* \text{ is reachable from } x \end{cases}$$

where the inf is computed from $x(0) = x$ and $u(t)$ that drives $x(t)$ from $x(0) = x$ to $x(T) = x^*$.

Instead here $S_{d,x^*}(x)$ coincides with the maximum available energy that can be extracted from the system when Σ is driven from x to x^*.

Now it is presented another technical proposition that will be used soon. The proof is not presented here but it can be found in [6].

Proposition 1. *Let $X = \mathbb{R}^n$ and $F : X \to \mathbb{R}$. Then the following assertions are equivalent:*

1. *the function F is quadratic, namely there exists a matrix $M \in \mathbb{R}^{n\times n}$ such that $F(x) = x^T M x$;*
2. *function F is continuous in 0 and satisfies the following identity (the parallelogram identity)*

$$F(x+y) + F(x-y) = 2F(x) + 2F(y), \quad \forall\, x, y \in X \tag{2.32}$$

With this proposition it is possible prove the following important lemma:

Lemma 1. *Let (Σ, w) be a linear, reachable and generalized dissipative system with a quadratic supply rate. Then $S_{d,0}(x)$ and $S_{r,0}(x)$ are quadratic functions and so there exist symmetric $\Pi_{d,0},\ \Pi_{r,0} \in \mathbb{R}^{n\times n}$ such that*

$$S_{d,0} = x^T \Pi_{d,0} x, \qquad S_{r,0} = x^T \Pi_{r,0} x \qquad \forall\, x \in \mathbb{R}^n$$

Proof. The proof for $S_{d,0}$ and $S_{r,0}$ are entirely similar so here the property is proved only for $S_{r,0}$. To do that it is possible to use proposition 1. So to prove that $S_{r,0}$ is quadratic it is equivalent to prove that:

1. $S_{r,0}$ is continuous in 0;
2. $S_{r,0}$ satisfy the parallelogram identity (2.32);

To prove the continuity at the origin it is sufficient to prove that there exist two matrices M and N such that

$$-z^T N z \leq S_{r,0}(x) \leq z^T M z \tag{2.33}$$

for all $z \in \mathbb{R}^n$.

First $S_{r,0}(0) = 0$ because $x^* = x = 0$ in definition 4. Let $\{e_1, \ldots, e_n\}$ be the canonical basis of $\mathbb{R}^n$. Then, by the reachability of the system there exist inputs $u_1, \ldots, u_n$, defined in $[0, T]$ such that u_i generates a state evolution $x_i(t)$ with $x_i(0) = 0$ and $x_i(T) = e_i$. Let $z \in \mathbb{R}^n$. Then $z = \sum_{i=1}^n z_i e_i$, where z_i is the i-th component of the vector z. But, for the linearity of the system, $u := \sum_{i=1}^n z_i u_i$ generates a state evolution $x(t)$ such that $x(0) = 0$ and $x(T) = z$. Then

$$\begin{aligned}\int_0^T w(u(t), x(t))\, dt &= \int_0^T w\left(\sum_{i=1}^n z_i u_i(t), \sum_{i=1}^n z_i x_i(t)\right) dt \\ &= \int_0^T \sum_{i=1}^n \sum_{j=1}^n z_i z_j \begin{bmatrix} x_i^T(t) & u_i^T(t) \end{bmatrix} \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \begin{bmatrix} x_j(t) \\ u_j(t) \end{bmatrix} = z^T M z\end{aligned}$$

where $M \in \mathbb{R}^{n\times n}$ has elements

$$M_{ij} := \int_0^T \begin{bmatrix} x_i^T(t) & u_i^T(t) \end{bmatrix} \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \begin{bmatrix} x_j(t) \\ u_j(t) \end{bmatrix} dt$$

Then the evolution from $0 \to z$ is such that

$$z^T M z = \int_0^T w\left(u\left(t\right), x\left(t\right)\right) dt \geq S_{r,0}\left(z\right) - S_{r,0}\left(0\right) = S_{r,0}\left(z\right).$$

With an analogous reasoning, we can find a set of inputs such that the evolution of the state starting from e_i goes to 0. It is possible to find a matrix N such that

$$z^T N z \geq S_{r,0}\left(0\right) - S_{r,0}\left(z\right) = -S_{r,0}\left(z\right)$$

To verify the identity (2.32), first consider the following fact

$$\int_0^T w\left(2u\left(t\right), 2x\left(t\right)\right) dt = 4\int_0^T w\left(u\left(t\right), x\left(t\right)\right) dt$$

which is true in the case that the first evolution starts from 0 and goes to $2x$ and the second from 0 to x. Hence

$$S_{r,0}\left(2x\right) = 4S_{r,0}\left(x\right). \tag{2.34}$$

Let now $x_1, x_2 \in \mathbb{R}^n$. For any $\epsilon > 0$ there exist $T_1, T_2 \geq 0$ and inputs u_1, u_2 such that

$$S_{r,0}\left(x_1\right) + \epsilon > \int_0^{T_1} w\left(u_1\left(t\right), x_1\left(t\right)\right) dt$$
$$S_{r,0}\left(x_2\right) + \epsilon > \int_0^{T_2} w\left(u_2\left(t\right), x_2\left(t\right)\right) dt.$$

Choose $T \geq max\left\{T_1, T_2\right\}$ and define the inputs $\bar{u}_1$ and $\bar{u}_2$ as

$$\bar{u}_1\left(t\right) = \begin{cases} 0 & \text{if } 0 \leq t \leq T - T_1 \\ u_1\left(t - T + T_1\right) & \text{if } T - T_1 < t \leq T \end{cases}$$
$$\bar{u}_2\left(t\right) = \begin{cases} 0 & \text{if } 0 \leq t \leq T - T_2 \\ u_2\left(t - T + T_2\right) & \text{if } T - T_1 < t \leq T \end{cases}$$

Let $\bar{x}_1, \bar{x}_2$ be the forced evolution of the states corresponding to inputs $\bar{u}_1, \bar{u}_2$. It is easy to

verify that $\bar{x}_1(T) = x_1, \bar{x}_2(T) = x_2$ and that

$$\int_0^{T_1} w(u_1(t), x_1)\, dt = \int_0^{T} w(\bar{u}_1(t), \bar{x}_1(t))\, dt \qquad \text{with } 0 \to x_1$$
$$\int_0^{T_2} w(u_2(t), x_2)\, dt = \int_0^{T} w(\bar{u}_2(t), \bar{x}_2(t))\, dt \qquad \text{with } 0 \to x_2.$$

Consider now the inputs $\bar{u}_1 + \bar{u}_2$ and $\bar{u}_1 - \bar{u}_2$ for what, by the linearity, correspond the evolutions of the states $\bar{x}_1 + \bar{x}_2$ and $\bar{x}_1 - \bar{x}_2$. Then

$$\begin{aligned}
S_{r,0}(x_1 + x_2) + S_{r,0}(x_1 - x_2) \leq & \underbrace{\int_0^T w(\bar{u}_1(t) + \bar{u}_2(t), \bar{x}_1(t) + \bar{x}_2(t))\, dt}_{0 \to x_1 + x_2} + \\
& + \underbrace{\int_0^T w(\bar{u}_1(t) - \bar{u}_2(t), \bar{x}_1(t) - \bar{x}_2(t))\, dt}_{0 \to x_1 - x_2} \\
= & \int_0^T w(\bar{u}_1(t) + \bar{u}_2(t), \bar{x}_1(t) + \bar{x}_2(t)) \quad + \\
& + w(\bar{u}_1(t) - \bar{u}_2(t), \bar{x}_1(t) - \bar{x}_2(t))\, dt \\
= & 2 \left\{ \underbrace{\int_0^T w(\bar{u}_1(t), \bar{x}_1(t))\, dt}_{0 \to x_1} + \underbrace{\int_0^T w(\bar{u}_2(t), \bar{x}_2(t))\, dt}_{0 \to x_2} \right\} \\
< & 2 \{ S_{r,0}(x_1) + S_{r,0}(x_2) + 2\epsilon \}.
\end{aligned}$$

Since ϵ is arbitrary, it is possible to conclude that

$$S_{r,0}(x_1 + x_2) + S_{r,0}(x_1 - x_2) \leq 2 \{ S_{r,0}(x_1) + S_{r,0}(x_2) \}.$$

Now let $z_1 \coloneqq x_1 + x_2$ and $z_2 \coloneqq x_1 - x_2$. With the previous formula it is possible to obtain

$$\begin{aligned}
S_{r,0}(2x_1) + S_{r,0}(2x_2) = S_{r,0}(z_1 + z_2) + S_{r,0}(z_1 - z_2) \leq \\
\leq \{ S_{r,0}(z_1) + S_{r,0}(z_2) \} = 2 \{ S_{r,0}(x_1 + x_2) + S_{r,0}(x_1 - x_2) \}
\end{aligned}$$

Using (2.34) :

$$2\left\{S_{r,0}\left(x_1\right)+S_{r,0}\left(x_2\right)\right\} \leq S_{r,0}\left(x_1+x_2\right)+S_{r,0}\left(x_1-x_2\right)$$

and so

$$2\left\{S_{r,0}\left(x_1\right)+S_{r,0}\left(x_2\right)\right\} = S_{r,0}\left(x_1+x_2\right)+S_{r,0}\left(x_1-x_2\right)$$

□

With this preliminary lemma it is possible to introduce the previous cited theorem.

Theorem 1. *Let* (Σ, w) *a linear reachable system with quadratic supply rate. Then the following conditions are equivalent:*

1. (Σ, w) *is generalized dissipative;*
2. (Σ, w) *is generalized dissipative with quadratic energy function;*
3. *the matrix inequality*

$$\begin{bmatrix} Q-A^TP-PA & S-PB \\ S^T-B^TP & R \end{bmatrix} \geq 0 \tag{2.35}$$

admit a symmetric solution P.

To every solution $P = P^T$ *of* (2.35) *corresponds a quadratic energy function with sign* $S\left(x\right) = x^TPx$ *for the system and vice-versa for any quadratic energy function* $S\left(x\right) = x^TPx$ *for* (Σ, w) *corresponds a solution* $P = P^T$ *to* (2.35).

Proof.

2. $\Rightarrow$ **1.** It is trivial.

1. $\Rightarrow$ **2.** It is consequence of lemma 1.

2. $\Rightarrow$ **3.** Suppose that the system (Σ, w) is *generalized dissipative* with quadratic *signed energy* function $S\left(x\right) = x^TPx$. Then for any initial state $x\left(0\right)$ and for any $h > 0$:

$$\int_0^h w\left(u\left(t\right), x\left(t\right)\right) dt \geq x^T\left(h\right) Px\left(h\right) - x^T\left(0\right) Px\left(0\right) \qquad \text{for } x\left(0\right) \to x\left(h\right).$$

Dividing both side of the equation by h and taking the limit for h going to zero, the following result holds

$$w\left(u\left(0\right),x\left(0\right)\right)\geq\left(\frac{d}{dt}x^{T}\left(t\right)Px\left(t\right)\right)_{t=0}$$

and so

$$\begin{aligned}w\left(u\left(0\right),x\left(0\right)\right)&\geq\dot{x}^{T}\left(0\right)Px+x^{T}\left(0\right)P\dot{x}\left(0\right)\\&=\begin{bmatrix}x^{T}\left(0\right) & u^{T}\left(0\right)\end{bmatrix}\begin{bmatrix}A^{T}P+PA & PB\\B^{T}P & 0\end{bmatrix}\begin{bmatrix}x\left(0\right)\\u\left(0\right)\end{bmatrix}.\end{aligned}\tag{2.36}$$

Since inequality (2.36) holds for any $x\left(0\right)\in\mathbb{R}^{n}$ and for any $u\left(0\right)\in\mathbb{R}^{m}$, it follows that

$$\begin{bmatrix}Q-A^{T}P-PA & S-PB\\S^{T}-B^{T}P & R\end{bmatrix}\geq 0.$$

3. $\Rightarrow$ **2.** Suppose that $P=P^{T}$ satisfies (2.35). Also suppose that $u\left(t\right)$ is an arbitrary admissible input defined on $[0,T]$ and that $x\left(t\right)$ is the corresponding state evolution of the system Σ starting from x_{0}. Then for any $t\in[0,T]$ it holds:

$$\begin{aligned}&\begin{bmatrix}x^{T}\left(t\right) & u^{T}\left(t\right)\end{bmatrix}\begin{bmatrix}Q & S\\S^{T} & R\end{bmatrix}\begin{bmatrix}x\left(t\right)\\u\left(t\right)\end{bmatrix}\geq\\&\quad\geq x^{T}\left(t\right)A^{T}Px\left(t\right)+u\left(t\right)B^{T}Px\left(t\right)+x^{T}\left(t\right)PAx\left(t\right)+x^{T}\left(t\right)PBx\left(t\right)\end{aligned}$$

and from the definition of the supply rate and the equations of the system it follows that

$$w\left(u\left(t\right),x\left(t\right)\right)\geq\dot{x}^{T}\left(t\right)Px\left(t\right)+x^{T}\left(t\right)P\dot{x}\left(t\right)=\frac{d}{dt}x^{T}\left(t\right)Px\left(t\right)$$

By integrating on $[0,T]$

$$\underbrace{\int_{0}^{T}w\left(u\left(t\right),y\left(t\right)\right)dt}_{x_{0}\rightarrow x(T)}\geq x^{T}\left(T\right)Px\left(T\right)-x^{T}\left(0\right)Px\left(0\right)$$

which implies that $S(x) \coloneqq x^T P x$ is a *signed energy* function for (Σ, w).

□

Typically, the analysis of high-dimensional systems is complex and laborious, but the analysis is much simpler if these have a low-dimensional *dominant* behaviour for which it is possible to perform a model reduction and a simplified analysis. A well known situation is the linear case, where frequently few *dominant* poles capture the main properties of a possibly high-dimensional system. In this section we try to generalize this idea to the *non-linear* setting, where *multi-stability* or *limit cycles* can arise. It is not immediate to deal with these phenomena so we need to introduce some other tools.

This can be done by using *differential* analysis and linear *dissipation* inequalities. We will use this approach to derive a *p*-dominance theory, a theory that studies the convergence of the *non-linear* system to a *p*-dimensional dominant subspace.

In what follows, we introduce the preliminary concepts necessary for the p-dominant analysis. First, it is necessary to define the *inertia* of a matrix.

Definition 6. Let A be a symmetric matrix with real entries. Then A has *inertia* (p, z, m) if it has p negative eigenvalues, z eigenvalues equal to zero and m positive eigenvalues.

In the *p*-dominance theory, matrices of *inertia* $(p, 0, n - p)$ are fundamental. So, in order to simplify the notations, these matrices are denoted by matrices of *inertia* p.

In this thesis, *p*-dominance is first presented in the linear case and then extended to the non-linear framework.

2.3.1 Dominant Linear Time Invariant Systems

In this section we introduce the concept of *p*-dominance for LTI systems:

Definition 7. A linear system $\dot{x}(t) = Ax(t)$ is *p*-dominant if there exists a symmetric matrix P with *inertia p* such that

$$A^T P + PA \leq -\epsilon I \tag{2.37}$$

for some $\epsilon \geq 0$. The property is strict if $\epsilon > 0$.

Now, with the following proposition, it is possible to give other characterizations of the *p*-dominance. Concerning the algebraic condition (2.37), the next proposition links the *p*-dominance to the behaviour of the trajectories of the system.

Proposition 2. *The three following facts are equivalent:*

1. *a linear system $\dot{x}(t) = Ax(t)$ is strictly p-dominant;*
2. *the matrix A has p eigenvalues with strictly positive real part and $n - p$ eigenvalues with strictly negative real part;*
3. *there exists an invariant splitting of the vector space $\mathbb{R}^n = \mathcal{H}_p \oplus \mathcal{V}_{n-p}$ into dominant $\mathcal{H}_p$ and non-dominant $\mathcal{V}_{n-p}$ eigenspaces such that any solution of the linear system $\dot{x}(t) = Ax(t)$ can be written as $x(t) = x_p(t) + x_{n-p}(t)$ with $x_p(t) \in \mathcal{H}_p$, $x_{n-p}(t) \in \mathcal{V}_{n-p}$. Furthermore, there exist constants $0 < \underline{C} \leq 1 \leq \overline{C}$ and $\underline{\lambda} < 0 < \overline{\lambda}$ such that,*

$$\begin{aligned} \|x_p(t)\| &\geq \underline{C} e^{-\underline{\lambda} t} \|x_p(0)\| \\ \|x_{n-p}(t)\| &\leq \overline{C} e^{-\overline{\lambda} t} \|x_{n-p}(0)\| \end{aligned} \tag{2.38}$$

Proof.

2. $\Rightarrow$ **1.** We can consider the Jordan form associated to A (2.39). Without loss of generality, suppose that the Jordan form takes the following shape, with $J_u \in \mathbb{R}^{p \times p}$ and $J_s \in \mathbb{R}^{(n-p)\times(n-p)}$:

$$J = T^{-1} A T = \left[\begin{array}{c|c} J_u & 0 \\ \hline 0 & J_s \end{array}\right] \tag{2.39}$$

where

$$\mathcal{H}_p \coloneqq \langle v_1, \dots, v_p \rangle$$
$$\mathcal{V}_{n-p} \coloneqq \langle v_{p+1}, \dots, v_n \rangle$$

and

$$T = \left[\begin{array}{ccc|ccc} & & & & & \\ v_1 & \cdots & v_p & v_{p+1} & \cdots & v_n \\ & & & & & \end{array}\right],$$

with $\{v_1, \dots, v_n\}$ *generalized eigenvectors* and J_u is the block matrix containing all the Jordan blocks associated with the eigenvalues with negative real part and J_s is the one associated to the eigenvalues with positive real part. Then, because the matrices $-J_u$ and J_s are asymptotically stable, there exist two positive definite matrices P_u and P_s such that :

$$\begin{aligned} J_s^T P_s + P_s J_s &< 0 \\ \left(-J_u\right)^T P_u + P_u \left(-J_u\right) &< 0 \end{aligned} \tag{2.40}$$

with $P_u \in \mathbb{R}^{p \times p}$ and $P_s \in \mathbb{R}^{(n-p)\times(n-p)}$. Defining P as:

$$P \coloneqq \left[\begin{array}{c|c} -P_u & 0 \\ \hline 0 & P_s \end{array}\right]$$

it is immediate to verify that

$$\begin{aligned} J^T P + PJ &= \left[\begin{array}{c|c} J_u^T & 0 \\ \hline 0 & J_s^T \end{array}\right] \left[\begin{array}{c|c} -P_u & 0 \\ \hline 0 & P_s \end{array}\right] + \left[\begin{array}{c|c} -P_u & 0 \\ \hline 0 & P_s \end{array}\right] \left[\begin{array}{c|c} J_u & 0 \\ \hline 0 & J_s \end{array}\right] \\ &= \left[\begin{array}{c|c} -J_u^T P_u + \left(-P_u J_u\right) & 0 \\ \hline 0 & J_s^T P_s + P_s J_s \end{array}\right] < 0 \end{aligned}$$

where the last inequality derives directly from (2.40). Then

$$\begin{aligned} J^T P + PJ &= \left(T^{-1}AT\right)^T P + PT^{-1}AT = T^T A^T \left(T^{-1}\right)^T P + PT^{-1}AT \\ &= A^T \underbrace{\left(T^{-1}\right)^T PT^{-1}}_{P'} + \underbrace{\left(T^{-1}\right)^T PT^{-1}}_{P'} A < 0 \end{aligned} .$$

Now

$$P' = \left(T^{-1}\right)^T P T^{-1} \Rightarrow \left(P'\right)^T = \left(T^{-1}\right)^T P T^{-1} = P' \tag{2.41}$$

and so also P' is symmetric. Moreover also P' has *inertia p*. This is because P' is congruent to P and by Sylvester's law of inertia [12] two congruent symmetric matrices have the same *inertia*.

1. $\Rightarrow$ **2.** Consider now a change of basis $A = TJT^{-1}$. Such that

$$J = \left[\begin{array}{c|c|c} J_u & 0 & 0 \\ \hline 0 & J_0 & 0 \\ \hline 0 & 0 & J_s \end{array}\right] \tag{2.42}$$

where $J_u \in \mathbb{R}^{r\times r}$, $J_0 \in \mathbb{R}^{w\times w}$ and $J_s \in \mathbb{R}^{q\times q}$ with $r+q+w=n$. With J_u, J_0 and J_s, respectively, the Jordan blocks associated to the eigenvalues with positive, zero or negative real part. As before, inequality (2.37) imply the existence of a symmetric matrix

$$P = \left[\begin{array}{c|c|c} P_u & \star & \star \\ \hline \star & P_0 & \star \\ \hline \star & \star & P_s \end{array}\right]$$

with $P_u \in \mathbb{R}^{r\times r}$, $P_0 \in \mathbb{R}^{w\times w}$ and $P_s \in \mathbb{R}^{q\times q}$. Such that

$$J^T P + PJ < 0 \tag{2.43}$$

holds. First we prove that (2.43) implies that $w = 0$. Consider

$$\left[\begin{array}{c|c|c} J_u^T & 0 & 0 \\ \hline 0 & J_0^T & 0 \\ \hline 0 & 0 & J_s^T \end{array}\right] \left[\begin{array}{c|c|c} P_u & \star & \star \\ \hline \star & P_0 & \star \\ \hline \star & \star & P_s \end{array}\right] + \left[\begin{array}{c|c|c} P_u & \star & \star \\ \hline \star & P_0 & \star \\ \hline \star & \star & P_s \end{array}\right] \left[\begin{array}{c|c|c} J_u & 0 & 0 \\ \hline 0 & J_0 & 0 \\ \hline 0 & 0 & J_s \end{array}\right] < 0$$

$$\left[\begin{array}{c|c|c} J_u^T P_u + P_u J_u & \star & \star \\ \hline \star & J_0^T P_0 + P_0 J_0 & \star \\ \hline \star & \star & J_s^T P_s + P_s J_s \end{array}\right] < 0 \quad ,$$

but, by Sylvester's criterion this entails:

$$J_u^T P_u + P_u J_u < 0 \tag{2.44a}$$
$$J_0^T P_0 + P_0 J_0 < 0 \tag{2.44b}$$
$$J_s^T P_s + P_s J_s < 0. \tag{2.44c}$$

We next show that equation (2.44b) implies $w = 0$. If $w \neq 0$ then J_0 would have one or more zeros eigenvalues and/or one or more purely imaginary eigenvalues. Suppose, by contradiction, that $w \neq 0$ and consider the case that at least one eigenvalue of J_0 is equal to zero. Then it is always possible choose the basis T such that, from (2.44b):

$$\begin{aligned} J_0^T P_0 + P_0 J_0 &= \left[\begin{array}{c|ccc} 0 & 0 & \cdots & 0 \\ \hline \mathbf{v}^T & & \star & \end{array}\right] P_0 + P_0 \left[\begin{array}{c|c} 0 & \mathbf{v} \\ \hline 0 & \\ \vdots & \star \\ 0 & \end{array}\right] \\ &= \left[\begin{array}{c|ccc} 0 & 0 & \cdots & 0 \\ \hline \star & & \star & \end{array}\right] + \left[\begin{array}{c|c} 0 & \star \\ \hline 0 & \\ \vdots & \star \\ 0 & \end{array}\right] = \left[\begin{array}{c|c} 0 & \star \\ \hline \star & \star \end{array}\right] \end{aligned} \tag{2.45}$$

with $\mathbf{v} = \begin{bmatrix} v & 0 & \cdots & 0 \end{bmatrix}$ and $v = 0$ or $v = 1$. But, in this case the resulting matrix (2.45) is surely not positive definite and so we contradict (2.44b). Consider instead the case when J_0 has only purely imaginary eigenvalues, then in this case the J_0 represents the *Real Jordan Form* associated at the eigenvalues with zero real part, and the Jordan block associated to a pair of conjugate eigenvalues $\lambda = \pm i\omega$, with $\omega \in \mathbb{R}$, has the form

$$\begin{bmatrix} 0 & \omega \\ -\omega & 0 \end{bmatrix}$$

Then from (2.44b)

$$
\begin{aligned}
J_0^T P_0 + P_0 J_0 = &\left[\begin{array}{cc|ccc} 0 & -\omega & 0 & \cdots & 0 \\ \omega & 0 & 0 & \cdots & 0 \\ \hline & \mathbf{V}^T & & & \\ 0 & 0 & & \star & \\ \vdots & \vdots & & & \\ 0 & 0 & & & \end{array}\right] \underbrace{\left[\begin{array}{cc|c} a & b & \star \\ b & c & \\ \hline \star & \star & \end{array}\right]}_{P_0} + \\
&+ \underbrace{\left[\begin{array}{cc|c} a & b & \star \\ b & c & \\ \hline \star & \star & \end{array}\right]}_{P_0} \left[\begin{array}{cc|ccc} 0 & \omega & & 0 & \cdots & 0 \\ -\omega & 0 & \mathbf{V} & 0 & \cdots & 0 \\ \hline 0 & 0 & & & & \\ \vdots & \vdots & & \star & & \\ 0 & 0 & & & & \end{array}\right] \\
= &\left[\begin{array}{cc|c} -\omega b & -\omega c & \star \\ \omega a & \omega b & \\ \hline \star & & \star \end{array}\right] + \left[\begin{array}{cc|c} -\omega b & \omega a & \star \\ -\omega c & \omega b & \\ \hline \star & & \star \end{array}\right] \\
= &\left[\begin{array}{cc|c} -2\omega b & \omega\,(a-c) & \star \\ \omega\,(a-c) & 2\omega b & \\ \hline \star & & \star \end{array}\right]
\end{aligned}
\tag{2.46}
$$

with $\mathbf{V} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ or $\mathbf{V} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$. But the characteristic polynomial of the left 2-dimensional upper block is, $\forall\, s \in \mathbb{C}$:

$$
s^2 - 4\omega^2 b^2 - \omega^2 (a-c)^2 = 0 \tag{2.47}
$$

that has

$$s = \pm\sqrt{4\omega^2 b^2 + \omega^2 (a-c)^2}$$

as roots, so always a positive and a negative solution. But by (2.44b) and the Sylster's criterion we required that all the the eigenvalues of the *principal minors* (the sub-matrix obtained cancelling the last j columns and rows) are definite negative. Here is evident that is not true; this leads to a contradiction. So we can conclude that $w = 0$ and J does not have eigenvalues with zero real part. Now we prove that:

$$\begin{aligned} J_u^T P_u + P_u J_u = (-J_u)^T (-P_u) + (-P_u)(-J_u) \quad &< 0 \quad \Rightarrow \quad P_u < 0 \quad \text{(2.48a)} \\ J_s^T P_s + P_s J_s \quad &< 0 \quad \Rightarrow \quad P_s > 0. \quad \text{(2.48b)} \end{aligned}$$

We prove only relation (2.48b), the proof for (2.48a) is totally equivalent. Consider the following dynamical system

$$\dot{y}(t) = J_s y(t) \quad y(t) \in \mathbb{R}^q \tag{2.49}$$

then we know from (2.42) that J_s has only eigenvalues with negative real part and so (2.49) is an asymptotically stable system. So, setting

$$J_s^T P_s + P_s J_s = -Q \tag{2.50}$$

with Q positive definite, we know, from the Lyapunov equation [13] theory, that the equation

$$J_s^T X + X J_s = -Q \tag{2.51}$$

admit a solution $X = P_s' > 0$. We now prove that this solution is unique. Consider two solutions of (2.51) P_s' and P_s'' then

$$J_s^T (P_s' - P_s'') + (P_s' - P_s'') J_s = 0 \tag{2.52}$$

and multiplying for $e^{J_s^T t}$ to the left and for $e^{J_s t}$ to the right we obtain

$$\begin{aligned} 0 &= e^{J_s^T t} J_s^T (P_s' - P_s'') e^{J_s t} + e^{J_s^T t} (P_s' - P_s'') J_s e^{J_s t} \\ &= \frac{d}{dt}\left(e^{J_s^T t} (P_s' - P_s'') e^{J_s t}\right) \quad \forall t \in \mathbb{R} \end{aligned} \tag{2.53}$$

so the matrix $e^{J_s^T t}\left(P_s' - P_s''\right)e^{J_s t}$ is a constant matrix for the time evolution, and is equivalent to the case $t = 0$. So

$$e^{J_s^T t}\left(P_s' - P_s''\right)e^{J_s t} = P_s' - P_s'' \quad \forall\, t \in \mathbb{R}.$$

Now, due to the asymptotic stability of the system (2.49) when $t \to +\infty$ the modes of the system go to zero and so we can conclude that

$$0 = P_s' - P_s'' \Rightarrow P_s' = P_s''.$$

So we have proved that the solution of (2.51) is unique. But at the same time P_s is the solution to (2.51) by definition (2.50) and so is the only solution and is also positive definite, $P_s > 0$. For (2.48a) we consider the asymptotically stable system $-J_u$ and we find the solution of the relatives Lyapunov equation $-P_u$ must be positive definite and so $P_u > 0$.

Now we state a useful lemma [12, Lemma 2] needed to conclude the proof.

Lemma 2. *Let $P \in \mathbb{R}^{n\times n}$ symmetric and W a subspace of $\mathbb{R}^n$ with dim $(W) = k$. If*

$$w^T P w > 0 \qquad \forall\, w \in W$$

then P has at least k positive eigenvalues (counted with their multiplicity).

The previous lemma holds also in the case that $w^T P w < 0$ for all the vectors in W, in this case P has at least k negative eigenvalues. At this point we consider the subspace of $\mathbb{R}^n$ given by $W_s = \langle e_{n-q+1}, \ldots, e_n\rangle$ where every $e_i,\ i = n-q, \ldots, n$ is the i-th vector of the canonical basis. Then

$$\forall\, \mathbf{x} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ x \end{bmatrix} \text{ with } x \in \mathbb{R}^q \ \mathbf{x}^T P\, \mathbf{x} = x^T P_s x > 0 \tag{2.54}$$

So by lemma 2 P has at least q positive eigenvalues. The same holds considering $W_u = \langle e_1, \ldots, e_r\rangle$ so P has at least r negative eigenvalues. In other words, denoting with p

the number of negative eigenvalues of P, P is such that

$$\begin{aligned} r &\leq p \\ q &\leq n - p, \end{aligned} \tag{2.55}$$

but because $n = r + q$ the inequalities (2.55) imply that

$$q \leq n - p \Rightarrow \cancel{n} - r \leq \cancel{n} - p \Rightarrow r \geq p, \tag{2.56}$$

by (2.55) $r \leq p$ and by (2.56) $r \geq p$ if and only if $r = p$ and, from that $q = n - p$.

2. $\Rightarrow$ **3.** Consider the p' distinct eigenvalues with positive real part and the q' the distinct eigenvalues with negative real part of A. Then we consider the generalized eigenspaces N_{λ_i} relative to the eigenvalue λ_i. It is possible to prove that every generalized eigenspace is A-invariant, i.e. $v \in N_{\lambda_i} \Rightarrow Av \in N_{\lambda_i}$, and that:

$$\mathbb{R}^n = N_{\lambda_1} \oplus \cdots \oplus N_{\lambda_{p'}} \oplus N_{\lambda_{p'+1}} \oplus \cdots \oplus N_{\lambda_{p'+q'}}.$$

Now we define

$$\mathcal{H}_p := N_{\lambda_1} \oplus \cdots \oplus N_{\lambda_{p'}} \tag{2.57}$$

$$\mathcal{V}_{n-p} := N_{\lambda_{p'+1}} \oplus \cdots \oplus N_{\lambda_{p'+q'}} \tag{2.58}$$

then consider a vector $v \in \mathcal{H}_p$, by (2.57) we can express v as

$$v = \sum_{i=1}^{p'} \alpha_i v_i \qquad \text{with } \alpha_i \in \mathbb{R}^n, v_i \in N_{\lambda_i} \forall i = 1, \ldots, p'$$

But due to the invariance of N_{λ_i}

$$Av = \sum_{i=1}^{p'} \alpha_i \underbrace{Av_i}_{\in N_{\lambda_i}}$$

the vector Av is a linear combination of vectors in N_{λ_i}, with $i = 1, \ldots, p'$. So also $\mathcal{H}_p$ is A-invariant. But it is possible to prove that this implies that $\mathcal{H}_p$ is also e^{At}-invariant for all $t > 0$. The same holds for $\mathcal{V}_{n-p}$. Finally the existence of $0 < \underline{C} \leq$

$1 \leq \overline{C}$ and $\underline{\lambda} < 0 < \overline{\lambda}$ such that

$$\begin{aligned} \|x_p(t)\| &\geq \underline{C} e^{-\underline{\lambda} t} \|x_p(0)\| \\ \|x_{n-p}(t)\| &\leq \overline{C} e^{-\overline{\lambda} t} \|x_{n-p}(0)\| \end{aligned}$$

with $x_p(t) \in \mathcal{H}_p$ and $x_{n-p}(t) \in \mathcal{V}_{n-p}$, follows from an analysis of the modes of the system associated to the eigenvalues with positive or negative real part.

3. $\Rightarrow$ **2.** If there exists an invariant splitting $\mathbb{R}^n = \mathcal{H}_p \oplus \mathcal{V}_{n-p}$ then consider a change of basis T such that

$$T = \left[\begin{array}{ccc|ccc} & & & & & \\ v_1 & \cdots & v_p & v_{p+1} & \cdots & v_n \\ & & & & & \end{array}\right],$$

with $v_i \in \mathcal{H}_p$ if $i = 1, \ldots, p$ and $v_i \in \mathcal{V}_{n-p}$ if $i = p+1, \ldots, n$. Then

$$J = T^{-1} A T = \left[\begin{array}{c|c} J_{\mathcal{H}} & 0 \\ \hline 0 & J_{\mathcal{V}} \end{array}\right]$$

Now consider $x(0) \in \mathcal{V}_{n-p}$ by relations (2.38) we can conclude that exists $\overline{\lambda} < 0$ and $0 < \overline{C} \leq 1$ such that

$$\|x(t)\| = \left\|e^{At} x(0)\right\| \leq \overline{C} e^{-\overline{\lambda} t} \|x(0)\| \xrightarrow[t \to +\infty]{} 0 \tag{2.59}$$

but because $e^{At} = T e^{Jt} T^{-1}$ and the fact that T and e^{Jt} are always invertible this implies that for $x(0) \in \mathcal{V}_{n-p}$

$$\left\| T e^{Jt} \underbrace{T^{-1} x(0)}_{\in \mathcal{V}_{n-p}} \right\| \xrightarrow[t \to +\infty]{} 0 \Rightarrow e^{Jt} T^{-1} x(0) \xrightarrow[t \to +\infty]{} 0 \tag{2.60}$$

But because $T^{-1}x\left(0\right) \in \mathcal{V}_{n-p}$ we can rewrite

$$T^{-1}x\left(0\right) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ v \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ v \end{bmatrix} \in \mathbb{R}^n \tag{2.61}$$

with v a $n-p$-column vector. So (2.60) implies

$$\left[\begin{array}{c|c} e^{J_{\mathcal{H}}t} & 0 \\ \hline 0 & e^{J_{\mathcal{V}}t} \end{array}\right] \begin{bmatrix} \mathbf{0} \\ v \end{bmatrix} = e^{J_{\mathcal{V}}t}v \xrightarrow[t\to+\infty]{} 0 \quad \forall\, v \in \mathcal{V}_{n-p} \tag{2.62}$$

But this guarantees that $J_{\mathcal{V}}$ has $n-p$ eigenvalues with real part strictly negative. With similar arguments it is possible prove that $J_{\mathcal{H}}$ is associated to p eigenvalues with positive real part.

□

By virtue of Proposition 2 we can say that p-dominance ensures a splitting between $n-p$ *transient* modes and p *dominant* modes. Only the p *dominant* modes dictate the asymptotic behaviour.

The (strict) inequality (2.37) can also be seen in terms of quadratic form $V\left(x\left(t\right)\right) = x\left(t\right)^T Px\left(t\right)$:

$$\begin{aligned} \dot{V}\left(x\left(t\right)\right) &= x\left(t\right)^T \left(A^T P + PA\right) x\left(t\right) \\ &\leq -\epsilon \left\|x\left(t\right)\right\|^2 . \end{aligned}$$

For $\epsilon > 0$ this implies that the two cones

$$\begin{aligned} \mathcal{K}^- &= \left\{x \in \mathbb{R}^n | V\left(x\right) \leq 0\right\}, \\ \mathcal{K}^+ &= \left\{x \in \mathbb{R}^n | V\left(x\right) \geq 0\right\} \end{aligned}$$

are strictly contracting either in backward or in forward time:

$$\begin{aligned} \forall\, t > 0 : & e^{-At}\mathcal{K}^+ \subset \mathcal{K}^+ \\ \forall\, t > 0 : & e^{At}\mathcal{K}^- \subset \mathcal{K}^- \end{aligned}$$

Indeed, consider $x(0) \in \mathcal{K}^-$. Then $V(x(0)) \leq 0$ and so

$$\begin{aligned} V(x(t)) - V(x(0)) &= \int_0^t \dot{V}(x(\tau))\, d\tau \\ &\leq -\epsilon \int_0^t \|x(\tau)\|^2 \, d\tau < 0. \end{aligned}$$

But this means $V(x(t)) < V(x(0)) \leq 0$ and so $x(t) \in \mathcal{K}^-$. Instead consider now $x(0) \in \mathcal{K}^+$. Then $V(x(0)) \geq 0$. But as before

$$\begin{aligned} V(x(-t)) - V(x(0)) &= \int_0^{-t} \dot{V}(x(\tau))\, d\tau \\ &= -\int_{-t}^0 \dot{V}(x(\tau))\, d\tau \\ &\geq \epsilon \int_{-t}^0 \|x(\tau)\|^2 \, d\tau > 0 \end{aligned}$$

Then $V(x(-t)) > V(x(0)) \geq 0$ that implies $x(t) \in \mathcal{K}^+$.

Now, since $\mathcal{V}_{n-p} \subset \mathcal{K}^+$, where $\mathcal{V}_{n-p}\langle v_{p+1}, \ldots, v_n\rangle$ is the subspace in space decomposition of proposition 2, if we take a $x(0) \in \mathcal{V}_{n-p}$ the quadratic form $V(x(t))$ is a Lyapunov function along this trajectory. In a certain sense the p-dominance guarantees the fact that there exists an "almost" Lyapunov function, that decreases along the trajectories of the system, but that is not positive definite. Essentially the subspace linked to the negative eigenvalues of P represents the subspace where $V(x(t))$ is not positive definite and so where the system can exhibit a non stable behaviour, instead the subspace linked to the positive eigenvalues of P represents the subspace for which $V(x(t))$ is positive definite, where the second Lyapunov's method can be applied and where the system has a stable behaviour. For $p = 0$, p-dominance is the classical property of exponential stability: all modes are *transient* and the asymptotic behaviour is 0-dimensional.

2.3.2 Differential analysis of dominant non-linear systems

As done in [7], we proceed introducing the *p*-dominance for the *non-linear* case. To do that we will use a *differential* approach. A *differential* method is based on the linearisation for the system dynamic along the trajectories. From that we can infer local properties for the *non-linear* system. For example there are techniques largely used in the stability analysis. When we apply the Lyapunov's first method to prove local stability properties of fixed points, we exactly use this approach.

Example 4: Consider a *non-linear* dynamical system

$$\dot{x}(t) = f(x(t))$$

and suppose that $x = 0$ is an equilibrium point. Define $J = \frac{\partial f}{\partial x}(0)$ as the Jacobian of the system evaluated at $x = 0$. From the Lyapunov's first method we deduce that any trajectory $v(t)$ of the linearised system

$$\dot{v}(t) = Jv(t)$$

at the fixed point $x = 0$ is an approximation of the infinitesimal displacement between a trajectory $\hat{x}(t)$ arising from an infinitesimal initial variation given by $\hat{x}(0) = v(0)$ with a trajectory starting at the equilibrium $x = 0$. Indeed, if we prove exponential stability of the linearisation, we can deduce that $\hat{x}(t)$ asymptotically converges to the fixed point $x = 0$.

The aim of the *differential* analysis in the context of this thesis is to exploit the properties of the linearised dynamics beyond the local stability analysis of attractors. So consider a *non-linear* dynamical system described by

$$\dot{x}(t) = f(x(t)), \tag{2.63}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable (f is continuous with its derivative also continuous over $\mathbb{R}^n$). For $x_0 \in \mathbb{R}^n$, let $\phi(t, x_0)$ be the solution to the system (2.63) at time t from initial condition x_0 at time 0.

In the following we will be interested to study the following problem. Given two initial conditions x_0 and y_0 of the system (2.63), take the segment connecting x_0 to y_0, described by

$$x_0 + \epsilon(y_0 - x_0), \quad \text{for } \epsilon \in [0, 1].$$

Take the state evolution subject to (2.63) at every point in the segment:

$$x(t,\epsilon) = \phi(t, x_0 + \epsilon(y_0 - x_0)). \tag{2.64}$$

The study of how the length of the segment changes with respect to the time, gives information about the behaviour of system and if the trajectories tend to *contract* each other with respect to the time. This consists in studying the norm of the quantity

$$\tilde{x}(t,\epsilon) \coloneqq x(t,\epsilon) - x(t,0) \tag{2.65}$$

If we look at $x(t,\epsilon)$ as function of ϵ, and we take the Taylor expansion around $\epsilon = 0$, then

$$x(t,\epsilon) \approx x(t,0) + \underbrace{\left.\frac{\partial x(t,\epsilon)}{\partial \epsilon}\right|_{\epsilon=0}}_{\coloneqq v(t)} \epsilon.$$

Then

$$\begin{aligned} \tilde{x}(t,\epsilon) &\approx v(t)\,\epsilon \\ v(t) &= \left.\frac{\partial x(t,\epsilon)}{\partial \epsilon}\right|_{\epsilon=0} \end{aligned}$$

where $v(t)$ represents the so-called *variational vector* and represents a infinitesimal displacement, at the same time, between two trajectories starting from two different initial conditions. With reference to the first Lyapunov's method,recalling example 4, the *variational vector* represents the displacement between two neighbouring trajectories in which one is the trivial trajectory starting at the equilibrium point. It is immediate to see that important features on the evolution of the system (2.63) can be derived studying the time evolution of

the *variational vector* $v\left(t\right)$. Indeed:

$$\begin{aligned}\dot{v}(t) &= \frac{\partial}{\partial t}\frac{\partial x(t,\epsilon)}{\partial \epsilon}\bigg|_{\epsilon=0} = \frac{\partial}{\partial \epsilon}\frac{\partial x(t,\epsilon)}{\partial t}\bigg|_{\epsilon=0} = \\ &= \frac{\partial f\left(x\left(t,\epsilon\right)\right)}{\partial \epsilon}\bigg|_{\epsilon=0} = \frac{\partial f(x\left(t,\epsilon\right))}{\partial x}\bigg|_{\epsilon=0}\underbrace{\frac{\partial x\left(t,\epsilon\right)}{\partial \epsilon}\bigg|_{\epsilon=0}}_{v(t)}. \\ &= \underbrace{\frac{\partial f(x\left(t,0\right))}{\partial x}}_{:=A(t)} v(t) = A(t)v(t)\end{aligned} \tag{2.66}$$

The characterization of the *p*-dominance of the *non-linear* system (2.63) is exactly given through the linear time varying system (2.66). The previous definition 7 of the *p*-dominance is given only for LTI systems. So the extension to this case passed through considering the state evolution of the linear time varying system (2.66). These reasoning brings to the following definition of *p*-dominance:

Definition 8. Given a non-linear system $\dot{x}\left(t\right) = f\left(x\left(t\right)\right)$ with $x\left(t\right) \in \mathbb{R}^n$ and denoting by $A\left(x\right) = \frac{\partial f}{\partial x}\left(x\right)$ the Jacobian of f evaluated at x. Then the dynamical system is *p*-dominant if there exists a symmetric matrix $P \in \mathbb{R}^{n\times n}$ with *inertia p* such that

$$A\left(x\right)^T P + PA\left(x\right) \leq -\epsilon I \quad \forall x \in \mathbb{R}^n \tag{2.67}$$

for some $\epsilon \geq 0$. The property is strict if $\epsilon > 0$.

Now before moving on it is necessary to introduce some others concepts. In particular it is necessary to introduce concepts of *differential geometry* as *tangent space* and *distribution*.

Definition 9 (Tangent space)**.** Given a C^1 curve $\gamma\left(t\right) : \left[t_0 - T, t_0 + T\right] \to \mathbb{R}^n$ with $t_0, T \in \mathbb{R}$, for which $\gamma\left(t_0\right) = x$ a *tangent vector* calculated at t_0 is :

$$v_x\left(\gamma\right) = \frac{d\gamma\left(t_0\right)}{dt} \in \mathbb{R}^n. \tag{2.68}$$

The *tangent space* of $\mathbb{R}^n$ at p is the set of *tangent vectors* at x, and it is denoted by $T_x\mathbb{R}^n$.

For the scope of this thesis the *tangent space* is equal to $\mathbb{R}^n$ at every point $x \in \mathbb{R}^n$.

Definition 10 (Distribution)**.** A *distribution*, in *differential geometry*, is a map that assigns a subspace of the tangent space to each point $x \in \mathbb{R}^n$:

$$\mathcal{H} \,:\, x \in \mathbb{R}^n \,\mapsto\, \mathcal{H}_x \subseteq T_x\mathbb{R}^n,$$

with $\mathcal{H}_x$ a subspace.

With the next definition the formal concept of *flow*, for an autonomous systems of ordinary differential equations, is introduced.

Definition 11 (Flow)**.** Be $f : \mathbb{R}^n \to \mathbb{R}^n$ Lipschitz-continuous function and $x : \mathbb{R} \to \mathbb{R}^n$ the solution to the initial value problem

$$\begin{cases} \dot{x}\left(t\right) & = f\left(x\left(t\right)\right) \\ x\left(0\right) & = x_0 \end{cases}.$$

Then $\phi\left(t, x_0\right) = x\left(t\right)$ is the *flow* for f.

To derive the following results it is needed to introduce a last fundamental concept: the *prolonged system*

$$\begin{cases} \dot{x}\left(t\right) & = f\left(x\left(t\right)\right) \\ \dot{v}\left(t\right) & = A\left(x\left(t\right)\right) v\left(t\right) \end{cases}, \tag{2.69}$$

where $A\left(x\left(t\right)\right) = \frac{\partial f}{\partial x}\left(x\left(t\right)\right)$ and $v\left(t\right)$ represent a vector in the *tangent space* $T_x\mathbb{R}^n$. With all these concepts it is now possible to derive the following theorem that represents the generalization to the *non-linear* case of the previous proposition 2.

Theorem 2. *Let* (2.63) *a p-dominant system with matrix P of inertia p. Then, for each $x \in \mathbb{R}^n$ there exists an invariant splitting $T_x\mathbb{R}^n = \mathcal{H}_x \oplus \mathcal{V}_x$ such that $\forall\, h \in \mathcal{H}_x$ and $\forall\, w \in \mathcal{V}_x$*

$$\begin{aligned} \left.\frac{\partial \phi\left(t, x_0\right)}{\partial x_0}\right|_{x_0 = x} h \in \mathcal{H}_{\phi(t,x)} \quad \forall\, t \in \mathbb{R} \\ \left.\frac{\partial \phi\left(t, x_0\right)}{\partial x_0}\right|_{x_0 = x} w \in \mathcal{V}_{\phi(t,x)} \quad \forall\, t \in \mathbb{R} \end{aligned} \tag{2.70}$$

where $\mathcal{H}$ and $\mathcal{V}$ are two distributions associated to the subspaces $\mathcal{H}_x$ and $\mathcal{V}_x$ with $\mathcal{H}_x$ of dimension p and $\mathcal{V}_x$ of dimension $n-p$. Furthermore, there exist constants $0 < \underline{C} \leq 1 \leq \overline{C}$

and $\underline{\lambda} < 0 < \overline{\lambda}$ *such that,*

$$\left\| \left. \frac{\partial \phi(t, x_0)}{\partial x_0} \right|_{x_0 = x} h \right\| \geq \underline{C} e^{-\underline{\lambda} t} \|h\| \qquad \forall x \in \mathbb{R}^n,\ h \in \mathcal{H}_x \tag{2.71a}$$

$$\left\| \left. \frac{\partial \phi(t, x_0)}{\partial x_0} \right|_{x_0 = x} w \right\| \leq \overline{C} e^{-\overline{\lambda} t} \|w\| \qquad \forall x \in \mathbb{R}^n,\ w \in \mathcal{V}_x \tag{2.71b}$$

Proof.

Invariant splitting : First consider the following quadratic function for the *variational system* of (2.69).

$$V(v(t)) \coloneqq v(t)^T P v(t).$$

Then by p-dominance

$$\dot{V}(v(t)) \leq -\epsilon \|v(t)\|^2 \tag{2.72}$$

As a consequence the two cone fields

$$\mathcal{K}^+(x) \coloneqq \{v \in T_x \mathbb{R}^n | V(v) \geq 0\}$$
$$\mathcal{K}^-(x) \coloneqq \{v \in T_x \mathbb{R}^n | V(v) \leq 0\}$$

are strictly contracting either in forward time or in backward, respectively. A *cone field* on $\mathbb{R}^n$ is a collection of cones $\mathcal{K}(x) \subset T_x \mathbb{R}^n$ for $x \in \mathbb{R}^n$. For any $p > 0$ and for any $v \in \mathcal{K}^-(x)$, the dissipation inequality (2.72) is true for all $x \in \mathbb{R}^n$ and all v on the boundary of $\mathcal{K}^-(x)$. Due to the fact that $\frac{\partial \phi(t, x_0)}{\partial x_0} v(0)$ represents the flow associated to the *variational vector* $v(t)$, so the infinitesimal displacement between two trajectories one starting from x the other at an infinitesimal different initial condition from x, in the tangent space $T_x \mathbb{R}^n$, we argue that

$$\begin{aligned} \frac{d}{dt} v(t) &= \frac{\partial}{\partial t} \frac{\partial \phi(t, x_0)}{\partial x_0} v(0) = \frac{\partial}{\partial x_0} \frac{\partial \phi(t, x_0)}{\partial t} v(0) = \frac{\partial}{\partial x_0} f(\phi(t, x_0)) v(0) \\ &= \frac{\partial f(x(t))}{\partial x} \underbrace{\frac{\partial \phi(t, x_0)}{\partial x_0} v(0)}_{v(t)} = A(x(t)) v(t), \end{aligned}$$

holds and

$$\forall\, t \geq 0 : \left.\frac{\partial \phi\left(t, x_0\right)}{\partial x_0}\right|_{x_0 = x} \mathcal{K}^{-}\left(x\right) \subseteq \mathcal{K}^{-}\left(x\right) \tag{2.73a}$$

$$\forall\, t > 0 : \left.\frac{\partial \phi\left(t, x_0\right)}{\partial x_0}\right|_{x_0 = x} \left(\mathcal{K}^{-}\left(x\right) \setminus \{0\}\right) \subset \mathcal{K}^{-}\left(x\right). \tag{2.73b}$$

Where it is used the notations that, given a set $\mathcal{S} \subseteq T_x \mathbb{R}^n$, then

$$\left.\frac{\partial \phi\left(t, x_0\right)}{\partial x_0}\right|_{x_0 = x} \mathcal{S} := \left\{ \left.\frac{\partial \phi\left(t, x_0\right)}{\partial x_0}\right|_{x_0 = x} v\left(0\right) | v\left(0\right) \in \mathcal{S} \right\}.$$

Defining $\lambda_{min}\left(P\right)$ the smallest eigenvalue of P, the dissipation inequality (2.72) also implies

$$\dot{V}\left(v\left(t\right)\right) \leq \epsilon_1 V\left(v\left(t\right)\right) \tag{2.74}$$

where $\epsilon_1 := \frac{\epsilon}{|\lambda_{min}(P)|} > 0$. This follows directly through the fact that

$$P + I\left|\lambda_{min}\right| \geq 0$$

So, if $v\left(0\right)$ belongs to the interior of $\mathcal{K}^{-}\left(x\right)$, by time-integration of inequality (2.74)

$$\frac{\dot{V}\left(v\left(t\right)\right)}{V\left(v\left(t\right)\right)} \geq \epsilon_1 \Rightarrow \int_0^t \frac{\dot{V}\left(v\left(t\right)\right)}{V\left(v\left(t\right)\right)} d\tau \geq \epsilon_1 \int_0^t d\tau \Rightarrow$$

$$\ln V\left(v\left(t\right)\right) - \ln V\left(v\left(0\right)\right) \geq \epsilon_1 t \Rightarrow \ln \frac{V\left(\left.\frac{\partial \phi(t,x_0)}{\partial x_0}\right|_{x_0=x} v\left(0\right)\right)}{V\left(v\left(0\right)\right)} \geq \epsilon_1 t \Rightarrow$$

$$\frac{V\left(\left.\frac{\partial \phi(t,x_0)}{\partial x_0}\right|_{x_0=x} v\left(0\right)\right)}{V\left(v\left(0\right)\right)} \geq e^{\epsilon_1 t} \text{ that holds } \forall t \geq 0.$$

The last estimate with equations (2.73) guarantees that there exist $T,\ \nu > 0$ such that

$$\frac{\left\| \left.\frac{\partial \phi(t,x_0)}{\partial x_0}\right|_{x_0=x} v\left(0\right) \right\|}{\left\| v\left(0\right) \right\|} \geq \nu \quad \forall\, t \geq T,\ x \in \mathbb{R}^n \text{ and } v\left(0\right) \in \mathcal{K}^{-}\left(x\right) \tag{2.75}$$

Likewise, for any $n-p>0$ and for any $v\in\mathcal{K}^{+}(x)$, proceeding as before, this time setting a new $\epsilon_2=\frac{\epsilon}{\lambda_{max}(P)}>0$ and integrating backward in time it is possible prove that exist $T,\ \nu>0$ such that

$$\frac{\left\|\left.\frac{\partial\phi(-t,x_0)}{\partial x_0}\right|_{x_0=x} v(0)\right\|}{\|v(0)\|}\geq\nu\quad\forall\, t\geq T,\ x\in\mathbb{R}^n \text{ and } v(0)\in\mathcal{K}^{+}(x)\,. \tag{2.76}$$

The fundamental steps, given the two conditions (3.16) and (3.17), is to proceed as in the proof [14, Theorem 1.2] to show that

$$\mathcal{H}_x:=\bigcap_{t\geq 0}\left.\frac{\partial\phi(t,x_0)}{\partial x_0}\right|_{x_0=\phi(-t,x)}\mathcal{K}^{-}(\phi(-t,x))\qquad\subset\mathcal{K}^{-}(x) \tag{2.77a}$$

$$\mathcal{V}_x:=\bigcap_{t\geq 0}\left.\frac{\partial\phi(-t,x_0)}{\partial x_0}\right|_{x_0=\phi(t,x)}\mathcal{K}^{+}(\phi(t,x))\qquad\subset\mathcal{K}^{+}(x) \tag{2.77b}$$

are invariant distributions of dimension p and $n-p$, respectively, that is,

$$\left.\frac{\partial\phi(t,x_0)}{\partial x_0}\right|_{x_0=x}\mathcal{H}_x\subseteq\mathcal{H}_{\phi(t,x)}\quad\forall\, t\in\mathbb{R}$$
$$\left.\frac{\partial\phi(t,x_0)}{\partial x_0}\right|_{x_0=x}\mathcal{V}_x\subseteq\mathcal{V}_{\phi(t,x)}\quad\forall\, t\in\mathbb{R}.$$

Exponential estimates: Due to (2.77a), for all $x\in\mathbb{R}^n$, $v(0)\in\mathcal{H}_x$ we have that $v(t)$ belongs to the interior of $\mathcal{K}^{-}(x)$. The estimate (2.71a) with $\underline{\lambda}=\frac{\epsilon_1}{2}$ follows from the fact that $-V(v(t))$ is positive definite in $\mathcal{H}_x$ and that $V(v(t))$ satisfies (2.72). Likewise, $v(0)\in\mathcal{V}_x$ implies that $v(t)$ belongs to the interior of $\mathcal{K}^{+}(x)$. The estimate (2.71b) with $\overline{\lambda}=\frac{\epsilon_2}{2}$ follows from the fact that $V(v(t))$ is positive definite in $\mathcal{V}_x$ and satisfies (2.72).

□

The interpretation of the theorem is very similar that in the linear case, substantially the linearised flow $\left.\frac{\partial\phi(t,x_0)}{\partial x_0}\right|_{x_0=x} v(0)$ in the *tangent space* $T_x\mathbb{R}^n$ admits an invariant splitting between $n-p$ transient modes and p dominant modes. With the p dominant modes that dictate the asymptotic behaviour of the system. For the next theorem it will be useful to in-

troduce the following *incremental* inequality that represent the integration of the inequality (2.72). Let $w(t)$ and $y(t)$ solutions of (2.63). Then

$$\begin{aligned}
\dot{V}(w(t)-y(t)) = & (w(t)-y(t))^T P\left(f(w(t))-f(y(t))\right)+ \\
& +\left(f(w(t))-f(y(t))\right)^T P(w(t)-y(t)) \\
= & (w(t)-y(t))^T P\int_0^1 \frac{\partial f}{\partial x}\left(sw(t)+(1-s)y(t)\right)ds(w(t)-y(t))+ \\
& +(w(t)-y(t))^T\int_0^1 \frac{\partial f}{\partial x}^T\left(sw(t)+(1-s)y(t)\right)dsP(w(t)-y(t)) \\
= & (w-y)^T\int_0^1 P\frac{\partial f}{\partial x}\left(sw(t)+(1-s)y(t)\right)+ \\
& +\frac{\partial f}{\partial x}^T\left(sw(t)+(1-s)y(t)\right)Pds(w(t)-y(t)) \\
\leq & (w(t)-y(t))^T\left(-\int_0^1 \epsilon I\,ds\right)(w(t)-y(t)) \\
= & -\epsilon\left\|w(t)-y(t)\right\|^2.
\end{aligned} \tag{2.78}$$

The inequality (2.78) represents the *finite* difference of two trajectories $w(t)$ and $y(t)$ starting from different initial conditions. In some sense it is the equivalent of the quantity $v(t)$ that in (2.66) is defined *differentially*, namely for an *infinitesimal* displacement of initial conditions. Some others preliminaries are now introduced to explain the following results.

Definition 12 (homeomorphism)**.** A function $Q : X \to Y$ between two topological spaces is a *homeomorphism* if it has the following properties:

- Q is one-to-one;
- Q is continuous;
- the inverse function Q^{-1} is continuous.

The previous definition is used for introducing the concept of *topologically equivalence*.

Definition 13. Let ϕ, ψ be flows in respective spaces A, B. Then ϕ and ψ are topologically equivalent if there is a homeomorphism $Q : A \to B$ such that

$$Q \circ \phi(t, \cdot) = \psi(t, \cdot) \circ Q \quad \forall\, t \in \mathbb{R}. \tag{2.79}$$

Topological equivalence is an equivalence relation on the class of flow; it formalizes the notion of "having the same qualitative dynamics". Finally we introduce the last important definition

Definition 14. A state p is a *ω-limit* point of $x\left(t\right)$ if there exist a sequence of times $\left\{t_n\right\}_{n\in\mathbb{N}}$ with $lim_{n\to+\infty}t_n=\infty$, such that:

$$lim_{n\to+\infty}x\left(t_n\right)=p$$

The set of all such p for fixed initial condition $x\left(0\right)=x$ is denoted by $\Omega\left(x\right)$ and is called the *ω-limit* set.

Theorem 3. *Let* (2.63) *a p-dominant system with matrix P of inertia p. Then the flow on any compact ω-limit set is topologically equivalent to a flow on a compact invariant set of a Lipschitz system in $\mathbb{R}^p$.*

For small values of p, theorem 3 severely constrains the possible attractors of the system.

Corollary 1. *Under assumptions of theorem 3, every bounded solution asymptotically converges to the following:*

1. *a unique fixed point if $p=0$;*
2. *a fixed point if $p=1$;*
3. *a simple attractor if $p=2$, that is, a fixed point, a set of fixed points and connecting arcs, or a limit cycle.*

A strict 0-dominant system is a contractive system. For the linear systems, the property is simply exponential stability, meaning hyperbolicity and contraction of the n transient modes to the 0-dimensional attractor. More generally 0-dominance ensures the existence of the *incremental* Lyapunov function $V\left(x\left(t\right)-y\left(t\right)\right)$. Then from (2.78), it implies exponential contraction of the difference between any two trajectories. The attractor of a 0-dominant system is necessarily a unique fixed point.

The property of 2-dominance provides the following generalization of the Poincaré-Bendixson theorem [15]:

Corollary 2. *For $p=2$, under assumptions of theorem 3, let $\mathcal{U}\subseteq\mathcal{R}^n$ be a compact forward invariant set that does not contain fixed points. Then, the ω-limit set of any point in $\mathcal{U}$ is a closed orbit.*

In this section the *Koopman operator* is presented. This instrument, that was developed by Koopman in [8], can be very useful to understand the behaviour of complex dynamical systems. The power of this operator is the fact that it can translate a non-linear, finite dimensional problem into a linear but infinite dimensional one. To deal with the infinite dimensional perspective several approximations are presented, for example in [16], [17] or in [18].

The *Koopman operator* can be derived both for the continuous and the discrete time systems. Even though the neural models presented in this thesis are all given in the continuous time, these are translated in their discrete version. This allows a simpler framework for the application of the *Koopman operator*, following the derivation in [19].

Consider a dynamical system

$$\dot{x}(t) = h(x(t)) \tag{2.80}$$

defined on $\mathbb{R}^n$ and with h a possibly *non-linear* but sufficiently regular function. We denote by $\phi(t, x_0)$ the state at time t of a trajectory solution for (2.80) that starts at time 0 at point x_0. Following the notation of the thesis, the family of functions $\phi(t, \cdot)$ is called *flow*.

We denote by g an arbitrary function, called an *observable* from $\mathbb{R}^n$ to $\mathbb{C}$, that belongs to a infinite-dimensional Hilbert space as $L^2(\mathbb{R}^n, \nu)$ with ν a proper measure. The value of this *observable* g given x_0 and t is

$$g(t, x_0) = g(\phi(t, x_0)).$$

Then the family of operators U^t, acting on the space of *observables*, parametrized by time t is defined by

$$U^t g(x_0) \coloneqq g(\phi(t, x_0))$$

Thus, for fixed time t, U^t maps the *observable* $g(x_0)$ to $g(t, x_0)$. The family of operators U^t indexed by time t is called the *Koopman operator* of the continous-time system (2.80). Sampling $\phi(t, x_0)$ for times $\tau, 2\tau, \ldots, n\tau, \ldots$ leads to the τ-mapping $T \coloneqq \phi(\tau, \cdot) : \mathbb{R}^n \to \mathbb{R}^n$ with the discrete dynamics

$$x(k+1) = Tx(k) \tag{2.81}$$

This preliminary is necessary to introduce the formal definition of the *Koopman operator* for

the discrete-time case.

We typically describe a discrete time system through its orbits $\left\{T^k(x)\right\}_{k=0}^{\infty}$, where x represents the initial condition. The idea of the *Koopman operator* is to focus on the evolution of an *output* function $g : \mathbb{R}^n \to \mathbb{C}$, the *observable*, and on its orbits $\left\{g(T^k(x))\right\}_{k=0}^{\infty}$. Formally the evolution of all *observables* is given by the *Koopman operator* associated with the system (2.81), defined next.

Definition 15 (*Koopman operator* (discrete-time))**.** Consider the function space $L^2(\nu)$ of *observables* $g : \mathbb{R}^n \to \mathbb{C}$ with ν a proper measure. The *Koopman operator* $U_T : L^2(\nu) \to L^2(\nu)$ associated with the map $T : \mathbb{R}^n \to \mathbb{R}^n$ is defined through the composition

$$U_T f = f \circ T \quad \forall\, f \in L^2(\nu)$$

In the following the subscript is omitted and we denote the operator by U.

The *Koopman operator* can be very powerful for two main advantages. First, it gives a *global* picture of the system, in contrast to the point-wise description in terms of orbits. Second, it provides a linear approach to the study of the (*non-linear*) system since it is a linear operator:

$$\begin{aligned} \forall\, c_1, c_2 \in \mathbb{C} \quad & \forall\, f_1, f_2 \in L^2(\nu) \\ U(c_1 f_1 + c_2 f_2) &= (c_1 f_1 + c_2 f_2) \circ T \\ &= c_1 f_1 \circ T + c_2 f_2 \circ T \\ &= c_1 U f_1 + c_2 U f_2 \end{aligned}$$

However the *Koopman operator* is infinite-dimensional. This is the main challenge to overcome when dealing with the operator-theoretic viewpoint. The focus in [20], [21], and much of the follow-up work, will be on solving this issue by focusing on the spectral objects-eigenvalues, eigenfunctions, and modes, to be defined below and reducing the dimensionality via finite-dimensional projections onto eigenspaces.

2.4.1 Koopman eigenvalues and eigenfunctions

As a linear operator a eigenvalues, eigenfunctions decomposition can be very useful to understand the behaviour of the operator. Indeed as in the matrix case, these objects resume some important properties of the original dynamic but their analysis is simpler. The eigenfunctions and eigenvalues of the *Koopman operator* are called *Koopman eigenfunctions* and *Koopman eigenvalues* for short. And for discrete-time systems, they are defined as follows.

Definition 16 (*Koopman eigenfunction* and *eigenvalue* (discrete-time))**.** An eigenfunction of the *Koopman operator* associated with the discrete-time map T is an *observable* $\psi_\mu \in L^2(\nu) \setminus \{0\}$ that satisfies

$$U\psi_\mu = \psi_\mu \circ T = \mu\psi_\mu$$

where $\mu \in \mathbb{C}$ is the corresponding eigenvalue.

Example 5 (Linear discrete-time system)*:* Consider the linear transformation $T(x) = Ax, x \in \mathbb{R}^n$, where A is a matrix with eigenvalues μ_j and corresponding left eigenvectors w_j. In this case the spectrum of the *Koopman operator*, so the set of the eigenvalues, contains the eigenvalues μ_j and the associated eigenfunctions are given by $\psi_{\mu_j}(x) = w_j^T x$. Indeed

$$U\psi_{\mu_j}(x) = \psi_{\mu_j}(Ax) = w_j^T Ax = \mu_j w_j^T x = \mu_j \psi_{\mu_j}(x)$$

From the nature of T these eigenfunctions inherit the linearity. This is a general fact, usually the eigenfunctions of the *Koopman operator* resume important characteristic and properties of the underlying system.

Because we deal with an infinite dimensional space, finding the *Koopman eigenfunctions* is typically a difficult problem and can not be always solved analytically. But there exist different methods that try to derive them with different techniques, some asymptotically exact some approximated. The first method presented is the generalized Laplace averages.

Generalized Laplace Averages The first exact method is based on obtaining Koopman eigenfunctions through the so-called *Generalized Laplace Averages* (GLA). The power of this method is that the result is exact, but nevertheless has some weaknesses. We present the method for a discrete-time system. As a first assumption we have to suppose that $\mu_1, \dots \mu_j$ are simple Koopman eigenvalues (so with algebraic multiplicity equal to one) with $1 \geq |\mu_1| \geq \cdots \geq |\mu_j|$ and there is no other points μ in the spectrum of U such that $|\mu| \geq |\mu_j|$.

Just here we encountered the first problem of this methodology, the eigenvalues must be known a priori, and the method finds only the eigenfunctions associated to given eigenvalues. Under the previous assumptions, for a bounded, continuous observable g, the GLA is given by

$$g_1^*(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mu_1^{-k} g\left(T^k(x)\right). \tag{2.82}$$

Then, iteratively for $j = 2, \ldots$ we define the GLA by

$$g_j^*(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mu_j^{-k} \left(g\left(T^k(x)\right) - \sum_{i=1}^{j-1} \mu_i^k g_i^*(x) \right) \tag{2.83}$$

The function g_j^* is the projection of g onto the space spanned by the eigenfunction ψ_{μ_j} and so it is an eigenfunction scaled by a factor $\lambda_j \in \mathbb{C}$. In this way we try to approximate the eigenfunctions. For $j = 2, \ldots$ the procedure subtracts from signal g the already obtained eigenvalues-eigenfunctions part of the signal g associated to the index less or equal to $j-1$, and repeating the "average" (2.82) on the unknown part of the signal g. We do not prove that (2.83) provide an eigenfunction but we write a sketch of the proof that (2.82) provide an eigenfunction. For that scope, we define the following operator

$$U_{\mu_1} = \mu_1^{-1} U$$

Then, for some function $g(x)$, consider

$$\begin{aligned}
U\left(\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} U_{\mu_1}^k g(x)\right) &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mu_1^{-k} U^k g\left(T(x)\right) \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mu_1^{-k} U^{k+1} g(x) \\
&= \mu_1 \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mu_1^{-(k+1)} U^{k+1} g(x) \\
&= \mu_1 \left[\lim_{n\to\infty} \frac{1}{n} \left(\mu_1^{-n} g\left(T^n(x)\right) - g(x) + \sum_{k=0}^{n-1} U_{\mu_1}^k g(x)\right)\right] \\
&= \mu_1 \left[\lim_{n\to\infty} \frac{1}{n} \left(\mu_1^{-n} g\left(T^n(x)\right) + \sum_{k=0}^{n-1} U_{\mu_1}^k g(x)\right)\right]
\end{aligned}$$

where the last equality derives from the boundedness of g. As done in [17] we can prove that the first term on the right side of the equality in the last line goes to zero. Then it is also possible to prove the convergence of this result, and so by that, we can prove that

$$Ug^*_{\mu_1}(x) = \mu_1 g^*_{\mu_1}(x).$$

But, as already mentioned, this method has some weaknesses. Principally we must know the eigenvalues a priori. Moreover the procedure can also bring to numerical problems.

2.4.2 Koopman modes

From the eigenfunctions of the *Koopman operator* we can obtain a complete basis for $L^2(\nu)$.This is the case when the dynamics are integrable and defined in a compact space. In this case, every *observable* that can be expanded with the eigenfunctions of the *Koopman operator*. More generally, this expansion leads to the notion of *Koopman modes*.

Definition 17 (Koopman mode expansion)**.** Suppose that the span of Koopman eigenfunctions $\{\psi_{\mu_j}\}_{j=1}^{\infty}$ densely fills the space $\mathscr{G} \subseteq L^2(\nu)$. The *Koopman mode expansion* of $g \in \mathscr{G}$ is given by

$$g = \sum_{j=1}^{\infty} \lambda_j \psi_{\mu_j} \tag{2.84}$$

The coefficients $\lambda_j \in \mathbb{C}$ are the *Koopman modes* related to the *observable* g.

Using the *Koopman mode expansion* and the definition of *Koopman eigenfunctions*, it is possible to write

$$U^k g = \sum_{j=1}^{\infty} \lambda_j \mu_j^k \psi_{\mu_j} \tag{2.85}$$

In this case $|\lambda_j|$ and $\angle\lambda_j$ yield the amplitude and phase, respectively, of a specific oscillation mode in the time evolution of *observable* g.

Until now, only scalar-valued *observables* are considered. However, it is often useful to deal with vector-valued *observables*. These are defined by rearranging the different *observables* in a vector $\mathbf{g}$:

$$\mathbf{g}(x) := \begin{bmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_p(x) \end{bmatrix},$$

where each terms $g_i \in L^2(\nu)$. In this case the action of the *Koopman operator* is defined component-wise as :

$$U^k \mathbf{g}(x) := \begin{bmatrix} U^k g_1(x) \\ U^k g_2(x) \\ \vdots \\ U^k g_p(x) \end{bmatrix}.$$

As before it is possible to expand each individual *observable* in terms of eigenfunctions and

so in this case:

$$\mathbf{g}(x) = \sum_{j=1}^{\infty} \lambda_j \psi_{\mu_j}$$

where now each *Koopman mode* $\lambda_j \in \mathbb{C}^p$. Given that definition, it is possible to consider the vector identity function $\mathrm{Id}(x) = x$. In this case the dependence on the initial conditions is given by the eigenfunctions and the *Koopman mode expansion* provides the orbits of the system for any initial condition. This is illustrated in the following example.

Example 6 (Linear discrete-time system (continued))*:* Consider again the linear transformation $T(x) = Ax$ with $x \in \mathbb{R}^n$ as in the example 5. An orbit of the system is given by

$$T^k(x) = \sum_{j=1}^{n} \lambda_j \left(w_j^T x\right) \mu_j^k$$

where λ_j and w_j are the right and the left eigenvectors of A, respectively. This represents the evolution of $\mathrm{Id}(x) = x$ under the action of the *Koopman operator*. Comparing it with (2.84) and using that $\psi_{\mu_j}(x) = w_j^T x$ it follows that the *Koopman modes* of the identity function are the right eigenvectors λ_j. Then the *Koopman mode expansion* is finite, owing to the linear dynamics and the specific observable Id that is also linear in the state.

The next example instead presents the application of the *Koopman operator* to the *nonlinear* case.

Example 7: Consider the system

$$x(k+1) = T(x(k)) = \begin{bmatrix} a x_1(k) \\ b x_2(k) + (b - a^2) x_1^2(k) \end{bmatrix} \tag{2.86}$$

with $x = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T$, $a, b \in [0,\, 1)$. The system has one equilibrium at the origin $x = 0$ that is also stable, indeed the Jacobian matrix $J = \frac{\partial T}{\partial x}(x)\big|_{x=0}$, is

$$J = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} \tag{2.87}$$

with eigenvalues that have absolute value less than 1. Now we try to find a spectral decomposition of the *Koopman operator* associated with system (2.86). We proceed by trial and error driven by the idea that the *Koopman operator* encodes fundamental properties of the

original system. We are searching for a scalar-function $\psi_\mu(x)$ such that

$$U\psi_\mu = \mu\psi_\mu. \tag{2.88}$$

We try by setting $\psi_\mu = c_1 x_1$ with $c_1 \in \mathbb{R}$. Then (2.88) implies

$$\begin{aligned} U\psi_\mu(x) &= \mu\psi_\mu(x) \quad \forall x \in \mathbb{R}^2 \\ \Rightarrow \psi_\mu(T(x)) &= \mu\psi_\mu(x) \\ \Rightarrow c_1(T(x))_1 &= \mu c_1 x_1 \\ \Rightarrow c_1 a x_1 &= \mu c_1 x_1 \end{aligned}$$

that holds for all $x \in \mathbb{R}^2$ if and only if $\mu = a$. So the pair $\mu = a$ and $\psi_\mu = c_1 x_1$ represents a eigenvalue-eigenfunction pair. We now try setting $\psi_\mu = c_1 x_1 + c_2 x_2$. Then

$$\begin{aligned} U\psi_\mu(x) &= \mu\psi_\mu(x) \quad \forall x \in \mathbb{R}^2 \\ \Rightarrow c_1(T(x))_1 + c_2(T(x))_2 &= \mu\psi_\mu(x) \\ \Rightarrow c_1(T(x))_1 + c_2(T(x))_2 &= \mu(c_1 x_1 + c_2 x_2) \\ \Rightarrow c_1 x_1(a-\mu) + c_2 x_2(b-\mu) + c_2\left(b-a^2\right)x_1^2 &= 0 \end{aligned}$$

but is evident that there do not exist c_1, c_2, μ, different from zero, such that the last equality holds for all x_1 and x_2. So we try with the eigenfunction candidate $\psi_\mu = c_1 x_1^2 + c_2 x_2$

$$\begin{aligned} U\psi_\mu(x) &= \mu\psi_\mu(x) && \forall x \in \mathbb{R}^2 \\ \Rightarrow c_1(T(x))_1^2 + c_2(T(x))_2 &= \mu\left(c_1 x_1^2 + c_2 x_2\right) \\ \Rightarrow c_1 a^2 x_1^2 + c_2\left(b x_2 + \left(b-a^2\right)x_1^2\right) &= \mu\left(c_1 x_1^2 + c_2 x_2\right) \\ \Rightarrow x_1^2\left(c_1 a^2 + c_2 b - c_2 a^2 - \mu c_1\right) + c_2 x_2(b-\mu) &= 0 && \text{setting } c_1 = c_2 = 1 \\ \Rightarrow x_1^2\left(\cancel{a^2} + b - \cancel{a^2} - \mu\right) + x_2(b-\mu) &= 0 \end{aligned}$$

that holds for all $x \in \mathbb{R}^2$ if $\mu = b$. So we have find another eigenvalue-eigenfunction pair $\mu = b$ and $\psi_\mu(x) = x_2 + x_1^2$. Not only, it is possible to prove, by similar computations, that also $\mu = a^2$ and x_1^2 is another eigenvalue-eigenfunction pair. Consider now the following

vector of *observables* eigenfunctions.

$$\psi(x) = \begin{bmatrix} \psi_a(x) \\ \psi_b(x) \\ \psi_{a^2}(x) \end{bmatrix} = \begin{bmatrix} x_1 \\ x_1^2 + x_2 \\ x_1^2 \end{bmatrix} \tag{2.89}$$

Then

$$U^k\psi = \begin{bmatrix} U^k\psi_a \\ U^k\psi_b \\ U^k\psi_{a^2} \end{bmatrix} = \begin{bmatrix} a^k\psi_a \\ b^k\psi_b \\ a^{2k}\psi_{a^2} \end{bmatrix} = \underbrace{\begin{bmatrix} a^k & 0 & 0 \\ 0 & b^k & 0 \\ 0 & 0 & a^{2k} \end{bmatrix}}_{:=A^k} \begin{bmatrix} \psi_a \\ \psi_b \\ \psi_{a^2} \end{bmatrix}. \tag{2.90}$$

Take now the function $\mathrm{Id}(x) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ then Id admit a Koopman mode expansion as

$$\mathrm{Id} = \lambda_1\psi_a + \lambda_2\psi_b + \lambda_3\psi_{a^2} = \underbrace{\begin{bmatrix} | & | & | \\ \lambda_1 & \lambda_2 & \lambda_3 \\ | & | & | \end{bmatrix}}_{:=\Lambda} \psi \tag{2.91}$$

with $\lambda_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $\lambda_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ and $\lambda_3 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$. Finally we can rewrite the dynamics of the system as

$$\begin{aligned} x(k) &= T^k(x(0)) = U^k\mathrm{Id}(x(0)) = \left(U^k\Lambda\psi\right)(x(0)) = \Lambda\left(U^k\psi\right)(x(0)) \\ &= \Lambda A^k\psi(x(0)) \end{aligned}$$

In this way the system is no more propagated through the *non-linear* map T, but through a *non-linear* function of the initial state and a linear evolution over the time k.

Consider an N-dimensional linear subspace $\mathscr{G}_N \subset L^2(\nu)$ spanned by an orthonormal basis of functions $\{\xi_j\}_{j=1}^N$. The fundamental idea to approximate the *Koopman operator* consists in considering every function in the space $\mathscr{G}_N$ as a finite combination of elements of the basis. So if we know the propagation of every element of the basis through the *Koopman operator* and if this propagation belongs to $\mathscr{G}_N$, then we can express the resulting elements as a finite linear combination of the elements of the basis. In this way we can express the action of *Koopman operator* on a function, represented by its "coordinates" in $\mathscr{G}_N$, as a linear transformation of exact the "coordinates" representing the function in the basis $\{\xi_j\}_{j=1}^N$. The critical step of this procedure consists in requiring that the function deriving from the *Koopman operator* action on the elements of the basis belongs to $\mathscr{G}_N$. This is not always the case, and so to fix this we can resort to the projection operator $\Pi : L^2(\nu) \to \mathscr{G}_N$ defined as follows. Let $g \in L^2(\nu)$, then

$$\Pi(g) \coloneqq \sum_{i=1}^{N} \langle g, \xi_i \rangle \xi_i. \tag{2.92}$$

The projection operator (2.92) can be decomposed as

$$\Pi = \Xi^T \Gamma \tag{2.93}$$

where the operator $\Gamma : L^2(\nu) \to \mathbb{C}^N$ can be seen as a bounded linear map that yields the coordinates of $g \in L^2(\nu)$ in the basis $\{\xi_j\}_{j=1}^N$ namely

$$\Gamma(g) \coloneqq \begin{bmatrix} \langle g, \xi_1 \rangle \\ \vdots \\ \langle g, \xi_N \rangle \end{bmatrix} \qquad \forall\, g \in L^2(\nu) \tag{2.94}$$

and

$$\Xi \coloneqq \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_N \end{bmatrix}$$

so that

$$\Xi^T\Gamma = \begin{bmatrix}\xi_1 \cdots \xi_N\end{bmatrix} \begin{bmatrix}\langle g, \xi_1\rangle \\ \vdots \\ \langle g, \xi_N\rangle\end{bmatrix} = \sum_{i=1}^{N} \langle g, \xi_i\rangle \, \xi_i.$$

Notice that

$$\Gamma\Xi^T = \begin{bmatrix}\Gamma(\xi_1) \cdots \Gamma(\xi_N)\end{bmatrix} = \left[\begin{array}{c|c|c} \langle \xi_1, \xi_1\rangle & & \langle \xi_N, \xi_1\rangle \\ \vdots & \vdots & \vdots \\ \langle \xi_1, \xi_N\rangle & & \langle \xi_N, \xi_N\rangle \end{array}\right] = I$$

which implies

$$\Gamma g = \Gamma\Xi^T\Gamma g = \Gamma\Pi g \qquad \forall\ g \in L^2(\nu)$$

In particular, due to the fact that $L^2(\nu)$ is a Hilbert space and Π is an orthogonal projection, then

$$\Gamma g = 0 \qquad \forall g \perp \mathscr{G}_N$$

So, combining what we have seen until now we can derive the finite dimensional version of the *Koopman operator* as follows

$$U_N = \Pi U : \mathscr{G}_N \xrightarrow{U} L^2(\nu) \xrightarrow{\Pi} \mathscr{G}_N. \tag{2.95}$$

Now using (2.93) in (2.95).

$$U_N g = \Pi U \Pi g = \Xi^T \Gamma U \Xi^T \Gamma g = \Xi^T \mathbf{U} \Gamma g \qquad g \in \mathscr{G}_N \tag{2.96}$$

where

$$\mathbf{U} : \mathbb{C}^N \to \mathbb{C}^N \qquad \text{and} \qquad \mathbf{U}a = \Gamma U \Xi^T a \tag{2.97}$$

is the matrix representation of U_N. It is called the *Koopman matrix* of the system and corresponds to the action of Koopman operator on *observables* $g \in \mathscr{G}_N$ in the coordinates Γg related to the basis $\{\xi_j\}_{j=1}^N$. Since $\Xi^T\Gamma g = \Pi = g$ for $g \in \mathscr{G}_N$ it also follows from the equality (2.96) that

$$\Gamma U g = \mathbf{U}\Gamma g \qquad g \in \mathscr{G}_N$$

If $g = \xi_j$ then $\Gamma\xi_j = e_j$ where e_j is the jth unit vector, and (2.4.3) implies that

$$\mathbf{U}e_j = \Gamma U \xi_j \tag{2.98}$$

Hence the jth column of $\mathbf{U}$ contains the coordinates of $\Pi U \xi_j = U_N \xi_j$ in the basis of functions.

Spectral Properties The hope is that the eigenfunctions and eigenvalues of U_N, which are denoted by $\tilde{\psi}_{\mu_j}$ and $\tilde{\mu}_j$ respectively, approximate those of U. The eigenfunction $\tilde{\psi}_{\mu_j}$ is not necessarily a projection of ψ_{μ_j}. For instance $\tilde{\psi}_{\mu_j} = \Pi\psi_{\mu_j} = \psi_{\mu_j}$ when U commutes with Π, in which case the subspace $\mathscr{G}_N$ is invariant under the action of U. The equality $U_N\tilde{\psi}_{\mu_j} = \tilde{\mu}_j\tilde{\psi}_{\mu_j}$ with (2.93) and (2.96) implies that $\Xi^T\mathbf{U}\Gamma\tilde{\psi}_{\mu_j} = \tilde{\mu}_j\Xi^T\Gamma\tilde{\psi}_{\mu_j}$, or equivalently

$$\mathbf{U}\Gamma\tilde{\psi}_{\mu_j} = \tilde{\mu}_j\Gamma\tilde{\psi}_{\mu_j} \tag{2.99}$$

Thus the eigenvalues of U_N are the eigenvalues of the Koopman matrix $\mathbf{U}$ and the coordinates of the corresponding eigenfunctions in the basis of functions are the right eigenvectors of $\mathbf{U}$.

2.4.4 Data-driven methods

We can use data-driven methods to obtain the spectral properties of the dynamics (2.80) under the action of the *Koopman operator*. But these methods can be divided into those aiming to give a finite-dimensional matrix approximation of the *Koopman operator*, described in 2.4.3, such as the *Extended Dynamic Mode Decomposition* (EDMD), and methods based on generalized Laplace averages (GLA) presented in 2.4.1. The principal difference between these methods consist in the fact that GLA methods approximate directly the eigenfunctions [22] of the *Koopman operator* while the finite-dimensional matrix approximation we tries to approximate the more general operator action. Here we present the EDMD method.

Extended Dynamic Mode Decomposition

Using the same notation used in section 2.4.3 now we present the *Extended Dynamic Mode Decomposition*(EDMD). The EDMD procedure requires:

1. a data set of snapshot pairs $\{(x_m, y_m)\}_{m=1}^{M}$ where $x_m, y_m \in \mathbb{R}^n$ with $y_m = T(x_m)$;
2. N-dimensional linear subspace $\mathscr{G}_N \subset L^2(\nu)$ spanned by the orthonormal basis of functions $\{\xi_j\}_{j=1}^{N}$.

As in the derivation 2.4.3 here we define the following vector-valued function:

$$\Xi(x) = \begin{bmatrix} \xi_1(x) \\ \vdots \\ \xi_N(x) \end{bmatrix}.$$

The data set needed is typically constructed from multiple simulation or from the collection of experimental data. Indeed, if the data were given as a single time series, then for a given state x_m at the instant m, then $y_m = T(x_m) = x_{m+1}$ is the next snapshot in the time series.

Recall that, given an arbitrary positive measure ν on $\mathbb{R}^n$, the space $L^2(\nu)$ is the Hilbert space of all measurable functions $f : \mathbb{R}^n \to \mathbb{C}$ satisfying

$$\|f\|_{L^2(\nu)} := \sqrt{\int_{\mathbb{R}^n} |f(x)|^2 \, d\nu(x)} < \infty. \tag{2.100}$$

Given the data points $x_1, \ldots, x_M$ it is possible define the empirical measure $\hat{\mu}_M(x)$ by

$$\hat{\mu}_M(x) = \frac{1}{M} \sum_{m=1}^{M} \delta_{x_m}(x),$$

where $\delta_{x_m}(x)$ is the *Dirac measure* centred at x_m. In particular, the integral of a function f with respect to $\hat{\mu}_M$ is given by

$$\int_{\mathbb{R}^n} f(x)\, d\hat{\mu}_M(x) = \frac{1}{M} \sum_{m=1}^{M} f(x_m).$$

The EDMD procedure try to generate a finite section of the *Koopman operator* $\mathbf{U} \in \mathbb{C}^{N \times N}$ solving a least squares problem formulated in the following way. If $g \in \mathscr{G}_N$, then can be written as

$$g(x) = \sum_{i=1}^{N} \gamma_i \xi_i(x) = \Xi^T(x)\, \Gamma(g),$$

with $\gamma_i = (\Gamma(g))_i$ i-th components of the vector defined in (2.94). Since $\mathscr{G}_N$ is typically not an invariant subspace under the action of the *Koopman operator*, it holds that

$$Ug = \underbrace{U_N g}_{\in \mathscr{G}_N} + \underbrace{r}_{\in L^2(\nu) \setminus \mathscr{G}_N}$$

where U_N is the finite-dimensional approximation define in (2.95). Now setting $\nu = \hat{\mu}_M$ it

is possible evaluate the quantity:

$$\begin{aligned}
\|Ug - U_N g\|^2_{L^2(\hat{\mu}_M)} &= \int_{\mathbb{R}^n} (Ug - U_N g)^* (x) (Ug - U_N g)(x) \, d\hat{\mu}_M(x) \\
&= \frac{1}{M} \sum_{m=1}^{M} |(Ug - U_N g)(x_m)|^2 \\
&= \frac{1}{M} \sum_{m=1}^{M} |(g \circ T)(x_m) - (U_N g)(x_m)|^2 \\
&= \frac{1}{M} \sum_{m=1}^{M} \left| \left(\Xi^T \Gamma(g) \circ T\right)(x_m) - (U_N g)(x_m) \right|^2 \\
&= \frac{1}{M} \sum_{m=1}^{M} \left| \Xi^T (T(x_m)) \Gamma(g) - (U_N g)(x_m) \right|^2 \\
&= \frac{1}{M} \sum_{m=1}^{M} \left| \Xi^T (y_m) \Gamma(g) - \left(\Xi^T \mathbf{U} \Gamma(g)\right)(x_m) \right|^2 \\
&= \frac{1}{M} \sum_{m=1}^{M} \left| \left(\Xi^T (y_m) - \Xi^T (x_m) \mathbf{U}\right) \Gamma(g) \right|^2
\end{aligned}$$

if we choose $g = \xi_j \, \forall \, j = 1, \ldots, N$, then $\Gamma \xi_j = e_j$. Hence, we define

$$\mathbf{U} = \begin{bmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_N \\ | & & | \end{bmatrix},$$

then the problem of simultaneously minimizing, for all j, all the residual

$$\|U\xi_j - U_N \xi_j\|^2_{L^2(\hat{\mu}_M)} = \frac{1}{M} \sum_{m=1}^{M} \left| \xi_j (y_m) - \Xi^T (x_m) \mathbf{u}_j \right|^2 \quad \forall \, j = 1, \ldots, N$$

setting

$$\boldsymbol{\Xi}(Y) = \begin{bmatrix} \Xi^T (y_1) \\ \vdots \\ \Xi^T (y_M) \end{bmatrix}, \quad \boldsymbol{\Xi}(X) = \begin{bmatrix} \Xi^T (x_1) \\ \vdots \\ \Xi^T (x_M) \end{bmatrix}$$

is equivalent to the following matrix minimization problem:

$$\underset{\mathbf{U}\in\mathbb{C}^{N\times N}}{\arg\min} \left\|\boldsymbol{\Xi}\left(Y\right)-\boldsymbol{\Xi}\left(X\right)\mathbf{U}\right\|_F$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The solution, as done in [18] becomes:

$$\mathbf{U}=\left(\frac{1}{M}\sum_{m=1}^{M}\Xi^T\left(x_m\right)^*\Xi^T\left(x_m\right)\right)^{\dagger}\left(\frac{1}{M}\sum_{m=1}^{M}\Xi^T\left(x_m\right)^*\Xi^T\left(y_m\right)\right), \tag{2.101}$$

where $\cdot^{\dagger}$ denotes the pseudoinverse. As a result, $\mathbf{U}$ is a finite dimensional approximation of U that maps $g\in\mathscr{G}_N$ to some other $\hat{g}\in\mathscr{G}_N$ by minimizing the residuals at the data points. As a consequence, if v_j is the j-th eigenvector of $\mathbf{U}$ with eigenvalue μ_j, then the EDMD approximation of an eigenfunction of $\mathbf{U}$ is

$$\psi_{\mu_j}=\Xi^T v_j$$

Now we apply the numerical method, just derived, in the context of example 7. In this example, and in the rest of the thesis, we have to choose an orthonormal basis of a N-dimensional subspace of $L^2\left(\nu\right)$, for a suitable measure ν. We decide to choose a basis of *Hermite polynomials*. Indeed it is possible to prove that these functions are a basis for problems defined on $\mathbb{R}^n$ (for an exhaustive treatment of this argument see [18]). For us it is sufficient to know that in our context, so with $x\in\mathbb{R}^2$, the basis of the subspace $\mathscr{G}_N$ is defined as the product of the *Hermite polynomials*

$$H_i\left(y\right)\qquad y\in\mathbb{R},\ i\in\mathbb{N}$$

in a single variable, so that $\forall\,\xi_j\in\{\xi_j\}_{j=1}^N$

$$\xi_j\left(x\right)=H_i\left(x_1\right)H_w\left(x_2\right)\qquad\text{for some } i,w\in\mathbb{N}.$$

Example 8*:* Consider the discrete-time dynamical system (2.86). To apply the EDMD procedure, we need a dataset and an orthonormal basis $\{\xi_j\}_{j=1}^N$. For the dataset we consider a set of snapshots $\{(x_m,y_m)\}_{m=1}^M$ with $y_m=T\left(x_m\right)$, $M=50$ and x_m generated randomly with a uniform distribution in the interval $(0,10)$. Also, we consider two distinct bases $\{\xi_j\}_{j=1}^N$ with :

1. $N=4$;

2. $N = 9$.

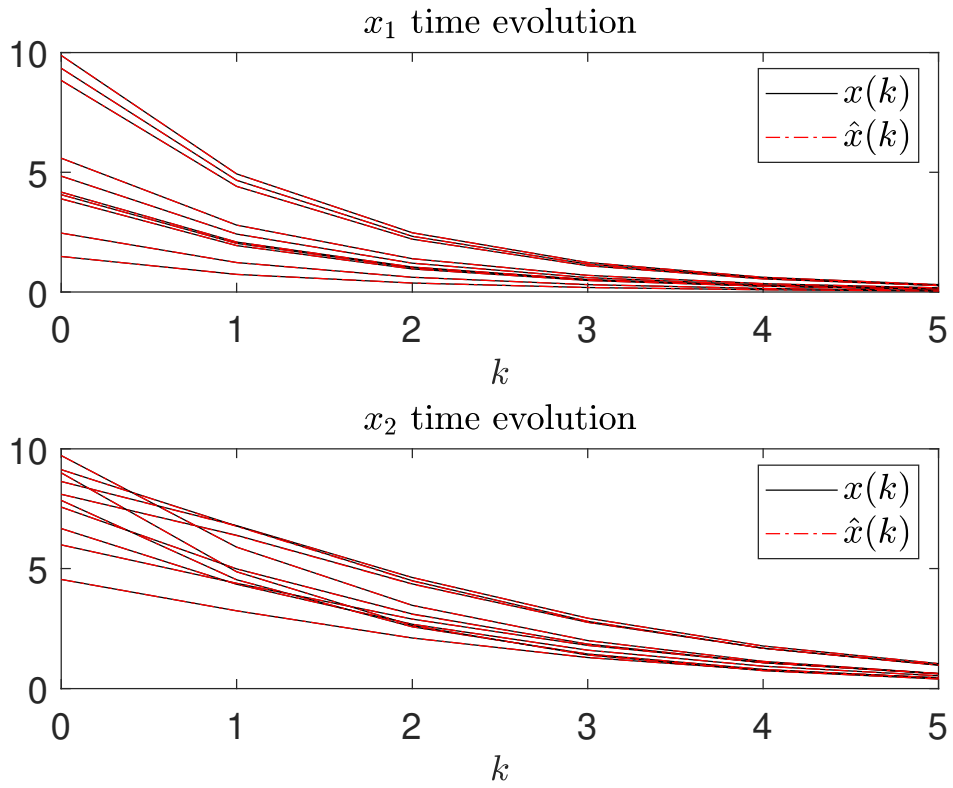


Figure 2.7: Plot of true evolution of $x\,(k)$ and the $\hat{x}\,(k)$, for ten trajectories starting from random initial conditions. The different trajectories are substantially indistinguishable. In this case $N = 4$.

In the first case, we consider the product of the first two *Hermite polynomials*, that are :

$$H_0\,(x_i) = 1$$
$$H_1\,(x_i) = x_i$$

with $i = 1, 2$. Note that in this case, we do not have a quadratic term in x_1. Instead, in the second case, we consider, as a basis, the product of the first three *Hermite polynomials*:

$$H_0\,(x_i) = 1$$
$$H_1\,(x_i) = x_i$$
$$H_2\,(x_i) = x_i^2 - 1$$

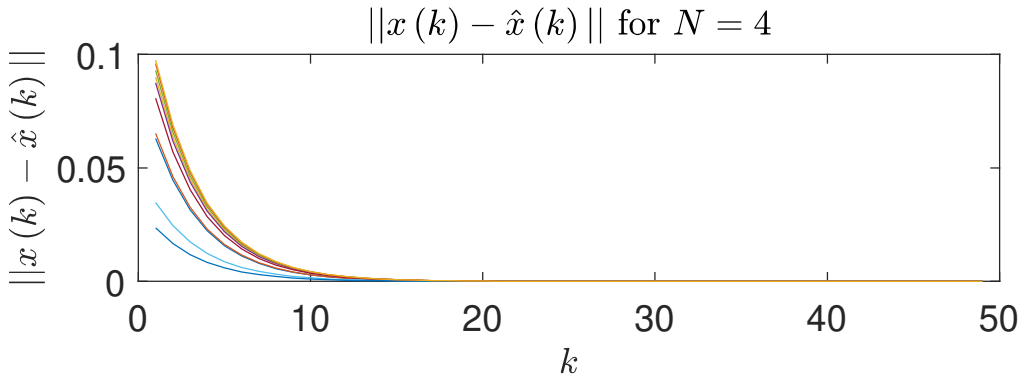


Figure 2.8: The error at every instant k between the true dynamic $x(k)$ and the approximated one $\hat{x}(k)$ for ten different trajectories starting from different initial condition. In this case $N = 4$

In this case, there is a function $\xi_j(x)$ that contains quadratic terms in x_1. This is important because, as we have already seen in example 7, in this case, we can express the dynamics of the *Koopman operator* through the application of the operator to the identity function Id that can be expressed as a linear combination of eigenfunctions of the *Koopman operator* that, if $N = 9$, live in $\mathscr{G}_N$. In the next simulation, figure 2.7, we plot the true dynamic of the system and the one propagated by (2.101) using the same reasoning of example 7. First, we can see as

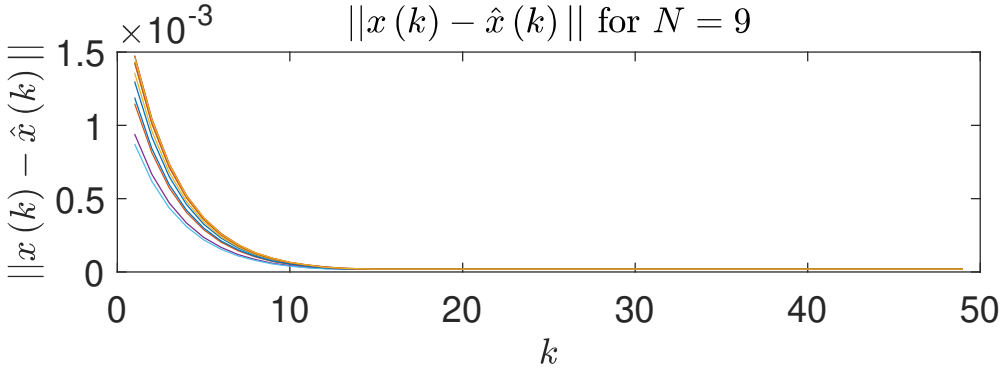


Figure 2.9: The error at every instant k between the true dynamic $x(k)$ and the approximated one $\hat{x}(k)$ for ten different trajectories starting at different initial condition. In this case $N = 9$.

in the two situations the finite dimensional approximations of the *Koopman operator* seem to work well, actually from the plot is not possible to distinguish the approximation from the true evolution of the system. So consider the two plots 2.8 and 2.9. These are the plots of the errors $\|x(k) - \hat{x}(k)\|$, between the true $x(k)$ and the approximated $\hat{x}(k)$. From these plots, the difference between the two is more appreciable and we can see how the choice of a

richer and more powerful basis $\{\xi_j\}_{j=1}^{N}$ can give better results.

In this section the concept of *non-normality*, *spectra* and *pseudospectra* are presented. In particular the focus is on the distinctive behaviour of *non-normal* matrices, their transient effects and the *non-normal* dynamics. To have a detail derivation of all the characteristics of this property it is possible to read [23]. After this section it should be clear why it is natural to think that should exists a link between the excitability property of the neuron and the degree of *non-normality* of its linear representation.

Once given the definition of a *non-normal* matrix, the concept of the *pseudospectra* it is presented. This is essential to understand the degree of *non-normality*.

The definition of the *non-normality* is given by the fact that a matrix is not *normal*. Then the definition of a *normal* matrix is the following :

Definition 18 (Normal matrix)**.** A matrix $\mathbf{A} \in \mathbb{C}^{N\times N}$ is *normal* if it commutes with its conjugate transpose $\mathbf{A}^*$ that is if:

$$\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^* \tag{2.102}$$

One of the most important and useful instrument known to understand the property of a matrix is the eigenvalue analysis. But this in certain applications could be misleading. For example the *non-normality* property introduces some strange behaviours that are not capture by the eigenvalues analysis. One of the most used tool to study *non-normality* is the *pseudospectra*.

Let $\|\cdot\|$ denoting the 2-norm on $\mathbb{C}^N$ and the associated induced norm on $\mathbb{C}^{N\times N}$, the space of complex $N \times N$ matrices, given by:

Definition 19. Let $\mathbf{A} \in \mathbb{C}^{N\times N}$ and $x \in \mathbb{C}^N$, then the induced norm of $\mathbf{A}$ is defined as

$$\|\mathbf{A}\| = \sup_{x\neq 0} \frac{\|\mathbf{A}x\|}{\|x\|} = \sup_{\|x\|=1} \|\mathbf{A}x\| \tag{2.103}$$

Then it is well known that the eigenvalues of a matrix are the set of complex number z that make the matrix $z\mathbf{I} - \mathbf{A}$ singular. But the eigenvalues are not defined in a robust way, an arbitrarily small perturbation of the entries of $\mathbf{A}$ can change drastically the singularity of $z\mathbf{I} - \mathbf{A}$. To deal with this problem, first consider, intuitively, that if the matrix $z\mathbf{I} - \mathbf{A}$ is almost singular then $\left\|(\mathbf{A} - z\mathbf{I})^{-1}\right\|$ is large. To clarify this fact consider the following example.

Example 9: Consider the following square matrix $\mathbf{D} \in \mathbb{C}^{N \times N}$ such that

$$\mathbf{D} = \begin{bmatrix} \epsilon & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} \tag{2.104}$$

Then $\mathbf{D}$ is clearly invertible unless $\epsilon = 0$. Then defining

Definition 20 (Singular value). Consider a complex matrix $\mathbf{M} \in \mathbb{C}^{m \times n}$. We define the singular value as the square root of the non-zero eigenvalues of $\mathbf{M}\mathbf{M}^*$, where $\cdot^*$ denotes the transposition-conjugation operation, arranged in descending order, namely

$$\sigma_i(\mathbf{M}) \coloneqq \sqrt{\lambda_i(\mathbf{M}\mathbf{M}^*)} = \sqrt{\lambda_i(\mathbf{M}^*\mathbf{M})}.$$

It is possible to prove that given $\mathbf{A} \in \mathbb{C}^{N \times N}$

$$\|\mathbf{A}\| = \sigma_{max}(\mathbf{A}) \tag{2.105}$$

where $\sigma_{max}(\mathbf{A})$ denotes the maximum singular value of $\mathbf{A}$. In this case

$$\mathbf{D}^{-1} = \begin{bmatrix} \frac{1}{\epsilon} & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} \tag{2.106}$$

and, if $\epsilon < 1$, then $\sigma_{max}(\mathbf{D}^{-1}) = \frac{1}{\epsilon}$ and so

$$\left\|\mathbf{D}^{-1}\right\| = \frac{1}{\epsilon} \tag{2.107}$$

that in the case $\epsilon \to 0 \;\Rightarrow\; \|\mathbf{D}^{-1}\| \to \infty$.

This is a practical example that explains the fact of characterizing the "degree" of singularity of a matrix by looking at the norm of its inverse, when it exists.

At this point the notion of *ϵ-pseudospectrum* can be introduced.

Definition 21 (*ϵ-pseudospectrum*)**.** Let $\mathbf{A} \in \mathbb{C}^{N \times N}$ and $\epsilon > 0$ be arbitrary. The *ϵ-pseudospectrum*

$\sigma_\epsilon(\mathbf{A})$ of $\mathbf{A}$ is the set of $z \in \mathbb{C}$ such that

$$\left\|(\mathbf{I}z - \mathbf{A})^{-1}\right\| > \frac{1}{\epsilon}. \tag{2.108}$$

Conventionally if $z \in \sigma(\mathbf{A})$, where $\sigma(\mathbf{A})$ is the spectrum of $\mathbf{A}$, then $\left\|(\mathbf{I}z - \mathbf{A})^{-1}\right\| = \infty$.

Now the natural question is if $\left\|(\mathbf{I}z - \mathbf{A})^{-1}\right\|$ is large precisely when z is close to an eigenvalue of $\mathbf{A}$. This is true if the matrix $\mathbf{A}$ is normal but the importance of pseudospectra arises for matrices that are far from normal, indeed for which $\left\|(\mathbf{I}z - \mathbf{A})^{-1}\right\|$ may be large even when z is far from the spectrum, see figure 2.10 where we plot the schematic pseudospectra of

$$\mathbf{A} = \begin{bmatrix} 1+2i & 0 & 0 \\ 0 & 2+3i & 0 \\ 0 & 0 & 3+1.5i \end{bmatrix} \text{ to the left and}$$

$$\mathbf{A} = \begin{bmatrix} 1+2i & \frac{3}{4} & 0 \\ 0 & 2+3i & 30 \\ 0 & 0 & 3+1.5i \end{bmatrix} \text{ to the right.}$$

Given the definition, the *ϵ-pseudospectrum* can be characterized also by the following result via two new equivalences that will be useful in the next analysis.

Theorem 4. *For any matrix $\mathbf{A} \in \mathbb{C}^{N\times N}$ and $\epsilon > 0$, the following four statements are equivalent:*

- *$\sigma_\epsilon(\mathbf{A})$ is the ϵ-pseudospectrum of $\mathbf{A}$;*
- *$\sigma_\epsilon(\mathbf{A})$ is the set of $z \in \mathbb{C}$ such that*

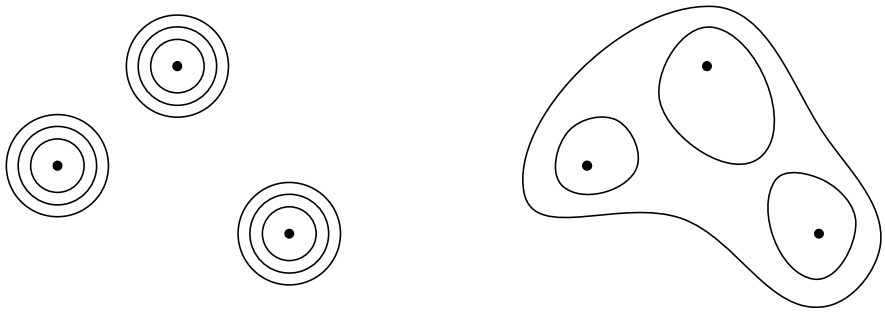
$$z \in \sigma(\mathbf{A} + \mathbf{E}) \tag{2.109}$$

  for some $\mathbf{E} \in \mathbb{C}^{N\times N}$ with $\|\mathbf{E}\| < \epsilon$;

- *$\sigma_\epsilon(\mathbf{A})$ is the set of $z \in \mathbb{C}$ such that*

$$\|(\mathbf{I}z - \mathbf{A})\,\mathbf{v}\| < \epsilon \tag{2.110}$$

  for some $\mathbf{v} \in \mathbb{C}^N$ with $\|\mathbf{v}\| = 1$;



(a) *normal* **(b)** *nonnormal*

Figure 2.10: A schematic view of the level curves of the ϵ-pseudospectrum. The black points represent the eigenvalues of $\mathbf{A}$ and the contours represent the boundary of the $\sigma_\epsilon(\mathbf{A})$ for different values of ϵ.

if $\|\cdot\| = \|\cdot\|_2$, $\sigma_\epsilon(\mathbf{A})$ *is the set of* $z \in \mathbb{C}$ *such that*

$$\sigma_{min}(\mathbf{I}z - \mathbf{A}) < \epsilon \tag{2.111}$$

where $\sigma_{min}(\mathbf{I}z - \mathbf{A})$ *denotes the smallest singular value of* $\mathbf{I}z - \mathbf{A}$

In other words the relation (2.109) says that the *ϵ-pseudospectrum* is the set of numbers that are eigenvalues of some perturbed matrix $\mathbf{A} + \mathbf{E}$ with $\|\mathbf{E}\| < \epsilon$.

From this and the first definition it follows that the *pseudospectra* associated with various ϵ are nested sets,

$$\sigma_{\epsilon_1}(\mathbf{A}) \subseteq \sigma_{\epsilon_2}(\mathbf{A}) \qquad 0 < \epsilon_1 \leq \epsilon_2$$

and that the intersection of all the *pseudospectra* is the spectrum,

$$\bigcap_{\epsilon>0} \sigma_\epsilon(\mathbf{A}) = \sigma(\mathbf{A}).$$

In all these relations the number z is an *ϵ-pseudoeigenvalue* of $\mathbf{A}$, and $\mathbf{v}$ appearing in (2.110) is the corresponding *ϵ-pseudoeigenvector*. So the set of the *ϵ-pseudoeigenvalues* is called the *ϵ-pseudospectrum*.

Now it is possible to link the notion of *ϵ-pseudospectrum* to the concept of *normality* and

nonnormality. First consider the following additional characterization of a *normal* matrix.

Theorem 5. *Given a matrix* $\mathbf{A} \in \mathbb{C}^{N\times N}$ *then:*

$$\mathbf{A} \textit{ is normal } \Leftrightarrow \mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{*} \tag{2.112}$$

where $\mathbf{U}$ *unitary, namely* $\mathbf{U}\mathbf{U}^{*} = \mathbf{I}$*, and* $\mathbf{\Lambda}$ *diagonal matrix of eigenvalues.*

This is useful because if $\mathbf{U}$ is unitary, then

$$\left(\mathbf{I}z - \mathbf{U}\mathbf{A}\mathbf{U}^{*}\right)^{-1} = \left[\mathbf{U}\left(\mathbf{I}z - \mathbf{A}\right)\mathbf{U}^{*}\right]^{-1} = \mathbf{U}\left(\mathbf{I}z - \mathbf{A}\right)^{-1}\mathbf{U}^{*}$$

and therefore

$$\left\|\left(\mathbf{I}z - \mathbf{U}\mathbf{A}\mathbf{U}^{*}\right)^{-1}\right\| = \left\|\left(\mathbf{I}z - \mathbf{A}\right)^{-1}\right\| \quad \forall\, z \in \mathbb{C}.$$

Thus the *resolvent* norm, where the *resolvent* is the matrix $\left(\mathbf{I}z - \mathbf{A}\right)^{-1}$, is invariant with respect to unitary similarity transformations, which implies that the same is true for the *pseudospectra*:

$$\sigma_{\epsilon}\left(\mathbf{A}\right) = \sigma_{\epsilon}\left(\mathbf{U}\mathbf{A}\mathbf{U}^{*}\right) \quad \forall\, \epsilon \geq 0.$$

For a normal matrix, the ϵ*-pseudospectrum* is just the union of the open ϵ-balls around the points of the spectrum. In other words the resolvent norm satisfies

$$\left\|\left(\mathbf{I}z - \mathbf{A}\right)^{-1}\right\| = \frac{1}{\operatorname{dist}\left(z,\, \sigma\left(\mathbf{A}\right)\right)},$$

where $\operatorname{dist}\left(z,\, \sigma\left(\mathbf{A}\right)\right)$ denotes the usual distance of a point to a set in the complex plane. The next theorem is presented to specify better this concept. Consider the following sets:

$$\Delta_{\epsilon} = \left\{z \in \mathbb{C} \,:\, |z| < \epsilon\right\}$$

and

$$\sigma\left(\mathbf{A}\right) + \Delta_{\epsilon} = \left\{z \,:\, z = z_1 + z_2,\, z_1 \in \sigma\left(\mathbf{A}\right),\, z_2 \in \Delta_{\epsilon}\right\}.$$

Then it is possible to prove the following result:

Theorem 6 (*Pseudospectra* of a normal matrix)**.** *For any* $\mathbf{A} \in \mathbb{C}^{N\times N}$*,*

$$\sigma_{\epsilon}\left(\mathbf{A}\right) \supseteq \sigma\left(\mathbf{A}\right) + \Delta_{\epsilon} \quad \forall\, \epsilon > 0, \tag{2.113}$$

and if $\mathbf{A}$ *is normal, then*

$$\sigma_\epsilon(\mathbf{A}) = \sigma(\mathbf{A}) + \Delta_\epsilon \quad \forall\, \epsilon > 0. \tag{2.114}$$

Conversely (2.114) *implies that* $\mathbf{A}$ *is normal.*

Now suppose that $\mathbf{A}$ is diagonalizable but not necessarily normal, and let $\mathbf{V} \in \mathbb{C}^{N\times N}$ be a *matrix of eigenvectors* of $\mathbf{A}$. Then the condition number of this basis of eigenvectors, is defined

$$\kappa(\mathbf{V}) \equiv \|\mathbf{V}\| \left\|\mathbf{V}^{-1}\right\| = \frac{\sigma_{max}(\mathbf{V})}{\sigma_{min}(\mathbf{V})}, \tag{2.115}$$

where $\sigma_{max}(\mathbf{V})$ and $\sigma_{min}(\mathbf{V})$ are the largest and the smallest singular values of $\mathbf{V}$. In general, $\kappa(\mathbf{V})$ may be any number larger or equal to 1, and the value $\kappa(\mathbf{V}) = 1$ is possible if and only if $\mathbf{A}$ is normal.

The condition number of $\mathbf{V}$ provides an upper bound for the condition numbers of the individual eigenvalues of $\mathbf{A}$. This fact is stated in the next theorem.

Theorem 7 (Bauer-Fike)**.** *Suppose* $\mathbf{A} \in \mathbb{C}^{N\times N}$ *is diagonalizable,* $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$*. Then for each* $\epsilon > 0$*, with* $\|\cdot\| = \|\cdot\|_2$*,*

$$\sigma(\mathbf{A}) + \Delta_\epsilon \subseteq \sigma_\epsilon(\mathbf{A}) \subseteq \sigma(\mathbf{A}) + \Delta_{\epsilon\kappa(\mathbf{V})} \tag{2.116}$$

Proof. The first inclusion was established in (2.113). For the second is necessary to compute the *resolvent* of $\mathbf{A}$

$$(\mathbf{I}z - \mathbf{A})^{-1} = \left(\mathbf{I}z - \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}\right)^{-1} = \left[\mathbf{V}(\mathbf{I}z - \Lambda)\mathbf{V}^{-1}\right]^{-1} = \mathbf{V}(\mathbf{I}z - \Lambda)^{-1}\mathbf{V}^{-1}$$

which implies

$$\left\|(\mathbf{I}z - \mathbf{A})^{-1}\right\| \leq \kappa(\mathbf{V}) \left\|(\mathbf{I}z - \mathbf{\Lambda})^{-1}\right\| = \frac{\kappa(\mathbf{V})}{\operatorname{dist}(z,\, \sigma(\mathbf{A}))}$$

and the definition (2.108) completes the proof. □

2.5.1 Scalar measure of *non-normality*

We conclude the section presenting a scalar measurement of the *non-normality*. It is known that for any square matrix $\mathbf{A}$ there exists a Schur decomposition, defined as

Definition 22. If $\mathbf{A} \in \mathbb{C}^{N\times N}$ then $\mathbf{A}$ can be expressed as

$$\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^{*} \tag{2.117}$$

with $\mathbf{U}$ a *unitary* matrix and $\mathbf{T}$ an upper triangular matrix, which is called a *Schur form* of $\mathbf{A}$.

With the *Schur decomposition* is possible to express any square matrix $\mathbf{A}$ as

$$\mathbf{A} = \mathbf{U}\left(\mathbf{\Lambda} + \mathbf{R}\right)\mathbf{U}^{*} \tag{2.118}$$

with $\mathbf{\Lambda}$ a diagonal matrix and $\mathbf{R}$ strictly upper triangular. The idea to find a metric for the *non-normality* degree is based on the fact that if $\mathbf{A}$ is normal then $\mathbf{R} = 0$. So, because in general, the Schur decomposition is not unique, a first estimate of the degree of the *non-normality* of $\mathbf{A}$ can be given by

$$\mathrm{dep}\left(\mathbf{A}\right) = \min_{\text{all the Schur decomposition of } \mathbf{A}} \left\|\mathbf{R}\right\|. \tag{2.119}$$

But if we consider the Frobenius norm $\left\|\cdot\right\|_F$ it is possible to prove that

$$\left\|\mathbf{R}\right\|_F^2 = \left\|\mathbf{A}\right\|_F^2 - \left\|\mathbf{\Lambda}\right\|_F^2, \tag{2.120}$$

and thus

$$\mathrm{dep}_F\left(\mathbf{A}\right) = \sqrt{\left\|\mathbf{A}\right\|_F^2 - \left\|\mathbf{\Lambda}\right\|_F^2}. \tag{2.121}$$

It is possible to define various and different metrics for the *non-normality* but in this thesis we used only (2.121) as a measure of the *non-normality* of a matrix.

2.5.2 Transients and *pseudospectra*

In a linear dynamical system it may happen that the transient behaviour differs from the behaviour for large times. It is in these cases that the eigenvalues fail to capture the transients. Instead the *pseudospectra* can be very useful to quantify and detect these transient phenomena that the *spectra* do not capture. This fact is introduce and formalize in this section.

Now some necessary concepts are defined :

Definition 23 (Spectral radius)**.** Let $\sigma(\mathbf{A})$ be the spectrum of a matrix $\mathbf{A} \in \mathbb{C}^{N\times N}$. The spectral radius of $\mathbf{A}$ is defined as

$$\rho(\mathbf{A}) := \sup_{z\in\sigma(\mathbf{A})} |z| .$$

Definition 24 (ϵ-pseudospectral radius)**.** Let $\sigma_\epsilon(\mathbf{A})$ the *ϵ-pseudospectrum* of $\mathbf{A}$ then

$$\rho_\epsilon(\mathbf{A}) := \sup_{z\in\sigma_\epsilon(\mathbf{A})} |z|$$

is the ϵ-pseudospectral radius of $\mathbf{A}$.

In the following we present some useful bounds on the transient response of a linear dynamical system that are based on the *pseudospectra*. For discrete time dynamical system it is known that the evolution of the system

$$x(k+1) = \mathbf{A}x(k)$$

is given by

$$x(k) = \mathbf{A}^k x(0)$$

So the main objective is to understand the behaviour of the quantity $\left\|\mathbf{A}^k\right\|$ as function of k. It is known that the asymptotic growth rate of $\left\|\mathbf{A}^k\right\|$ is determined by $\rho(\mathbf{A})$, the spectral radius of $\mathbf{A}$. Indeed the equation, called Gelfand's formula [24], tells that

$$\lim_{k\to\infty} \left\|\mathbf{A}^k\right\|^{\frac{1}{k}} = \rho(\mathbf{A}) \tag{2.122}$$

holds for any matrix $\mathbf{A}$. The transient behaviours is, in a certain sense, related to the "intermediate" values of k from the case $k = 0$ and $k \to \infty$.

To understand this, in the following we present two useful bounds based on the ϵ-*pseudospectrum* that give an interesting estimate of $\left\|\mathbf{A}^k\right\|$ for the intermediate values of k. Before to introduce these bounds, it is necessary to state a fundamental theorem that represents the relationships between $\mathbf{A}^k$ and $(\mathbf{I}z - \mathbf{A})^{-1}$.

Theorem 8. *Let $\mathbf{A} \in \mathbb{C}^{N\times N}$. There exist $\gamma > 0$ and $M \geq 1$ such that*

$$\left\|\mathbf{A}^k\right\| \leq M\gamma^k \qquad \forall\, k \geq 0. \tag{2.123}$$

Any $z \in \mathbb{C}$ with $|z| > \gamma$ is in the resolvent set of $\mathbf{A}$, so all the z for what $(\mathbf{I}z - \mathbf{A})^{-1}$ exists, and the resolvent for such z is given by the convergent series

$$(\mathbf{I}z - \mathbf{A})^{-1} = z^{-1}\left(\mathbf{I} + z^{-1}\mathbf{A} + \left(z^{-1}\mathbf{A}\right)^2 + \cdots\right) \tag{2.124}$$

Conversely, for any $k \geq 0$,

$$\mathbf{A}^k = \frac{1}{2\pi i}\int_\Gamma z^k\,(\mathbf{I}z - \mathbf{A})^{-1}\,dz \tag{2.125}$$

where Γ is any closed contour enclosing $\sigma(\mathbf{A})$ in its interior.

For a proof of the previous theorem see [25]. The first upper bound is given in the following theorem

Theorem 9. *If $\mathbf{A}$ is a matrix and $k \geq 0$ is arbitrary, then for any $\epsilon > 0$*

$$\left\|\mathbf{A}^k\right\| \leq \frac{\rho_\epsilon(\mathbf{A})^{k+1}}{\epsilon} \tag{2.126}$$

Proof. In (2.125) take Γ as the circle around the origin of radius $\rho_\epsilon(\mathbf{A})$. Then

$$\begin{aligned}
\left\|\mathbf{A}^k\right\| &= \left|\frac{1}{2\pi i}\right|\left\|\int_\Gamma z^k\,(\mathbf{I}z - \mathbf{A})^{-1}\,dz\right\| = \frac{1}{2\pi}\left\|\int_\Gamma z^k\,(\mathbf{I}z - \mathbf{A})^{-1}\,dz\right\| \\
&\leq \frac{1}{2\pi}\int_\Gamma \left\|z^k\,(\mathbf{I}z - \mathbf{A})^{-1}\right\|dz = \frac{1}{2\pi}\int_\Gamma |z|^k\left\|(\mathbf{I}z - \mathbf{A})^{-1}\right\|dz \\
&\leq \frac{1}{2\pi}\frac{\rho_\epsilon(\mathbf{A})^k}{\epsilon}\int_\Gamma dz = \frac{1}{\cancel{2\pi}}\frac{\rho_\epsilon(\mathbf{A})^k}{\epsilon}\cancel{2\pi}\rho_\epsilon(\mathbf{A}) = \frac{\rho_\epsilon(\mathbf{A})^{k+1}}{\epsilon}
\end{aligned} \tag{2.127}$$

□

This theorem gives an upper bound on the $\left\|\mathbf{A}^k\right\|$ based on the *ϵ-pseudospectrum*. Instead the following theorem states a lower bound on $\left\|\mathbf{A}^k\right\|$.

Theorem 10. *Assume there exists $z \in \mathbb{C}$ with $|z| > 1$ such that*

$$K \coloneqq (|z| - 1) \left\|(\mathbf{I}z - \mathbf{A})\right\|^{-1} > 1.$$

Then

$$\sup_{k \geq 0} \left\|\mathbf{A}^k\right\| \geq |z| \, K - |z| + 1 > K. \tag{2.128}$$

Moreover we have

$$\sup_{k \geq 0} \left\|\mathbf{A}^k\right\| \geq (\rho_\epsilon(\mathbf{A}) - 1) \frac{\rho_\epsilon(\mathbf{A}) - \epsilon}{\epsilon} \geq \frac{\rho_\epsilon(\mathbf{A}) - 1}{\epsilon} \tag{2.129}$$

for all $\epsilon > 0$.

Proof. Let $r = |z|$ and $\sup_{k \geq 0} \left\|\mathbf{A}^k\right\| = M$. Then by (2.124) and theorem 8

$$\begin{aligned}
\frac{rK}{r-1} &= r \left\|(\mathbf{I}z - \mathbf{A})^{-1}\right\| = r \left\|z^{-1}\left(\mathbf{I} + z^{-1}\mathbf{A} + \left(z^{-1}\mathbf{A}\right)^2 + \cdots\right)\right\| \\
&= \frac{\cancel{|z|}}{\cancel{|z|}} \left\|\left(\mathbf{I} + z^{-1}\mathbf{A} + \left(z^{-1}\mathbf{A}\right)^2 + \cdots\right)\right\| \\
&\leq \|\mathbf{I}\| + \frac{1}{|z|}\|\mathbf{A}\| + \frac{1}{|z|^2}\left\|\mathbf{A}^2\right\| + \cdots \\
&\leq 1 + \sum_{k=1}^{\infty} \frac{M}{|z|^k} = 1 + M \sum_{k=1}^{\infty} r^{-k} = 1 + \frac{M}{r-1}
\end{aligned} \tag{2.130}$$

which implies theorem (2.128). Indeed by (2.130)

$$\frac{rK}{r-1} \leq 1 + \frac{M}{r-1} \Rightarrow \sup_{k \geq 0} \left\|\mathbf{A}^k\right\| = M \geq rK - r + 1.$$

Now fixing $\left\|(\mathbf{I}z - \mathbf{A})^{-1}\right\|$ and taking the corresponding largest-modulus value of z, then $|z| = \rho_\epsilon(\mathbf{A})$ and so

$$\begin{aligned}
\left\|(\mathbf{I}z - \mathbf{A})^{-1}\right\| = \frac{K}{|z| - 1} &\Rightarrow K = (|z| - 1) \left\|(\mathbf{I}z - \mathbf{A})^{-1}\right\| \\
&\Rightarrow K = (\rho_\epsilon(\mathbf{A}) - 1) \frac{1}{\epsilon}
\end{aligned}$$

and so inequality (2.129) holds directly by bound (2.128). □

The bounds we presented are useful to characterize the transient behaviour of a system. Moreover the bounds (2.126) and (2.129) tell that the transient norm of the matrix $\mathbf{A}^k$, that in a certain sense characterize the "impulsivity" of the linear system $\mathbf{A}$, is determined by the quantity $\rho_\epsilon(\mathbf{A})$: the greater is the ϵ-pseudospectral radius, the 'greater' is the response of the system.

Now from theorem 7 it is evident that, to increase the ϵ-pseudospectral radius, it is necessary resort to the *non-normality*. Indeed, given two matrices with the same spectrum, one *normal* and the other *non-normal*, by theorem 6, the *normal* matrix has always smaller ϵ-pseudospectral radius, and so has a less "impulsive" transient behaviour. This fact is the fundamental link between the *non-normality* and the transient behaviour of a linear system.

3
Neuron models

In this chapter, the models of the neuron activity of interest are presented. In particular, in section 3.1 the fundamental work of Hodgkin and Huxley [3] is summarized. This work is a milestone for the further development of the knowledge about the neurons. All the past and present mathematical conductance-based models of the neuron are based on that work. In addition to the Hodgkin-Huxley model, another model, the FitzHugh-Nagumo, is presented. This model, that represents a planar version of the Hodgkin-Huxley model, is much more tractable but, remains a good neuron model. Many other models have been proposed in the literature. For example a researcher, who has worked extensively on neuronal modelling, is Rodolphe Sepulchre. He studied in depth conductance-based models, and one of the models he analysed is considered in the section 3.2. This *non-linear* model consists in a generalization of the *FitzHugh-Nagumo* model and represents a good starting point for a further analysis.

One of the most important models that explains the neuron behaviour is the one developed by A. Hodgkin and A. Huxley in [3]. This is a model that describes how potentials of neurons are initiated and propagated. More precisely, the authors develop a model to explain the ionic mechanisms underlying the initiation and propagation of action potentials in the squid giant axon and it is based on the representation of the neuronal model as an RC electrical circuit as illustrated in figure 3.1.

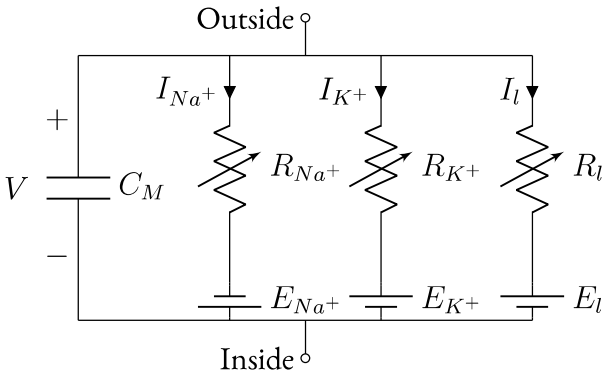


Figure 3.1: Basic components of Hodgkin-Huxley model

Neurons can generate a signal in response to an external stimulus, because they have a membrane that can maintain a voltage difference between the extracellular and intracellular medium. The cellular membrane is therefore modelled as a capacitor. But the voltage across the membrane can also vary because there are ions (primarily sodium (Na^+), potassium (K^+) and calcium (Ca^{2+})) that can flow from the extracellular into the intracellular medium, or vice-versa, thanks to specific transmembrane proteins called ion channels. The natural model of each ionic channel is a resistor in series with a battery. The battery represents the equilibrium potential at which the channel is "closed". Calcium and sodium have an equilibrium potential higher than the resting potential of the neuron, instead is lower for the potassium ions. Therefore, an inward current of sodium or calcium tends to depolarize the membrane voltage, whereas an (outward) flow of potassium tends to hyper-polarizes voltage across the membrane.

But also this model has some weaknesses. For example, in their derivation, Hodgkin and Huxley considered only two ionic currents: a depolarizing sodium current and a hyperpolarizing potassium current. This because calcium channels were discovered soon after their

seminal work. The sum of all remaining currents are combined in a passive leakage current. With the previous considerations, it is possible to derive a circuit that represents the electric model of the neuron. Solving the circuit, in figure 3.1, using directly the First Kirchhoff's law, we obtain

$$C\frac{dV}{dt}(t) = I_{Na^+}(t) + I_{K^+}(t) + I_l(t) + I_{app}(t) \tag{3.1}$$

where I_{app} is an externally applied current. By the Ohm's law, it is possible to derive the following set of equations:

$$I_{Na^+}(t) = g_{Na^+}(t)(V(t) + E_{Na^+}) \tag{3.2}$$

$$I_{K^+}(t) = g_{K^+}(t)(V(t) - E_{K^+}) \tag{3.3}$$

$$I_l(t) = g_l(t)(V(t) - E_l) \tag{3.4}$$

where $g_{Na^+}(t)$, $g_{K^+}(t)$ and $g_l(t)$ are the respective conductances of the relative resistance.

We want to highlight that Hodgkin and Huxley were able to separate experimentally the contribution of sodium and potassium currents in order to model, independently, the voltage dependence of their respective conductances. From (3.2) it is not completely clear where is the *non-linearity* of the model, because this is inherited directly by the voltage dependence of ionic conductances.

But the true weakness of this model is that the various conductances depends on many parameters that are difficult to choose a priori. And for these reasons the system is difficult to study. Moreover, it cannot be solved analytically. To overcome this issue, several simplified neuronal models have also been developed, such as the FitzHugh-Nagumo model and a generalized version of this model discussed in 3.2.

The Hodgkin-Huxley model has been presented in the previous section 3.1. We also have to highlight that some problems arise when we want to analyse it because of its complexity. So we need a simplified version of this model.

In this direction, Rodolphe Sepulchre, proposed different models that try to simplify the analysis of the Hodgkin-Huxley model and, at the same time, to better characterize the neuron behaviour. One of these models is presented in [4] and corresponds to a generalized version of the FitzHugh-Nagumo model which, in turn is a planar version of the Hodgkin-Huxley model,

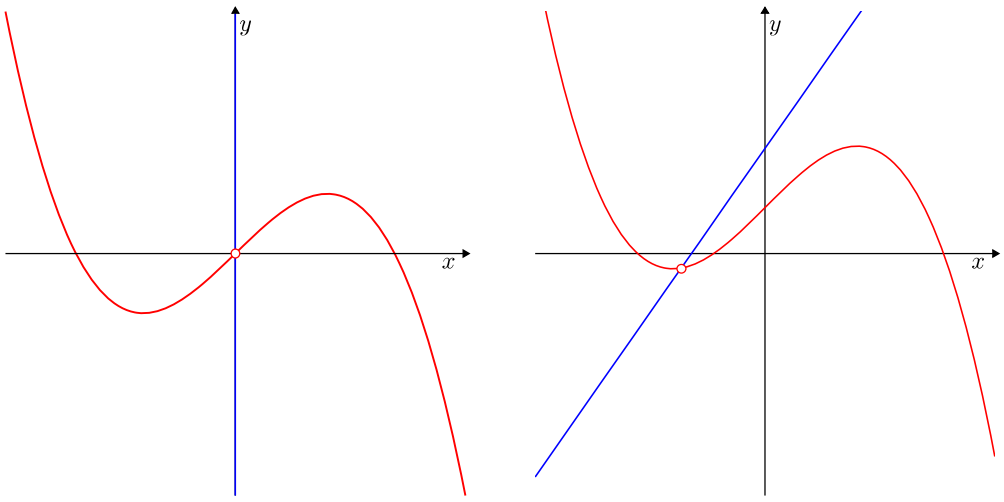
$$\begin{cases} \dot{v}(t) = v(t) - \dfrac{v(t)^3}{3} - n(t) + I_{app} & \text{(3.5a)} \\ \dot{n}(t) = \epsilon(v(t) + a - bn(t)) & \text{(3.5b)} \end{cases}$$
$$v(t),\ n(t),\ I_{app},\ a,\ b \in \mathbb{R},$$

where $v(t)$ is the action potential and $n(t)$ represent a recovery variable with a linear dynamics that provides a slow negative feedback and tends to recover the system to its rest state and $\epsilon > 0$. In this case I_{app} does not represent properly an applied current, but the magnitude of the constant current corresponding to an external stimulus. Finally, a and b are two parameters. More in detail, equation (3.5a) represent the *fast dynamics* of the membrane potential $v(t)$, whereas equation (3.5b) describes the evolution of the recovery variable $n(t)$ that aggregates the gating of various ionic channels.

The first who proposed this model was FitzHugh [26]. He started from the famous Van der Pol model

$$\begin{cases} \dot{v}(t) = v(t) - \dfrac{v(t)^3}{3} - n(t) & \text{(3.6a)} \\ \dot{n}(t) = \epsilon v(t) & \text{(3.6b)} \end{cases}$$

in order to explain the basic properties of excitability, as exhibited by the more complex Hodgkin-Huxley equations, derived model (3.5). We can immediately see that equations (3.6) correspond to the FitzHugh-Nagumo model (3.5) in the case $a = b = I_{app} = 0$. The difference between the Van der Pol and FitzHugh-Nagumo equations can be also seen from the nullclines, defined as



(a) Van der Pol nullclines and the unique unstable point in the origin plotted in red.

(b) FitzHugh-Nagumo nullclines and the unique unstable point in the origin plotted in red. In this case $I_{app} = 0.5$, $a = 0.8$ and $b = 0.7$.

Figure 3.2: Plot of the nullclines for the van der Pol and the FitzHugh-Nagumo models.

Definition 25. Consider a system of ordinary differential equations:

$$\begin{cases} \dot{x}_1(t) &= f_1(x_1(t), \ldots, x_n(t)) \\ &\vdots \\ \dot{x}_n(t) &= f_n(x_1(t), \ldots, x_n(t)) \end{cases}$$

with $x_1(t), \ldots, x_n(t) \in \mathbb{R}$. Then the j-nullcline is the set

$$\{(x_1, \ldots, x_n) \in \mathbb{R}^n : f_j(x_1, \ldots, x_n) = 0\}.$$

The nullclines of the models (3.5) and (3.6) are depicted in figure 3.2a. The first has a vertical line and a cubic that intersects in a single equilibrium point, which is always unstable (figure 3.2a). To see this it is sufficient to resort to the Lyapunov's first method. Indeed, the Jacobian J evaluated at the unique equilibrium point $v = n = 0$ is

$$J = \begin{bmatrix} \frac{\partial \dot{v}}{\partial v} & \frac{\partial \dot{v}}{\partial n} \\ \frac{\partial \dot{n}}{\partial v} & \frac{\partial \dot{n}}{\partial n} \end{bmatrix}_{(0,0)} = \begin{bmatrix} 1 & -1 \\ \epsilon & 0 \end{bmatrix}$$

that has always a positive eigenvalue and so the system is always unstable. The FitzHugh-Nagumo model (3.5), instead has the same cubic nullcline, that this time can be translated by a constant parameter I_{app}, but a more sophisticated linear nullcline in order to resemble a real neuron. This model also has only one equilibrium point, but displays, a threshold phenomenon based on a parameter change that recalls a "current stimulation". To ensure that the system has only an equilibrium point, the parameter b is positive and chosen sufficiently big to avoid multiple interconnections between the nullclines. Also, the slope of the linear nullclines is fundamental to determine the stability conditions of the equilibrium point. To understand this it is interesting to resort to Lyapunov's first method. So consider the Jacobian of the system (3.5) as a function of the equilibrium point (v_0, n_0) so $J(v_0, n_0)$. This is

$$J(v_0, n_0) = \begin{bmatrix} 1 - v_0^2 & -1 \\ \epsilon & -\epsilon b \end{bmatrix}. \tag{3.7}$$

Now the interesting thing is to see the characteristic polynomial, function of $s \in \mathbb{C}$, of this matrix

$$s^2 + \left[\left(v_0^2 - 1\right) + \epsilon b\right] s + \epsilon \left[b\left(v_0^2 - 1\right) + 1\right]. \tag{3.8}$$

In (3.8) v_0, the v coordinate of the equilibrium point, can be set by moving I_{app} and a. So at first it is possible to see, resorting to Cartesio's rule, that for the family of the FitzHugh-Nagumo system (3.5), that have an equilibrium point in the descending branches of the cubic, for which $1 - v_0^2 < 0$, independently from $b > 0$ the equilibrium point is asymptotically stable. Instead, in the case that the equilibrium point is in the ascending branch of the cubic, so where $1 - v_0^2 > 0$, the stability of the system depends strongly on b. Indeed, again from the Cartesio's rule applied to (3.8), the system is asymptotically stable if

$$\begin{cases} b > \dfrac{1 - v^2}{\epsilon} & \text{(3.9a)} \\ b < \dfrac{1}{1 - v^2}, & \text{(3.9b)} \end{cases}$$

so if b is

$$\frac{1 - v^2}{\epsilon} < b < \frac{1}{1 - v^2}.$$

which is possible only in the case $\epsilon > \left(1 - v_0^2\right)^2$. It is also interesting to see how the ϵ parameter, which regulates the system into a *fast* and a *slow dynamic* plays a fundamental role. Meanwhile, the smaller is ϵ, the greater is the temporal separation between the dynamics.

Typically ϵ is small, and so the right side of inequality (3.9a) can be very high. On the contrary, with (3.9b) we require that b remain sufficiently small. And, if we choose I_{app} and a such that the equilibrium point is on the ascending branch, the class of systems that are asymptotically stable parametrized by b tends to become smaller and smaller and the equilibrium point tends to be unstable. Not only, but we can also prove resorting to the *geometric singular perturbation theory* 2.1.2, the presence of a *limit cycle*. In the end, the two terms a and I_{app} play the same role. Formally, all the properties explained below are reached by adding only a constant term to the equations. However, because the physiologists usually are used to this, both equations have a constant term: the one in the first equation mimicking the experimental injection of external current into the membrane; the one in the second equation ensures that the equilibrium point for $I_{app} = 0$ (no stimulation) lies on the right descending branch, and hence it is stable.

A fundamental characteristic of this type of models is the fact that the two dynamics act at very different scales of time. And so, as we have seen before, this opens the door to the study by using of the *Geometrical Singular Perturbation Theory* that will be fundamental in the analysis of the generalized version of the FitzHugh-Nagumo model.

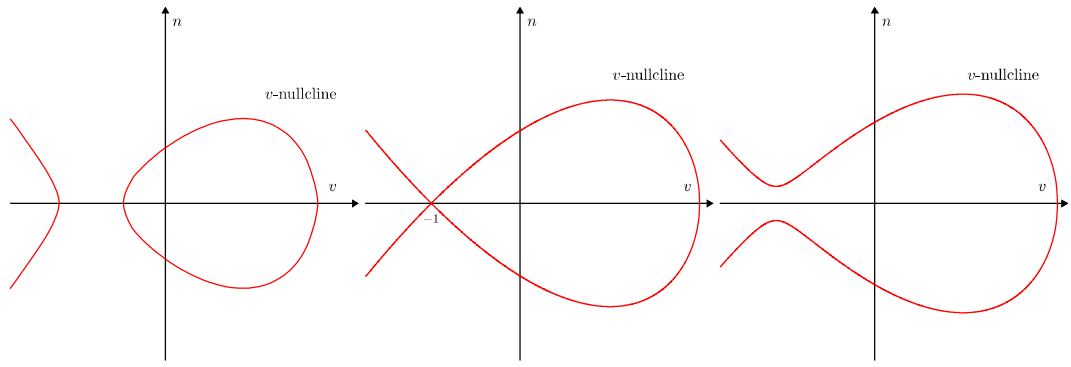
Now the generalized FitzHugh-Nagumo model [4] is presented. The following system of equations determines the evolution of the model

$$\begin{cases} \dot{v}(t) = v(t) - \dfrac{v(t)^3}{3} - n(t)^2 + I_{app} & \text{(3.10a)} \\ \dot{n}(t) = \epsilon(n_\infty(v(t) - v_0) + n_0 - n(t)) & \text{(3.10b)} \end{cases}$$

where

$$n_\infty(v(t)) := \frac{2}{1 + e^{-5v(t)}} \tag{3.11}$$

The dynamics (3.10a) is identical to the one of the FitzHugh-Nagumo model (3.5a), with the only difference that in (3.10a) we substitute the linear term n with a quadratic term n^2 . This yields the modification of the nullcline $\dot{v} = 0$ as illustrated in figure 3.3. With the pres-



(a) $I_{app} < I^*$ (b) $I_{app} = I^*$ (c) $I_{app} > I^*$

Figure 3.3: v-nullclines of (3.10) for different values of I_{app}.

ence of the quadratic term n^2, the v-nullcline is considerably changed. In particular the shape of the nullcline depends strongly on the applied current I_{app}, that in the case $I_{app} = I^* := \frac{2}{3}$ represents a *transcritical bifurcation organizing center*. For our scope the most interesting situation is the one plotted in figure 3.3c, which corresponds to the case $I_{app} > I^*$. Instead, in 3.4 we represent the n-nullcline for the case $v_0 = n_0 = 0$. The parameters v_0 and n_0 can be modified to translate up/down or left/right the n-nulllcine. With respect to the classical FitzHugh-Nagumo model (3.5), we have sigmoid n-nullcline. In this way, we model the first-order relaxation of the ionic current. It is also interesting to notice that it can be proved that

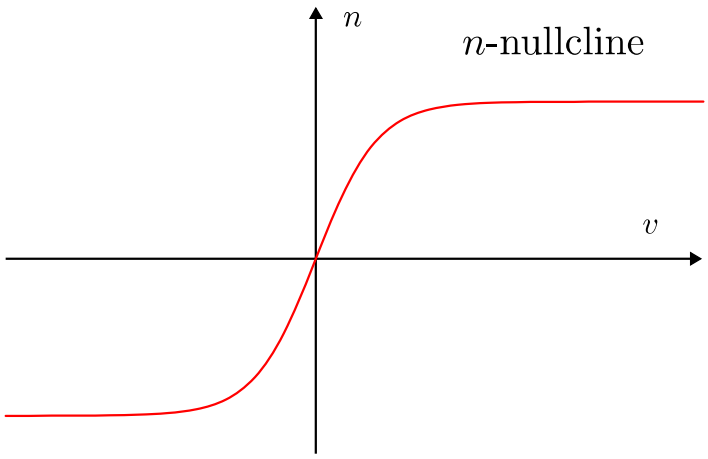


Figure 3.4: The new n-nullcline is plotted with $v_0 = 0$ and $n_0 = 0$.

the region

$$M_{n_0} := \{(v, n) \in \mathbb{R}^2 : n \in (n_0, n_0 + 2)\} \tag{3.12}$$

is attractive and invariant for the dynamics (3.10).

The model (3.10) has three free parameters (I_{app}, n_0, v_0). The power of this model is the fact that changing these parameters, it is possible to reproduce five types of excitability that answer different physiological questions. For the scope of this thesis, such refined characterization is less important, because other properties are investigated and so the analysis is restricted only to two types of excitability discussed in [4].

The first type of excitability consists in setting the parameters so as to have a dynamical system with a unique stable equilibrium point that can exhibit a large and rapidly changing response. This is the main type of excitability considered. Another interesting type of excitability is the one linked to the presence of oscillating phenomena and limit cycles. Using the language of the paper [4], the first type of excitability is called the type III instead the second is called the type IV.

Type III excitability was studied and observed in the squid giant axons[27], auditory brain stem [28], isolated axons from Carcinus maenas [29]. The v and the n-nullcliens associated to this type of excitability are plotted in figure 3.5. If type III is an excitable mechanism explainable also with the standard FitzHugh-Nagumo model (3.5), type IV, instead, requires the generalized version of the model (3.10). Indeed, in this case the quadratic term $n\,(t)^2$ of (3.10a) plays a fundamental role. We present this type here because it exhibits non-trivial behaviour such as the *multistability* and the presence of *limit cycles*. Examples of neurons

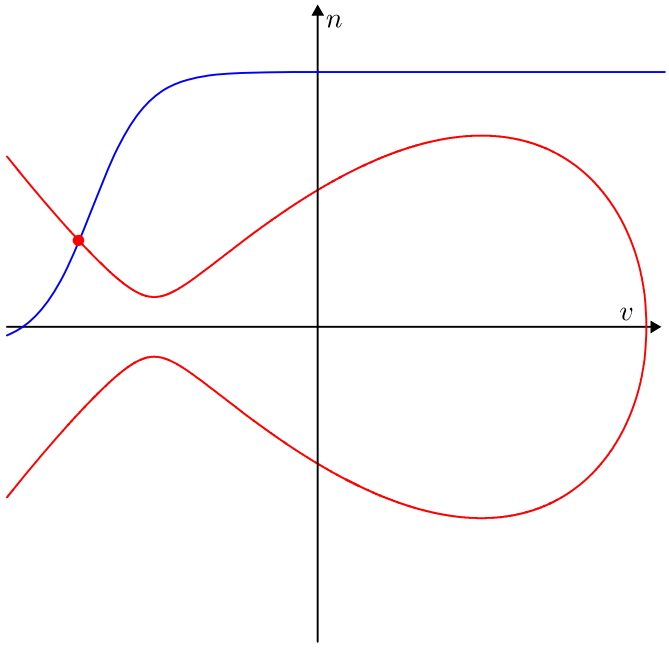


Figure 3.5: v-nullclines and n-nullclines for a system that exhibits a type III excitability. The filled red circle represents the unique stable point.

exhibiting type IV excitability include: subthalamic nucleus neurons [30], thalamo-cortical reticular and relay neurons with deinactivated T-type calcium current (hyperpolarized state) [31], dopaminergic neurons [32], superficial pyramidal neurons [33]. The nullclines associated to this type of excitability are depicted in figure 3.6. Now we proceed with a qualitative study of the trajectories in type III and IV of excitability.

In case of type III of excitability the system admits a unique equilibrium point and it is possible prove that this is stable. In figure 3.7 it is possible to see a simulation of a trajectory starting with initial condition around the equilibrium. Below, in the same figure, we plot only the action potential $v(t)$ as function of time. The situation is much different in type IV conditions. Also in this case there is an only equilibrium point, but now the equilibrium is unstable and it is possible to prove that around that point there exists a *limit cycle*. To see that we will use the techniques of section 2.1.2. Equations (3.10) represent a *slow-fast system* and the critical manifold is the set

$$M_0 = \left\{ (v, n) \in \mathbb{R}^2 : v - \frac{v^3}{3} - n^2 + I_{app} = 0 \right\}. \tag{3.13}$$

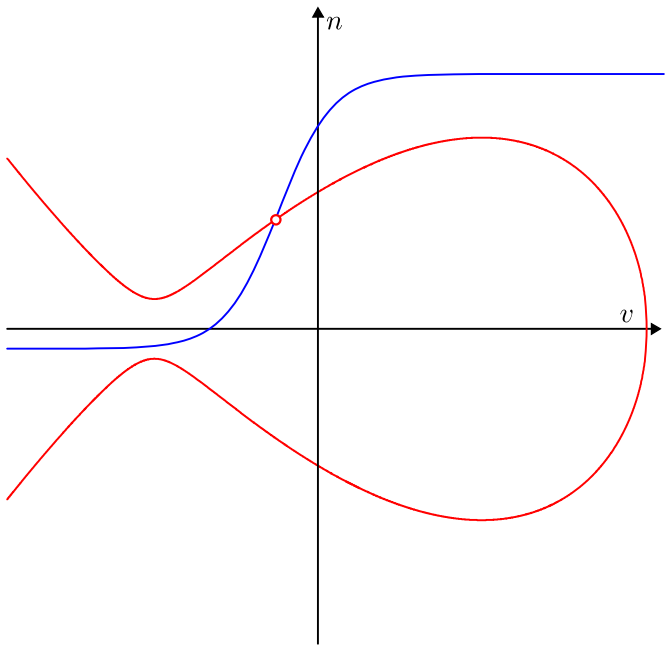


Figure 3.6: v-nullclines and n-nullclines for a system that exhibits a type IV excitability. The red circle represents the unstable equilibrium point.

We start studying the evolution of the *layer problem*

$$\dot{v}(t) = f(v(t), n, I_{app}) = v(t) - \frac{v(t)^3}{3} - n^2 + I_{app}. \tag{3.14}$$

We study the system by a qualitative study of the function $f(v, n, I_{app})$ with n and I_{app} fixed. The partial derivative with respect to v is

$$\frac{\partial f}{\partial v}(v, n, I_{app}) = 1 - v^2,$$

this means that the system has two critical points $v = \pm 1$. And since

$$\frac{\partial^2 f}{\partial v^2}(v, n, I_{app}) = -2v \tag{3.15}$$

then $v = -1$ is a minimum and $v = 1$ is a maximum. Now we want to understand how the function f changes under the translation acted by n. To do that we evaluate the following

quantities:

$$f(-1, n, I_{app}) = I_{app} - \frac{2}{3} - n^2 \tag{3.16}$$

$$f(1, n, I_{app}) = I_{app} + \frac{2}{3} - n^2. \tag{3.17}$$

The study of the sign of the two functions (3.16) and (3.17) permits us to understand how many equilibriums exist in the *layer problem* and what is their nature. So, with the help of figure 3.8, we can distinguish three different cases:

1. $-\sqrt{I_{app} - \frac{2}{3}} < n < \sqrt{I_{app} - \frac{2}{3}}$ the minimum (3.16) and the maximum (3.17) of the function (3.14) are positive, so there is only one stable equilibrium;

2. $n < -\sqrt{I_{app} + \frac{2}{3}}$ and $\sqrt{I_{app} + \frac{2}{3}} < n$ the minimum (3.16) and the maximum (3.17) are negative, so there is only one stable equilibrium;

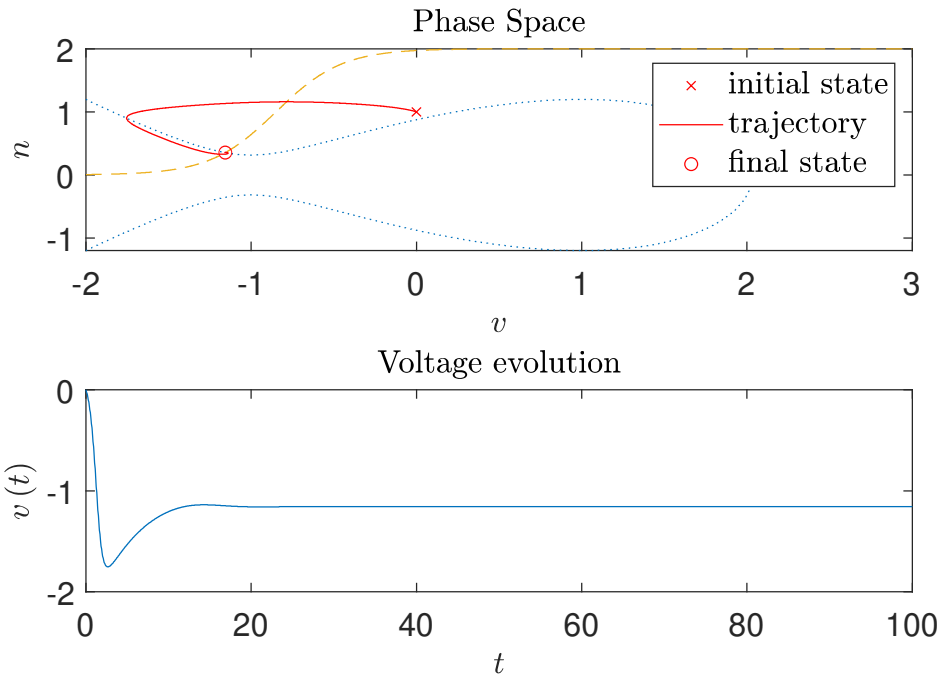


Figure 3.7: Simulation for a trajectory with system (3.10) in a set-up of type III of excitability. In the upper plot we have plotted the orbit in the phase space, in the second plot instead we have plot only the action potential $v(t)$ as function of time.

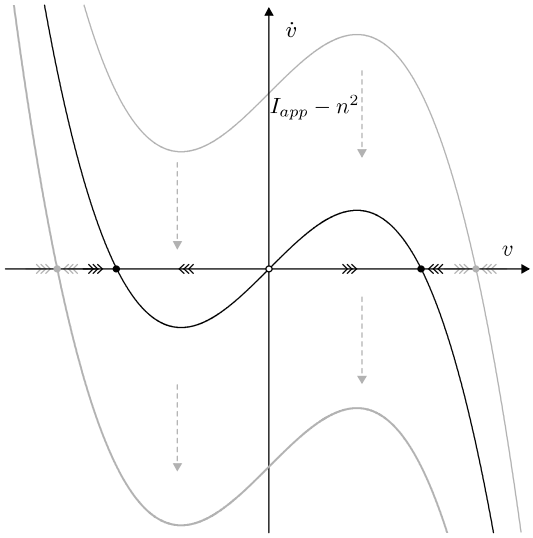


Figure 3.8: Study of the *layer problem* (3.14) under the translation $I_{app} - n^2$ acting by n. In the figure the three possible cases are depicted. On the axis v we plot the stable equilibrium points as filled circles and unstable equilibrium points as circle. The arrow indicates the direction of the motion of the *layer dynamic* (3.14).

3. $-\sqrt{I_{app} + \frac{2}{3}} < n < -\sqrt{I_{app} - \frac{2}{3}}$ or $\sqrt{I_{app} - \frac{2}{3}} < n < \sqrt{I_{app} + \frac{2}{3}}$ for the function (3.14) the minimum (3.16) is negative and the maximum is positive (3.17), so there are three equilibrium points two stable and one unstable.

This implies that in the first and second case the *layer problem* has only one equilibrium point, instead in the third case we have three equilibrium points. In this way we have completely characterized the *layer problem* and the *fast dynamic* of the system except that for the points where $\frac{\partial f}{\partial v} = 0$ and $n = \pm\sqrt{I_{app} \pm \frac{2}{3}}$. These are *fold* points and so, in the following, we can solve this problem using the results of section 2.1.2. But now, we pass to study the *reduced problem*

$$\begin{cases} 0 = v\left(t\right) - \dfrac{v\left(t\right)^3}{3} - n\left(t\right)^2 + I_{app} & \text{(3.18a)} \\ \dot{n} = n_\infty\left(v\left(t\right) - v_0\right) + n_0 - n\left(t\right). & \text{(3.18b)} \end{cases}$$

Equation (3.18a) constrains the system to move in the v-nullcline, and so we can study the property of the flow on this set separately for the upper and the down parts of the v-nullcline. In particular, since the region M_{n_0} is invariant, and in type IV of excitability M_{n_0} contains only the upper branch of the v-nullcline, we study the reduced flow only in this region, that

consists in the pairs $(v(t), n(t))$ that satisfy:

$$n(t) = \sqrt{v(t) - \frac{v(t)^3}{3} + I_{app}}. \tag{3.19}$$

So, denoting

$$p(v(t)) = \sqrt{v(t) - \frac{v(t)^3}{3} + I_{app}}.$$

we can substitute (3.19) into (3.18b) to obtain

$$\begin{aligned} \frac{d}{dt}\left(p(v(t))\right) &= n_\infty(v(t) - v_0) + n_0 - p(v(t)) \\ \frac{1}{2}\frac{1}{p(v(t))}\left(1 - v(t)^2\right)\dot{v}(t) &= n_\infty(v(t) - v_0) + n_0 - p(v(t)) \\ \dot{v}(t) &= 2\frac{n_\infty(v(t) - v_0) + n_0 - p(v(t))}{1 - v(t)^2}p(v(t)) \end{aligned} \tag{3.20}$$

The above equation is not define for $v = \pm 1$. Indeed in these points there are two *fold* points, and we can deduce the behaviour of the trajectories around these points by 2.1.2. But far away of these points the *reduced problem* can be solved by the following sign analysis:

	-1		E	1
$1 - v^2$	$-$	$+$	$+$	$-$
$w(v)$	$-$	$-$	$+$	$+$

Figure 3.9: Sign table for the sign study of the right side of the last equation in (3.20). The study of this function permits to understand the motion due to the *slow* dynamic on the critical manifold.

with

$$w(v)\, n_\infty(v - v_0) + n_0 - p(v).$$

The second line of the sign table derives directly from the fact that this is a type IV of excitability and $n_\infty(v(t) - v_0) + n_0 - p(v)$ is exactly the subtraction between the two nullclines. So also using 3.10 the analysis is complete. Table 3.9 explains the arrow along the critical manifold M_0 in the 3.10. Finally, the qualitative description of the trajectories of the system is completed by resorting to what we have derived about the *fold* points in 2.1.2. At this point 3.11 combines the previous analysis of the *fast* and the *slow* dynamic to give a full description

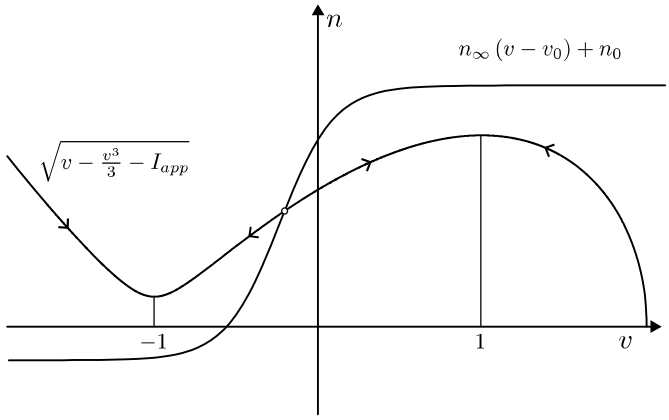


Figure 3.10: Study of the *reduced problem* (3.18) constrained to move in the critical manifold M_0. In the figure the two nullclines are depicted for the type IV of excitability and the unique equilibrium point. The arrow along the critical manifold represents the direction of the trajectories of the *reduced problem*. In the points $v = \pm 1$ of the critical manifold we have two *fold* points.

of the trajectories of the system. So it is possible to see how with this type of excitability we

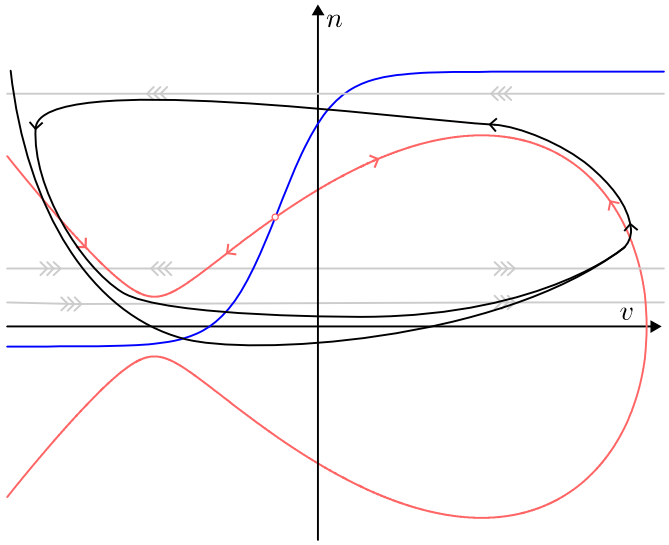


Figure 3.11: Qualitative behaviour of a trajectory for a system in type IV of excitability. The approximations of the *fast* and *slow* dynamics are plotted in grey or in red with triple or single arrows.

can induce a *limit cycle* behaviour. This fact is also proved by the following simulation see

figure 3.12.

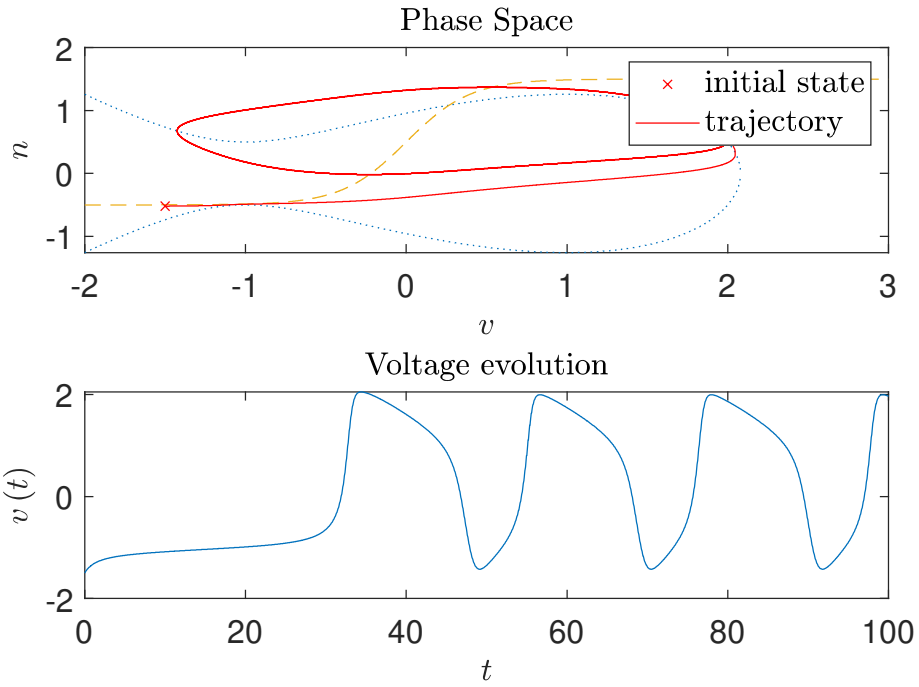


Figure 3.12: The simulation for a trajectory of system (3.10) in a set-up of type IV of excitability. In the upper plot we have plotted the orbit in the phase space. In the second plot, instead, we have plotted only the action potential $v\,(t)$ as function of time.

4

Conclusions

In this chapter, we present some conclusions about our analysis. In chapter 3, we have reported some of the models present in literature to model the behaviour of the neuron. In the last section 3.3 we have also presented a qualitative analysis of two types of excitability that can arise dealing with neuron models. One is linked to the presence of oscillator phenomenons that can be studied by resorting to some of the techniques presented in this thesis, such as the *geometric singular perturbation theory* presented in 2.1.2 or the p-dominance presented in section 2.3.

The other important type of excitability that is presented in the previous section is type III. In the following analysis, we try to linearise system (3.10) in a type III configuration, applying what we have seen for the *Koopman operator* theory presented in 2.4, and we try to study the *non-normality* of this system. As done in section 2.4.4, in particular using the same procedure of example 8, we implement the EDMD procedure. We used a dataset, $\{(x_m, y_m)\}_{m=1}^{M}$ of samples from the discrete version, as done in (2.81), of the Generalized FitzHugh-Nagumo model (3.10). The data are collected from 100 different simulations starting from different initial conditions generated at random from a uniform distribution with $v(0), n(0) \in (-0.1, 0.9)$. Every simulation lasts 200 samples and so the total number of samples is $M = 2000$. For the basis functions, as done in example 8, we used the cross product of a sequence of *Hermite polynomials*. We repeated the approximation process for a different number of basis functions, that go from $N = 25$ to $N = 200$. In figure 4.1 are plotted some true and approximated trajectories of model (3.10). From a visual analysis it is interesting to see how

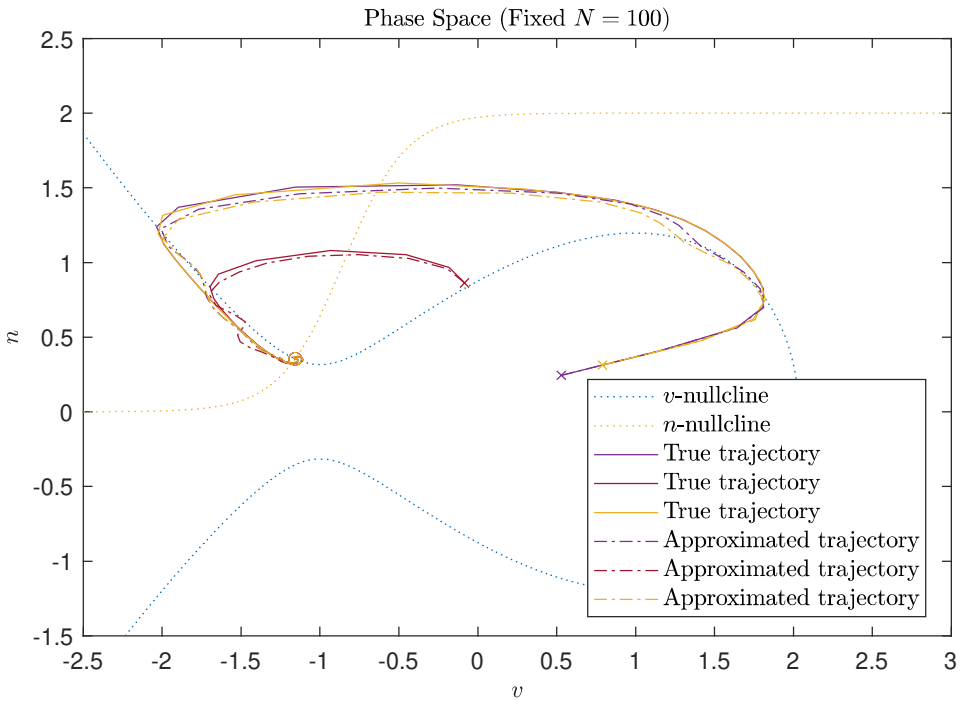


Figure 4.1: This figure represents the phase portrait of the model (3.10). In the same figure we have plotted the true trajectories of the system and the one approximated by the *Koopman operator* setting $N = 100$.

the *linear propagation* of the dynamic through the *Koopman operator* seems to work pretty well in approximating a *non-linear* dynamics as in model (3.10). All the figures presented in this chapter and all the comparisons between the true trajectory and the approximated are done on a new dataset different from $v\,(0)\,,\ n\,(0) \in (-0.1, 0.9)$.

In the simulation we have encountered a probably *overfitting* problem with increasing the value N. For that reason, the simulations with $N > 100$ are removed. In figure 4.2 instead, it is possible to compare the trajectories $v\,(k)$ and $n\,(k)$ as a function of the discrete-time k. The approximations seem to be good and the *Koopman operator* seems to capture the behaviour of the system. Finally, figure 4.3 represents a visual representation of the *Koopman* matrix $\mathbf{U}$ for different values of the dimension N. This plot is interesting because from that it is possible to understand the structure of the matrix $\mathbf{U}$. Indeed, $\mathbf{U}$ seems to be almost sparse with only a few values significantly different from zero. Finally, we try to understand the *non-normality* by the ϵ-pseudospectrum. This is made by plotting, in figure 4.4 the ϵ-pseudospectrum associated with every *Koopman matrix* $\mathbf{U}$. These seem to represent matrices far away from the *normality*, and seems that by increasing dimension N, the shapes of the

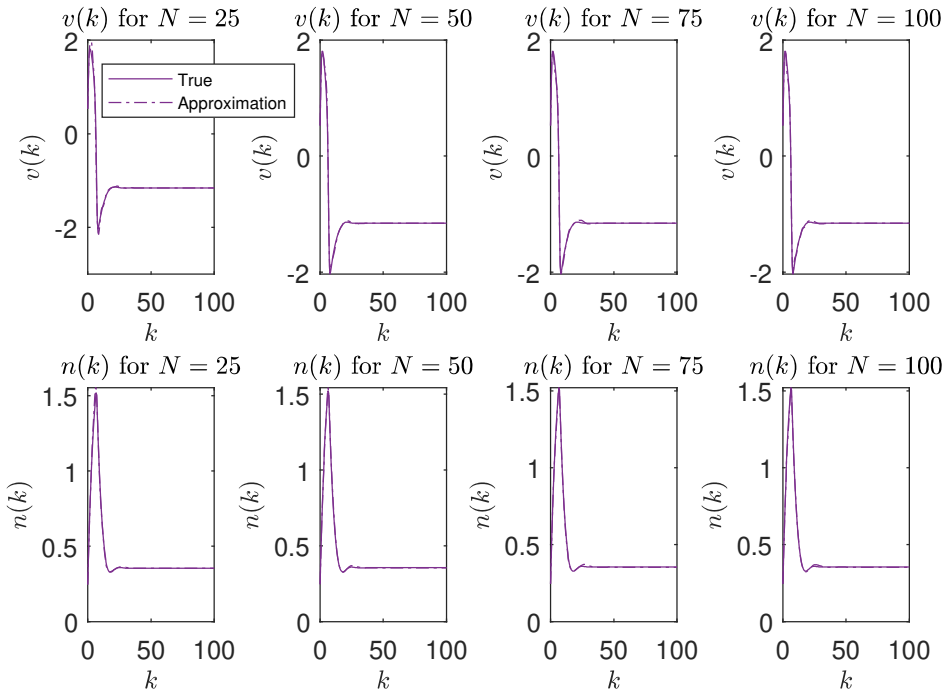


Figure 4.2: Here we plot the time evolution of the two quantities $v\,(k)$ and $n\,(k)$ for different values of N and the same trajectory. In this figure it possible see how the *Koopman* approximation perform with the increasing of the number of basis functions.

ϵ-pseudospectrum go away from the *normal* "situation". We denote $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{U}_3$ and $\mathbf{U}_4$ the *Koopman matrix* approximations with, respectively $N = 25,\ 50,\ 75,\ 100$. We conclude the chapter measuring, using the metric defined in (2.121), the *non-normality* of these matrices.

matr	$\mathrm{dep}_F\,(\cdot)$
$\mathbf{U}_1$	9.84×10^3
$\mathbf{U}_2$	6.20×10^4
$\mathbf{U}_3$	1.65×10^5
$\mathbf{U}_4$	1.78×10^5

(4.1)

From section 2.5.1 we know if $\mathrm{dep}_F\,(\cdot) = 0$, than the considered matrix is *normal*. From table (4.1) we can see that these matrices have big values of $\mathrm{dep}_F\,(\cdot)$. And this seem to confirm our intuition that a excitable system should have high degree of *non-normality*.

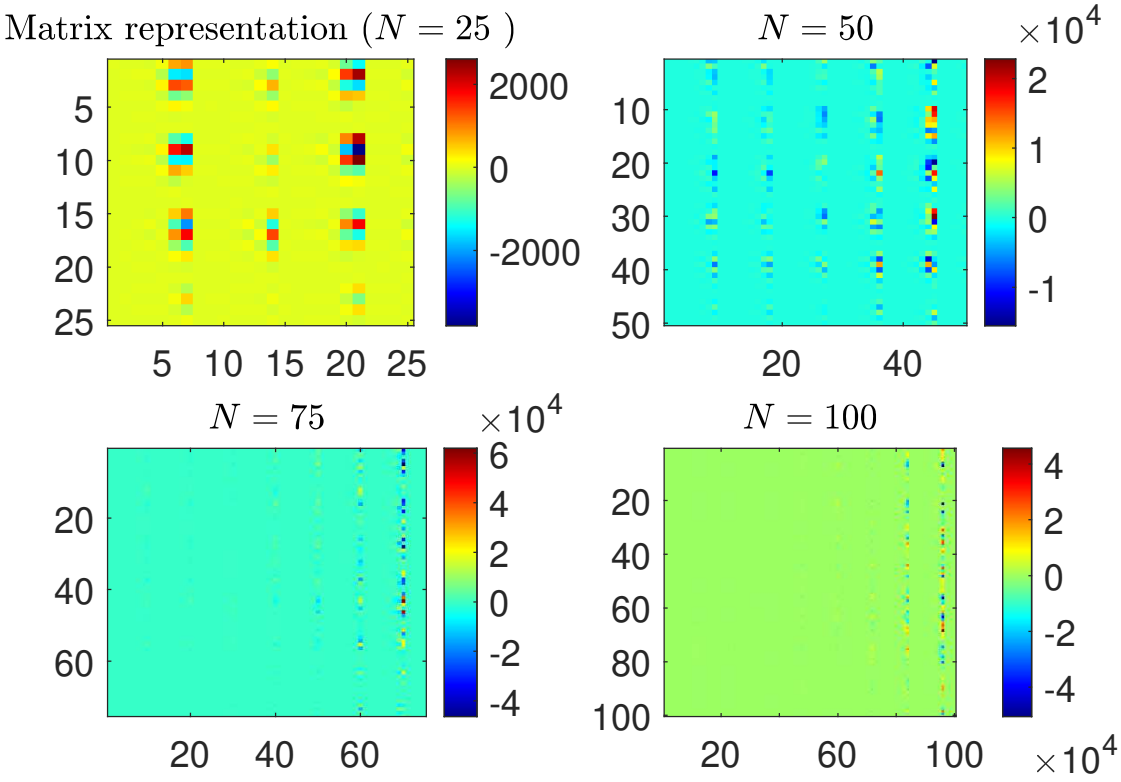


Figure 4.3: This figure represent a visual picture of the *Koopman matrix* for different vlaues of N. It is useful to understand the structure of the *Koopman matrix* and from that is possible to see how the matrix is substantially sparse.

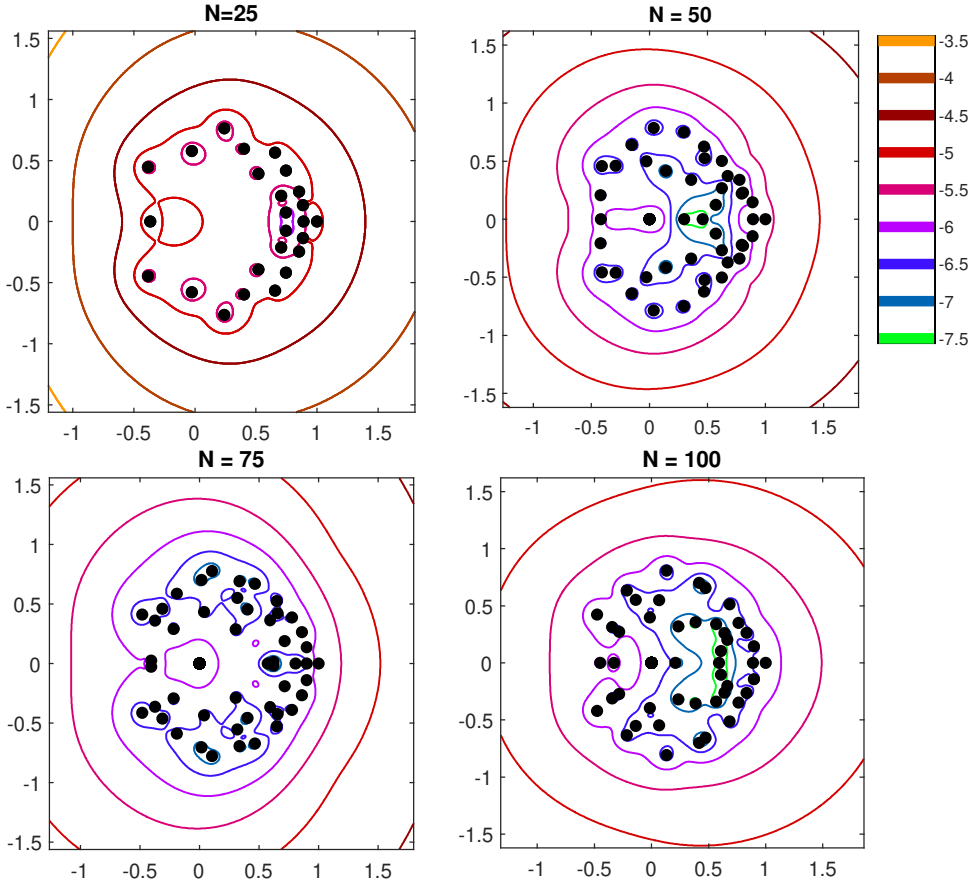


Figure 4.4: In this figure we plot the ϵ-pseudospectrum of the matrices $\mathbf{U}$ for different values of N. The black points are the eigenvalues and the coloured lines are the level curve associated to the ϵ-pseudospectrum. It is interesting to see how the ϵ-pseudospectrum seems become "larger" with the increasing of N.

References

[1] G. Drion, T. O'Leary, J. Dethier, A. Franci, and R. Sepulchre, "Neuronal behaviors: A control perspective," in *2015 54th IEEE Conference on Decision and Control (CDC)*, 2015, pp. 1923–1944.

[2] R. Sepulchre, "Spiking control systems," *Proceedings of the IEEE*, pp. 1–13, 2022.

[3] A. Hodgkin and A. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *Journal of Physiology*, vol. 117, pp. 500–544, 1952.

[4] A. Franci, G. Drion, and R. Sepulchre, "An organizing center in a planar model of neuronal excitability," *SIAM Journal on Applied Dynamical Systems*, vol. 11, no. 4, pp. 1698–1722, 2012. [Online]. Available: https://doi.org/10.1137/120875016

[5] C. K. R. T. Jones, *Geometric singular perturbation theory*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995, pp. 44–118. [Online]. Available: https://doi.org/10.1007/BFb0095239

[6] S. Zampieri, *Sistemi Dissipativi*, 03 2012.

[7] F. Forni and R. Sepulchre, "Differential dissipativity theory for dominance analysis," *IEEE Transactions on Automatic Control*, vol. 64, no. 6, pp. 2340–2351, 2019.

[8] B. O. Koopman, "Hamiltonian systems and transformation in hilbert space," *Proceedings of the National Academy of Sciences*, vol. 17, no. 5, pp. 315–318, 1931. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.17.5.315

[9] S. Wiggins, *Normally Hyperolic Invariant Manifolds in Dynamical Systems*, 01 1994, vol. 105.

[10] M. Krupa and P. Szmolyan, "Extending geometric singular perturbation theory to nonhyperbolic points—fold and canard points in two dimensions," *Society for Industrial and Applied Mathematics*, vol. 33, pp. 286–314, 09 2001.

[11] C. Chicone, *Ordinary Differential Equations with Applications*, 01 2006, vol. 34.

[12] G. T. Gilbert, "Positive definite matrices and sylvester's criterion," *The American Mathematical Monthly*, vol. 98, no. 1, pp. 44–46, 1991. [Online]. Available: http://www.jstor.org/stable/2324036

[13] E. Fornasini and G. Marchesini, "Appunti di teoria dei sistemi," 1992.

[14] S. Newhouse, M. Brin, B. Hasselblatt, and Y. Pesin, "Cone-fields, domination, and hyperbolicity, university in modern dynamical systems and applications," *pp*, pp. 419–433, 01 2004.

[15] *The Poincaré-Bendixson Theorem*. New York, NY: Springer New York, 2003, pp. 117–121. [Online]. Available: https://doi.org/10.1007/0-387-21749-5_10

[16] M. Korda and I. Mezić, "On convergence of extended dynamic mode decomposition to the koopman operator," *Journal of Nonlinear Science*, vol. 28, pp. 687–710, 2018.

[17] I. Mezic, "On numerical approximations of the koopman operator," 2020. [Online]. Available: https://arxiv.org/abs/2009.05883

[18] M. Williams, I. Kevrekidis, and C. Rowley, "A data-driven approximation of the koopman operator: Extending dynamic mode decomposition," *Journal of Nonlinear Science*, vol. 25, 08 2014.

[19] A. Mauroy, I. Mezić, and Y. Susuki, *The Koopman Operator in Systems and Control: Concepts, Methodologies, and Applications*, ser. Lecture Notes in Control and Information Sciences. Springer International Publishing, 2020. [Online]. Available: https://books.google.it/books?id=YUrSDwAAQBAJ

[20] I. Mezic, "Spectral properties of dynamical systems, model reduction and decompositions," *Nonlinear Dynamics*, vol. 41, pp. 309–325, 08 2005.

[21] I. Mezić and A. Banaszuk, "Comparison of systems with complex behavior," *Physica D: Nonlinear Phenomena*, vol. 197, no. 1, pp. 101–133, 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167278904002507

[22] H. Arbabi and I. Mezić, "Ergodic theory, dynamic mode decomposition, and computation of spectral properties of the koopman operator," *SIAM J. Appl. Dyn. Syst.*, vol. 16, pp. 2096–2126, 2017.

[23] L. Trefethen and M. Embree, *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*, 01 2005.

[24] Wikipedia contributors, "Spectral radius — Wikipedia, the free encyclopedia," 2022, [Online; accessed 9-August-2022]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Spectral_radius&oldid=1101685510

[25] E. Hille and R. Phillips, *Functional Analysis and Semi-groups*, ser. American Mathematical Society: Colloquium publications. American Mathematical Society, 1957, no. v. 31;v. 1957. [Online]. Available: https://books.google.it/books?id=hPpQAAAAMAAJ

[26] R. FitzHugh, "Impulses and physiological states in theoretical models of nerve membrane," *Biophysical Journal*, vol. 1, no. 6, pp. 445–466, 1961. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0006349561869026

[27] J. R. Clay, D. Paydarfar, and D. B. Forger, "A simple modification of the hodgkin and huxley equations explains type 3 excitability in squid giant axons," *Journal of The Royal Society Interface*, vol. 5, no. 29, pp. 1421–1428, 2008. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2008.0166

[28] Y. Gai, B. Doiron, V. Kotak, and J. Rinzel, "Noise-gated encoding of slow inputs by auditory brain stem neurons with a low-threshold k+ current," *Journal of neurophysiology*, vol. 102, pp. 3447–60, 10 2009.

[29] A. L. Hodgkin, "The local electric changes associated with repetitive action in a non-medullated axon," *The Journal of Physiology*, vol. 107, no. 2, pp. 165–181, 1948. [Online]. Available: https://physoc.onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.1948.sp004260

[30] N. E. Hallworth, C. J. Wilson, and M. D. Bevan, "Apamin-sensitive small conductance calcium-activated potassium channels, through their selective coupling to voltage-gated calcium channels, are critical determinants of the precision, pace, and pattern of action potential generation in rat subthalamic n...," *Journal of Neuroscience*, vol. 23, no. 20, pp. 7525–7542, 2003. [Online]. Available: https://www.jneurosci.org/content/23/20/7525

[31] J. Huguenard and D. Prince, "A novel t-type current underlies prolonged ca(2+)-dependent burst firing in gabaergic neurons of rat thalamic reticular nucleus," *Journal of Neuroscience*, vol. 12, no. 10, pp. 3804–3817, 1992. [Online]. Available: https://www.jneurosci.org/content/12/10/3804

[32] S. W. Johnson, V. Seutin, and R. A. North, "Burst firing in dopamine neurons induced by n-methyl-d-aspartate: Role of electrogenic sodium pump," *Science*, vol. 258, no. 5082, pp. 665–667, 1992. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1329209

[33] C. M. Gray and D. A. McCormick, "Chattering cells: Superficial pyramidal neurons contributing to the generation of synchronous oscillations in the visual cortex," *Science*, vol. 274, no. 5284, pp. 109–113, 1996. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.274.5284.109