Classical Topics in Dynamical Systems

Mario Natiello

Jörg Schmeling

Center for Mathematical Sciences, Lund University
Email address: `mario.natiello@math.lth.se`

Center for Mathematical Sciences, Lund University
Email address: `jorg.schmeling@math.lth.se`

2000 *Mathematics Subject Classification.* Primary

Key words and phrases. dynamical systems

To our families, for patiently putting up with us.

Contents

Foreword

The theory of dynamical systems has grown in the last 100 years from being a rather special branch of mathematics (within the realm of ordinary differential equations) into a rich and diverse topic, not only of mathematics but also of other sciences. The theory of dynamical systems is on one hand a field of its own in modern mathematics, but also a universal interdisciplinary method that is extensively used.

The mathematical theory of dynamical systems investigates those general structures which are the basis of evolutionary processes, more specifically time-evolution, i.e., how a general system is modified along time. This kind of processes is met in most scientific fields, from numerical simulations around number theory to the time-evolution of coexisting populations. The aim of the theory of dynamical systems is to give qualitative assertions about such processes.

Dynamical systems theory helps substantially to understand the applicability or limitations of mathematical models. Because evolutionary processes are met in so many seemingly different fields, their general nature is very complex and rich. General results arising from the mathematical structure of a problem are often subject to special restrictions depending on the nature of the system in consideration. The conditions under which one can ensure (æstethically) interesting properties are not always verifiable in the different praxes. Scientists in this field have the double task of developing the mathematical theory and understanding its relation with the specific features of the various applications.

This book considers classical topics of dynamical systems (nonlinear ordinary differential equations and difference equations), giving as well an overview of methods and results that are of more general nature. We will especially emphasise on the universality of this theory and try to indicate its wide applicability.

Organisation of the Book

The Book is structured as follows. Some general results that are frequently required in the theory of Dynamical Systems are listed in the Appendix. These include basic definitions such as those of topology, metric space, manifold, etc., and also some basic theorems such as Jordan Canonical Form Theorem or Banach's Fixed Point Theorem. Some of the results are just stated, others are also proved.

Chapter 1 introduces the basic ideas of dynamical systems, highlighting the discrete–time and continuous–time descriptions. Chapters 2 and 3 discuss the basic and subsequently some more advanced ideas about the solutions of systems of first–order ordinary differential equations (i.e., continuous–time systems)). Chapter 4 introduces the treatment of discrete–time systems. Chapter 5 advances in the qualitative theory of dynamical systems through the concept of stability, namely when a given singular solution can be expected to persist asymptotically. Chapter 6 deals with a cornerstone of the qualitative analysis of dynamical systems, the Center Manifold Theorem, preparing the field for one of the most important tools in applications, namely Local Bifurcation Theory, which is discussed in Chapter 7. Some mathematical tools related to the popular concepts of "chaos", "strange attractors", etc., is presented in Chapter 8. Finally, Chapter 9 deals with a beloved classical topic of Dynamical Systems Theory, namely Ergodic Theory.

Concerning references, we have chosen to rely on well-known textbooks in dynamical systems rather than presenting the original work, in cases where the textbook addresses the original sources.

Development

This book originated in Jörg's lecture notes for a course in the Faculty of Engineering at Lund University (Sweden). Mario revised the text and added material from his lecture notes. A bit of mutual revision followed over the years. As it is frequently the case, it took much longer to revise the book than to produce the first draft.

Target audience

The book is designed as a textbook for Engineering- and Science-students at Lund University (Sweden). Part of its added-value is that it is self-contained, so the reader may profit of the text already after the elementary algebra and analysis courses, covering all the way to basic and applied research level in the subject. Indeed, almost all results which are not *elementary* (with "elementary" we mean results discussed in the basic Algebra and Analysis (& other) courses available at essentially any Science- or Engineering faculty) have been stated and proved right here (at least a proof-design is given for a few cases). Hence, everything is here and the reader does not need to rush and see other books in order to complete the ideas we presented. However, we provide a list of references with books (or articles) that have inspired us, one way or the other, such as [**GH86, KH96, SNM96, Wig90**], or where some of the Exercises are discussed.

Acknowledgements

The authors thank the colleagues at the Centre for Mathematical Sciences, Lund University for having been able to discuss the contents of the book and other interesting subjects.

Lund, MMXXI.

CHAPTER 1

Basics

1.1. What is a Dynamical System?

The objects of dynamical systems theory are **space**, **time** and **evolution**.

1.1.1. Space. The mathematical description of a dynamical system is specified by giving the set of (potentially) observable states of the system. Each such state which characterises and identifies a system is regarded as a point in a set X called **phase space**. The points are considered as different states of the system.

Thus, phase space may consist of the set of all possible position and momentum values for a point mass in rectilinear movement, or of the intensity, phase and electric field polarisation of a laser device, the (vertical and horizontal) velocity amplitudes and temperature profile of a gas or the set of all possible numbers of predator and prey in a closed ecological environment.

In most applications, phase space has some additional structure, other than consisting simply of points. For example, it is quite natural to consider the distance between points to be relevant. Depending on the specific problem, phase space can be a *topological space*, a *measurable space*, a *manifold*, or some other structure [1]. E.g., in Hamiltonian dynamics the basic structure is a *symplectic*, even-dimensional manifold, while for the other examples above a proper space could be $\mathbb{R}^2 \times \mathbb{S}^1$, $\mathbb{R}^3$ and $\mathbb{N}^2$ (pairs of non-negative integers) respectively. See Appendix A.1 for a more precise description of the basic concepts.

Throughout this book we will assume that our space is not "too large". In particular we assume that phase space is **finite dimensional** and in many situations also **compact** (the most frequent occurrence of a non-compact phase space in the book will $\mathbb{R}^n$). Such assumptions are natural for most applications: when possible we focus on some limited features of the system (e.g., the number of predators/prey), and the possible extension of these variables is most frequently finite (e.g., the amount of prey may go down to zero, we call that an *extinction*, but it can never grow without bounds: after some "large" value both our models and the supporting physical system would collapse). These assumptions substantially help to develop a rich theory. Unfortunately

[1] A manifold is a space where each point has a neighbourhood equivalent to euclidean space. See Chapter 6.

they do not allow to apply these methods directly to the theory of *partial differential equations* where the phase space is an infinite dimensional function space.

1.1.2. Time. Time is the human means of characterising change. The goal in science is to characterise change away from any subjective opinion as much as possible, i.e., the scientist measures the variables defining phase space, and change is recognised because these variables have different values in different measuring situations. The concept of time used in dynamical systems theory is more often than not the natural time assumed in everyday life, i.e., the Galilean time given by ordinary accurate clocks.

Time is modeled by an additive group (or semi–group) G and the law of evolution will be specified. The most important groups (semi–groups) we consider will be $\mathbb{R}$, $\mathbb{R}^+$ (continuous time) or $\mathbb{Z}$, $\mathbb{N}$ (discrete time). There is a fundamental difference between continuous and discrete time which we will discuss in the next Subsection. Although all natural processes come to an end sooner or later, it is very practical to assume that the group is **non-compact**, i.e., that time may extend towards infinity, since this allows to define and study the notion of *asymptotic behaviour.*

1.1.3. Evolution. [**Arn89**] The evolution law is meant to assign to a state x the new state x_t after time $t \in G$ has elapsed. For the case of discrete time, t is computed in integer time-units. A time-unit is the minimal time-interval that is regarded as relevant for the problem. For example, in population problems it is customary to measure time in "generations". While for bacteria a new generation may be produced in a few hours, for many birds, large mammals and insects, generation time is usually one year. On the other hand, when the time-unit can be made arbitrarily small, as in the case of rapidly changing processes (electrical, chemical, gravitational, etc., processes), it is practical to describe the dynamics with continuous time. The dynamical properties are specified instantaneously, i.e., time is regarded as a real variable.

We note that under certain assumptions the evolution law does not change in time, i.e., it is not important at what moment of time we apply the evolution law. We call such a system **autonomous**. Formally, we can write

$$x_t = \phi_t(x)$$

where $\phi_t \colon X \to X$ is the evolution within t time-steps. Since there is no evolution in zero time steps we have for $t = 0$ that

$$\phi_0 = \mathrm{Id} \quad \text{i.e.,} \quad \phi_0(x) = x. \tag{1.1}$$

Furthermore, it is intuitive to assume that the evolution after $t + s$ time-steps is the same as going t steps and then continuing further s

additional steps. This can be formalised as

$$\phi_{t+s}(x) = \phi_s\left(\phi_t(x)\right). \tag{1.2}$$

REMARK 1.1. In general, the phase space X has some additional structures. It can be an Euclidean space (or, more generally, a manifold), a probability space, a metric space or others. Our main object of study will be Euclidean spaces. In this case it is natural to ask the evolution law to be consistent with the structure of the underlying phase space, i.e., if $X = \mathbb{R}^n$ then it is natural to ask that the laws ϕ_t are smooth. In general we expect that the evolution law **respects and preserves** the underlying structure.

Let us consider the discrete- and continuous-time cases in detail.

- $G = \mathbb{R}$ $(\mathbb{R}^+)$. In this case the dynamical system is called a **flow** (**semi–flow**). X is finite-dimensional and we are interested in obtaining the smooth function $x(t)$ giving the state of the system instantaneously for all times. We assume that the evolution law depends continuously on the time parameter t. The mathematically simplest case is $X = \mathbb{R}^n$. The instantaneous change in x is given by the evolution law, i.e., some continuous function $f(x, t)$. For autonomous systems the evolution law is given by a function $f(x)$ of the state only. Hence, autonomous, finite-dimensional dynamical systems are described by ordinary differential equations (ODE's) of the general form $x' = f(x)$, where the specific meaning of these symbols is to be given for each particular case. Formally, $\phi_t(x)$ reads as the integral of f over a time-interval t with initial condition x.
- $G = \mathbb{Z}, \mathbb{N}$. Describing the system in terms of its natural time-unit, "time" takes integer values. The evolution law can then be represented by the symbolic expression $x_{n+1} = F(x_n)$, where the function F describes the evolution law when going from generation n to the next generation $(n + 1)$. We call F a *map* or *mapping*. Autonomous evolution means now that F does not depend explicitly on n, but only on the state x. The property (1.2) implies that $\phi_n = F^n$ where $\phi_1 \equiv F$. Namely,
$$\phi_n(x) = \phi_{n-1}(\phi_1(x)) = \phi_1 \circ \phi_1 \circ \cdots \circ \phi_1(x) = \phi_1^n(x).$$
Moreover for $G = \mathbb{Z}$ we have
$$\phi_{-1}(\phi_1(x)) = \phi_1(\phi_{-1}(x)) = x$$
and $\phi_{-1} \equiv F^{-1}$ is the inverse map.

REMARK 1.2. There is a deep relation between the continuous- and discrete-time descriptions, basically arising from the fact that $\mathbb{Z}$ can be regarded as a subset of $\mathbb{R}$. Suppose that within a continuous-time description we choose to summarise the dynamics by making "readings" of the variables at fixed intervals of time. In such a case the flow defined by $f(x)$ induces a discrete-time dynamics with associated $F(x)$, where f and F are related by formal integration.

Remark 1.3. The use of capitals for maps and lower-case for flows is just a matter of notation. It was chosen for simplicity for the sake of the previous remark. In specific problems the notation is the choice of the researcher, the only condition being that the description is clearly, consistently and accurately specified.

1.2. General Features of Dynamical Systems Theory

1.2.1. Asymptotic Behaviour. [**GH86**] The main feature of dynamics distinguishing it from other areas is that one is interested in the asymptotic behaviour for arbitrarily large times. This is relevant because many natural systems when left alone, relating to the rest of the universe under controlled conditions, approach some sort of "steady state" where further change is undetectable. This steady state can be inferred from the mathematical description by taking limits as time "goes to infinity". The specific asymptotic questions depend on the underlying structure.

1.2.2. Stability. [**GH86**] One of the most important properties of a dynamical system is the concept of **stability**. This is a broad concept which we will further specify along the book. Without some sort of stability any model is questionable and numerical simulations are often not possible or improper.

There are two main characterisations of stability. The first one is the *inner stability*, meaning that the asymptotic behaviour does not drastically change if we make a small "error" in the starting position. This is sometimes called the **dependence on initial conditions**. The other notion of stability is called **structural stability**. It means that the behaviour of the entire system is not drastically changed if we slightly perturb (i.e., make a small change) the evolution law.

These facts have many important practical consequences. The first concept of stability allows us to make qualitative assertions about the system. Statements of the form "if the system is initially in *this* region of phase space, it will be found in the vicinity of such-and-such state after a given (sufficiently long) time".

The second concept of stability is relevant since in applications (numerical or theoretical) one always works with models, i.e., representations of the original problem. These models do not exactly describe the underlying processes up to 100%. There are always features that are simplified, disregarded or ignored, no matter how detailed the description. A good model, however, is in some sense sufficiently close to the underlying system. If this model is structurally stable then there is hope that the investigations of this model will lead to almost correct answers for the real system. The model is “not far from” the real system, and consequently the predictions of the model are not far from the real behaviour. In particular, any numerical simulation is a (discrete) dynamical system modeling the original system (and perhaps modeling the original model!) and the theory of convergence of algorithms is actually a stability theory in this sense.

1.2.3. Classification. For practical reasons, the mathematician has an interest in classifying all dynamical systems. The goal is to identify the qualitative properties of the time-evolution just by recognising some features in the mathematical description. Such programme is unfeasible in such generality. However, partial results do exist. One has to restrict the equivalence (classifying) relations or restrict the class of systems under study. Often only a local classification is possible.

1.3. Examples

We give here some examples of the structures we will encounter.

1.3.1. Topological Dynamics. [**SNM96**] Topology is the simplest mathematical structure where we can develop the concept of continuity. Topological properties, hence, deal with features that are persistent under *homeomorphisms* (continuous, invertible, 1–to–1 maps of phase space, e.g., coordinate changes). Hence, to establish such properties no derivatives are necessary.

In topological dynamics the phase space is usually a compact metric space and the evolution law is given by a family of homeomorphisms. This setup establishes the basis for more restrictive situations where other properties are built on top of the topological ones (smoothness, symplectic, algebraic, …). Many notions are defined topologically, i.e., in the continuous category. Some of those are: periodic orbits, recurrence, topological entropy, structural stability, attractors and invariant sets, … For example, it is clear that whether a solution $x(t)$ of a dynamical system defined for $t \in [t_0, t_0 + T]$ closes into itself at the end of the path or not (this is, if $x(t_0)$ equals $x(t_0 + T)$ or not) is a property that cannot depend on the choice of coordinates: it will persist under homeomorphisms. It is remarkable that many systems are stable in the topological but not in the smooth sense.

The analysis of asymptotic behaviour is highly related to topological properties. The simplest asymptotic feature of a system is to determine its *invariant sets*, i.e., the set(s) of points in phase space that are unaltered by the dynamics. In this context, the interesting topological properties are **transitivity and minimality**. These notions are different kinds of irreducibility of systems, i.e., we identify invariant regions of phase space that cannot be decomposed into a number of smaller independent invariant sets.

Asymptotics deals not only with invariant sets but also with the way the dynamics of a given initial condition may approach such set. One of the most important notions is the notion of *attractor*. Roughly speaking, an attractor ”collects” a great amount of trajectories, i.e., many trajectories in phase space get closer and closer to the attractor as time passes by.

Last but not least, topological dynamics has a profound connection to the theory of C^*–algebras. It also can be applied to questions in pure topology. It has a great record in the proof of the Poincaré conjecture in higher dimensions.

1.3.2. Measurable Dynamics and Ergodic Theory. [**Arn89**]

It is much too difficult to study the behaviour of any single orbit in a complex system. A way out is to study the behaviour of typical trajectories. This leads to statistical considerations. We are interested in *a.e.* behaviour with respect to an invariant measure. The expression *a.e.* stands for “almost everywhere” and it indicates properties that hold for all points (or all trajectories) of phase space except at most a “small” set. Also “small” has a specific definition: in this context it means a set of zero measure, describing collections of points that are in some sense “unimportant” from a statistical point of view.

So the setup is that the phase space is a *measure space* (a usual space where we can additionally define a measure, i.e., a way of computing the “size” of sets) and the evolution is measure preserving.

This is the setup of a stationary stochastic process. In fact, the ergodic theory contains the theory of stochastic processes. The difference is that it does not deal with independence, but it has a notion of equivalence of processes which makes the theory different. The questions we are interested in are of stochastic nature. Do we have a **law of large numbers**?, a **central limit theorem**? and so on. Many of these questions are studied and could be answered in an surprisingly general setting.

1.3.3. Smooth Dynamical Systems. In the same way as topology deals with continuity, smoothness is a sub-category where in addition it is possible to compute (infinitely many) derivatives. Here the phase space is now a manifold and the evolution is smooth, i.e., the solutions $x(t)$ are not only continuous functions, but also differentiable functions. The main advantage of the smooth structure is that we can linearise the system locally, by taking derivatives of the equations, i.e., much along the lines of Taylor expansions. In general, the study of linear systems is much simpler than the corresponding study of general systems (also called nonlinear).

Another feature of smooth dynamics is that it unites the methods from topological and measurable dynamics. It is quite striking that these more restrictive settings imply very strong general results in the smooth category. Such phenomena are called **rigidity**.

There are several more special subclasses of smooth systems. These include symplectic systems (celestial or Hamiltonian mechanics), algebraic systems, geodesic flows, and others.

1.3.4. Symbolic Dynamical Systems and Coding Theory. [**Wig90**] There is an important tool which is useful in the investigations of a large class of smooth, topological or measurable dynamical systems. In many cases, the system can be encoded into a symbolic system. For example, suppose we can identify two sets in phase space (which we label with the numbers 0 and 1), such that all trajectories intersect these sets as time goes by. A simplified way to describe the dynamics could be to list the intersections of each trajectory, e.g., to give a sequence $1001101\cdots$ of visits in time. When this is possible without ambiguities for all of phase space, each point can be associated to an (infinite) address out of an alphabet (in the example, a binary alphabet) which is coherent with the dynamics. We may then study the dynamics between intersections as a map in the space of binary strings. We "transfer" the dynamical study to a symbolic space. Under certain conditions, such space will carry the complete information of the original system. In those situations, the study of the system is largely simplified.

On the other hand, the intensive investigations of the symbolic dynamics (dynamics on infinite strings) brought a deep insight into the structure of those strings. This helped to develop the coding theory which deals with long (but finite) strings.

1.4. Exercises

Exercise 1.1. Show that in the case $G = \mathbb{Z}$ the evolution law is defined by $\phi_n = \phi_1^n$ (see Section 1.1).

Exercise 1.2. A linear, real-valued system $\frac{dx}{dt} = x$ has the associated flow $\phi_t(y) = e^t y$, $y \in \mathbb{R}$ in continuous time. Describe the dynamical system associated to the time-one map, i.e., the map connecting each point in phase space to its image after one unit of time has passed, $x_{n+1} = F(x_n)$. Show both the new discrete–time flow and $F(x)$ (see Section 1.1).

CHAPTER 2

Basic Ordinary Differential Equations

We state here the foundations of the theory of dynamical systems for continuous time. We start with some basic structure and further study the question of existence of solutions for systems of ordinary differential equations of the first order. We end the chapter with some remarks about linear ordinary differential equations.

Throughout the text we will use without distinction the notations $\dot{x}$ and $\frac{dx}{dt}$ to denote the time-derivative of a function x.

2.1. Preliminary Considerations

Newton stated the fundamentals of mechanics, the theory of motion. His principle states that the force acting on a particle equals its mass times the acceleration. Since acceleration is the second derivative of the position q of a particle with respect to time, we get

$$m\ddot{q} = F(q)$$

where m is the mass, $q(t)$ is the position of the point at time t and F is the acting force (here assumed to be autonomous, i.e., without explicit dependence on time). This is an example of an ordinary differential equation (ODE).

More generally, ODE's are equations of a function $x(t)$ that involve the time-derivatives $\dot{x}, \ddot{x}, \cdots, t$ and possibly other parameters. The highest derivative involved is called the order of the equation. A special but important case of differential equations occurs when we can solve the equation for the highest derivative. Thus, for an equation of order n of the form $d^n x/dt^n = f(x, \dot{x}, \cdots, t)$ we may introduce new variables

$$x_1 = \dot{x}, x_2 = \ddot{x}, \cdots, x_{n-1} = \frac{d^{(n-1)}x}{dt^{(n-1)}}$$

and recast this equation in vector form, obtaining a system of differential equations of first order:

$$\frac{d}{dt}(x, x_1, \cdots, x_{n-1}) = v(x, x_1, \cdots, x_{n-1}, t)$$

where $v\colon \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ is a vector function called the **vector field**. In case of Newton's equation, we get, with $p = \dot{q}$

$$(\dot{q}, \dot{p}) = \left(p, \frac{1}{m}F(q)\right).$$

The main question in systems of first-order ODE's is to find functions $t \mapsto x(t) \in \mathbb{R}^n$ such that $\dot{x}(t) = v(x(t), t)$. The autonomous case occurs when the vector field v does not depend on time, i.e., $\dot{x}(t) = v(x(t))$. Such a curve $t \to x(t)$ is called an **integral curve or trajectory** of the vector field v (also called an **orbit**). We will see that it is specified by one initial condition, $x(0)$, being the solution of a first-order equation. We can interpret the curve as showing how a state evolves in time.

2.1.1. Finding Differential Equations. Let us illustrate some ways in which differential equations arise in different problems. Assume we are given a family of curves. We want to find a differential equation which has this family as trajectories.

Example 2.1. We consider the family of curves in the plane

$$C_1 x + (y - C_2)^2 = 0.$$

with $y = y(x)$. C_1 and C_2 are two arbitrary constants specifying each member of the family. We differentiate twice and get:

$$C_1 + 2(y - C_2)y' = 0$$
$$2(y')^2 + 2(y - C_2)y'' = 0.$$

Now we use the resulting equations to eliminate the constants. This gives $C_1 = -2(y - C_2)y'$ and hence, $-2xy'(y - C_2) + (y - C_2)^2 = 0$. The second equation yields $y - C_2 = -\frac{(y')^2}{y''}$. This leads to

$$y' + 2xy'' = 0.$$

2.1.2. Simple Examples. We first consider the linear equation

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

with $v(x) = ax$. Here $x \in \mathbb{R}$ and a is a real constant, usually called a *parameter* of the equation.

From the theory of differential equations we know that given an initial condition $x_0 \in \mathbb{R}$, the function

$$x(t) = x_0 e^{at}$$

is a solution. In this way we find solutions to any initial condition. These are the only solutions. For if $y(t)$ is another solution with $y(0) = x(0) = x_0$ then

$$\frac{d}{dt}\left(y(t)e^{-at}\right) = \dot{y}(t)e^{-at} + y(t)\left(-ae^{-at}\right) = ay(t)e^{-at} - ay(t)e^{-at} = 0.$$

Hence $y(t)e^{-at}$ is a constant (it has zero time derivative) and therefore $y(t)e^{-at} \equiv x_0$. Hence, $y(t) = x_0 e^{at}$. Thus, we see that the initial condition uniquely defines the trajectory.

Let us study the nature of this solution in detail, for different values of a. Letting $a > 0$, the trajectories originating in the vicinity of 0 (for

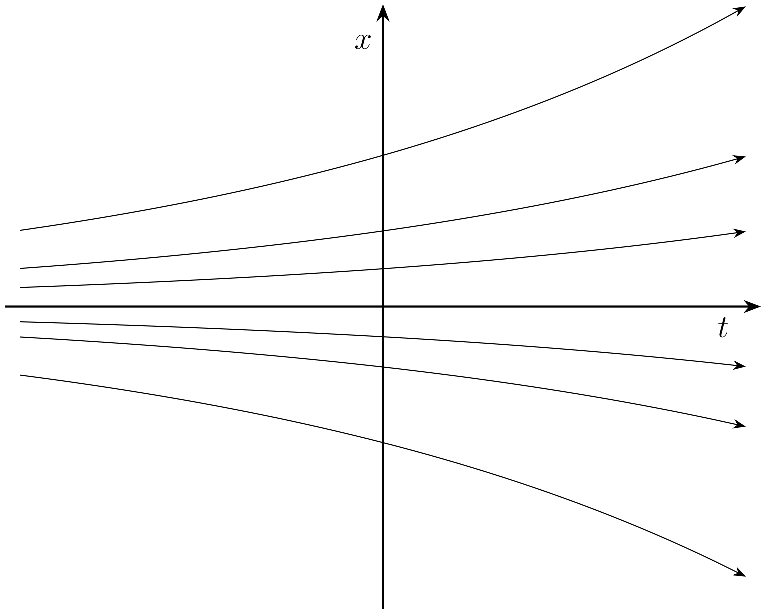


FIGURE 2.1. Solutions of a linear ODE for different initial conditions. Case $a > 0$.

large negative times) tend to infinity for large positive times, except for the solution $x \equiv 0$ (see Figure 2.1). For $a < 0$ the solutions "come from infinity" (for large negative times) and tend to the zero solution for large positive times (see Figure 2.2). These two behaviours are qualitatively different. On the contrary, the behaviour among e.g., all positive a is qualitatively similar. Indeed, defining a new variable $\tau = at$, the behaviour for all positive values of a reduces to $\bar{x}(t) = x_0 e^{\tau}$. A similar property is observed among all negative values of a. Small modifications of a, that do not change its sign, yield essentially the same dynamics, the solution curves coincide exactly after rescaling time.

For $a = 0$ (see Figure 2.3) there is a third behaviour, qualitatively different from the previous two. There is "no" dynamics, $x(t)$ is constant. However, in this case any modification of a away from zero yields qualitatively different solutions. We say that the system is **structurally unstable**: any small change of the parameter a away from zero can lead to one of the two previously classified types, and the three types of solutions cannot be mapped to each other by rescaling the coordinate or time. Note the substantial difference with the cases $a < 0$ or $a > 0$ where the type of solution is preserved under sufficiently small perturbations.

Next we consider the system of equations (vector equation)

$$\begin{aligned} \dot{x}_1(t) &= a_1 x_1(t) \\ \dot{x}_2(t) &= a_2 x_2(t). \end{aligned}$$

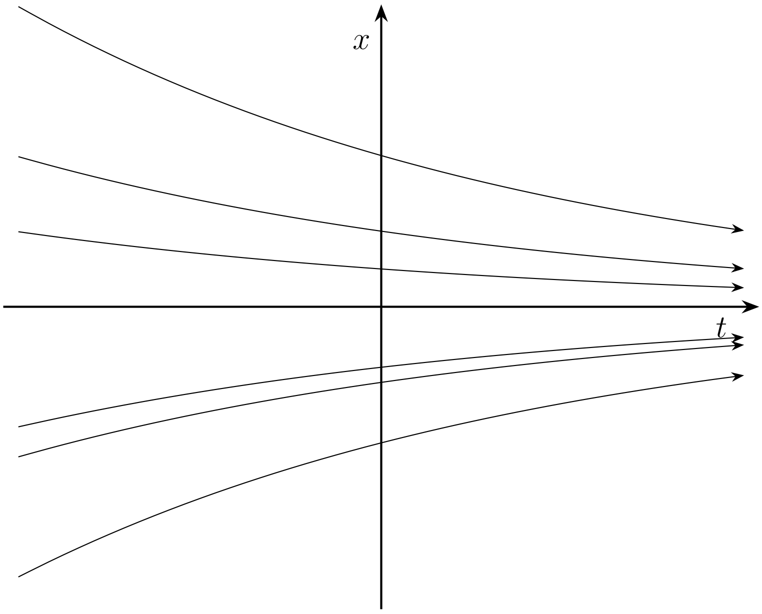


FIGURE 2.2. Solutions of a linear ODE for different initial conditions. Case $a < 0$.

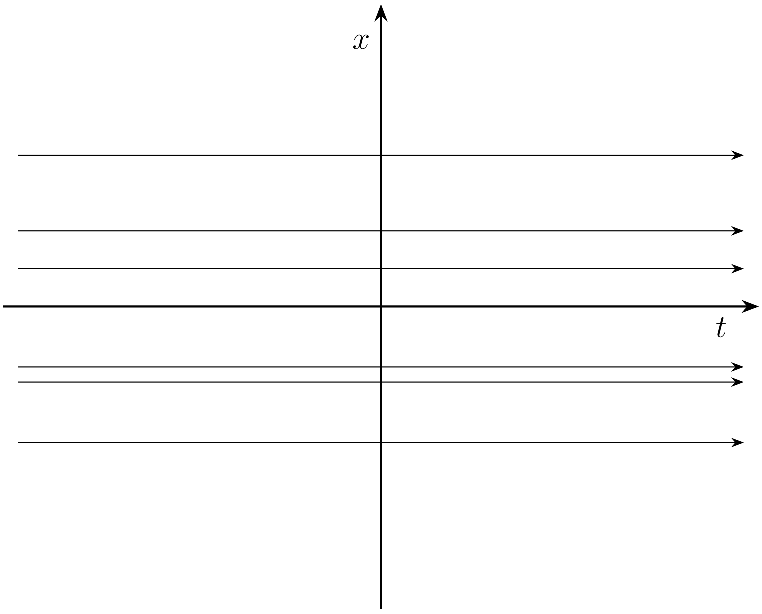


FIGURE 2.3. Solutions of a linear ODE for different initial conditions. Case $a = 0$.

These equations can be recast as a linear equation in $\mathbb{R}^2$, with $v(x) = Ax$, $x = (x_1, x_2) \in \mathbb{R}^2$ and

$$A = \begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix}, \tag{2.1}$$

a diagonal 2×2 matrix.

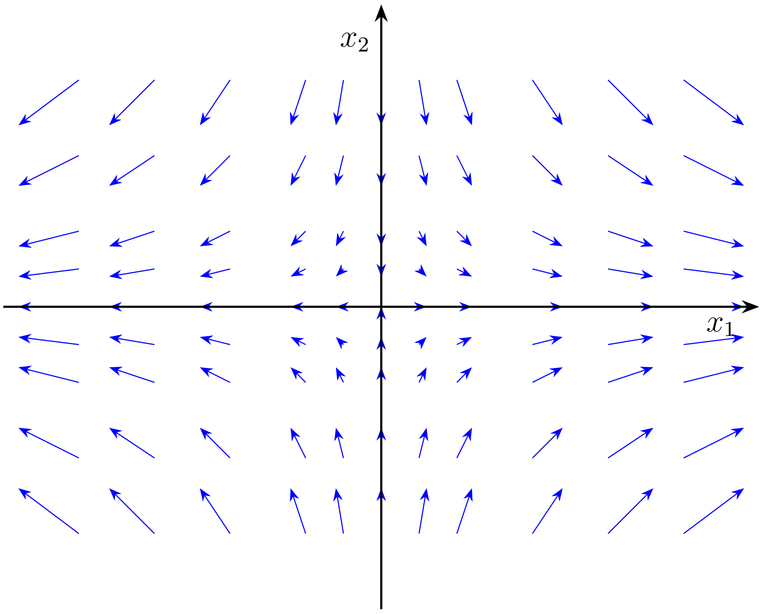


FIGURE 2.4. Graph representing the Vector field for eq. 2.1, with $a_1 = -a_2 = \log 2$.

Since both components are completely separated, the solutions are the juxtaposition of the individual solutions, following the previous example.

$$x_1(t) = x_1(0)e^{a_1 t} \quad x_2(t) = x_2(0)e^{a_2 t}.$$

Again the solutions exist and are uniquely determined by their initial conditions $(x_1(0), x_2(0))$.

Graphical tools to describe dynamical system are widely available in modern times, especially for 2-d systems, where relevant portions of phase space can be thoroughly depicted on a piece of paper. For example, the vector field of a 2-d system can be depicted by considering a grid covering phase space and assigning to each point on the grid, a small arrow representing the direction and length of the vector field. See Figure 2.4 for an example showing the linear vector field.

A useful tool in understanding dynamical systems is the **phase portrait**, where some typical trajectories are illustrated. If the vector field is sufficiently smooth, trajectories are also smooth curves. See Figure 2.5. Since trajectories are tangent to the vector field, it is easy to recognise the relation between both graphs in 2-d systems.

REMARK 2.1. The graphs of the trajectories, $(x_1(t), x_2(t))$, lie in $\mathbb{R}^3$ while the phase portrait lies in $\mathbb{R}^2$.

REMARK 2.2. This system of equations can be regarded as a **dynamical system**. The phase space is $\mathbb{R}^2$ and the time variable lies in $\mathbb{R}$. Let $x(0) = (x_1(0), x_2(0)) \in \mathbb{R}^2$ be the state at time $t = 0$. Then we

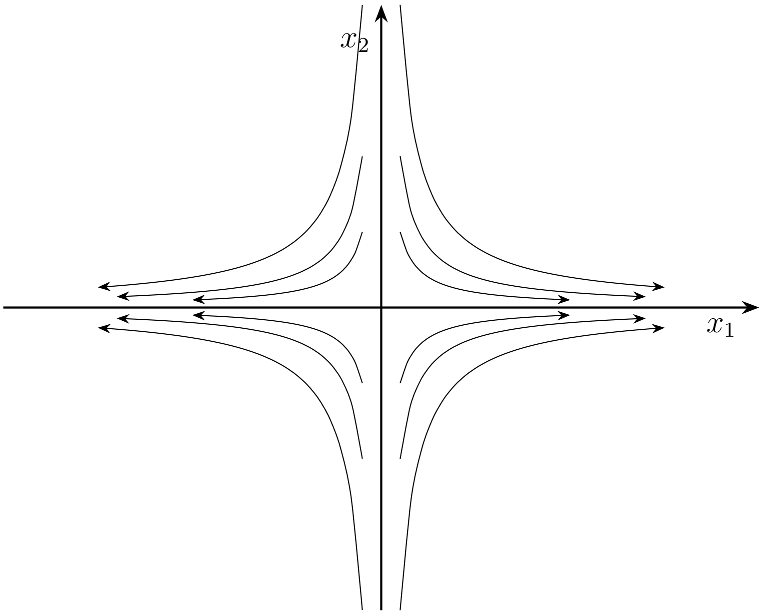


FIGURE 2.5. Phase portrait for eq. 2.1, with $a_1 = -a_2 = \log 2$.

have the unique solution $x(t) \equiv x(t, x(0))$. We can set

$$\phi_t(x_1(0), x_2(0)) := (x_1(t, x_1(0)), x_2(t, x_2(0)))\colon \mathbb{R}^2 \to \mathbb{R}^2.$$

This evolution law is called a **flow**. It is straightforward to check that conditions (1.1) and (1.2) are satisfied.

REMARK 2.3. For $A = \begin{pmatrix} \log 2 & 0 \\ 0 & -\log 2 \end{pmatrix}$ the solutions with non-zero first coordinate, i.e., $x_1(0) \neq 0$ are unbounded in the future (positive times). Also, all solutions with non-zero second coordinate, i.e., $x_2(0) \neq 0$, are unbounded in negative time. We have exactly one bounded solution $x(t) \equiv 0$ for the initial condition $x_1(0) = x_2(0) = 0$.

If we consider the linear system with matrix $-A$ the solution unbounded in the future and in the past are exchanged, as compared with the previous system and vice versa. Here we also have a qualitatively different behaviour.

REMARK 2.4. For the vector field $A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ we have that all solutions are of the form

$$(x_1(t, x_1(0)), x_2(t, x_2(0))) = (x_1(0)\cos t - x_2(0)\sin t, x_1(0)\sin t + x_2(0)\cos t)$$

and hence are periodic (see Figure 2.6).

The periodic case is similar to the case $a = 0$ in dimension one, since solutions never become unbounded for large (positive or negative) times. Note however that in all examples so far, the solutions stay finite after a finite amount of time has passed. This is not always the case.

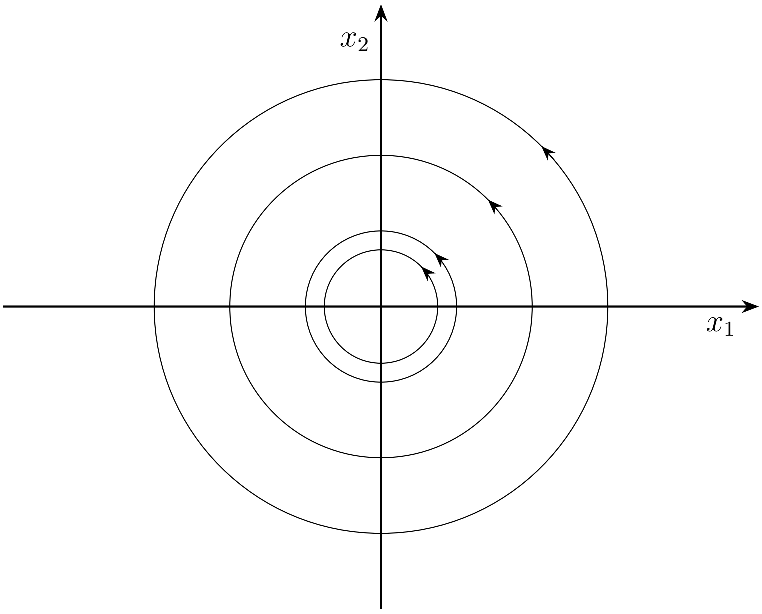


FIGURE 2.6. Linear ODE with periodic solutions.

EXAMPLE 2.2. Let $v(x) = 1 + x^2$. The integral curves have the form $x(t) = \tan(t + \tau)$ for $x(0) = \tan \tau$. These solutions escape to infinity after time $\pi/2 - \tau$ (see Figure 2.7).

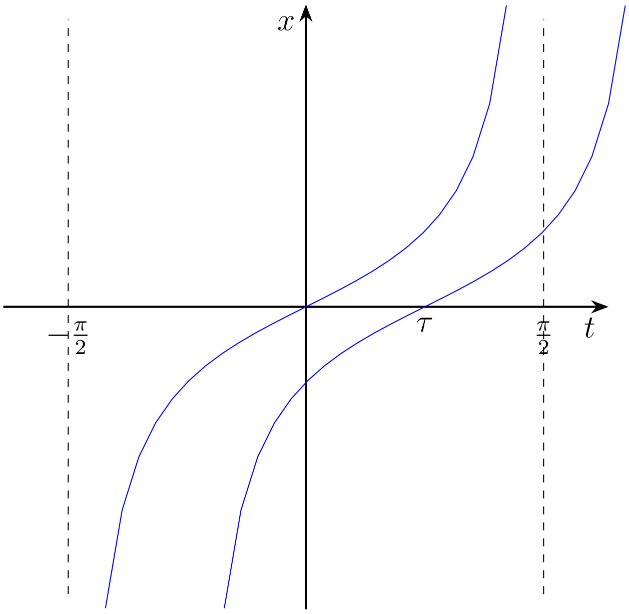


FIGURE 2.7. Unbounded solutions diverging in finite time.

REMARK 2.5. If A is not a diagonal matrix we can try to diagonalise it by a linear transformation $y = S^{-1}x$. If we succeed, this implies

$x = Sy$ and $\dot{x} = S\dot{y}$. Hence, $S\dot{y} = ASy$ or $\dot{y} = S^{-1}ASy$. Since $S^{-1}AS$ is diagonal, the phase portrait, solutions, etc., in the new coordinate system y, are just that of a diagonal matrix (see Figure 2.8). We end up with a set of uncoupled differential equations in one dimension.

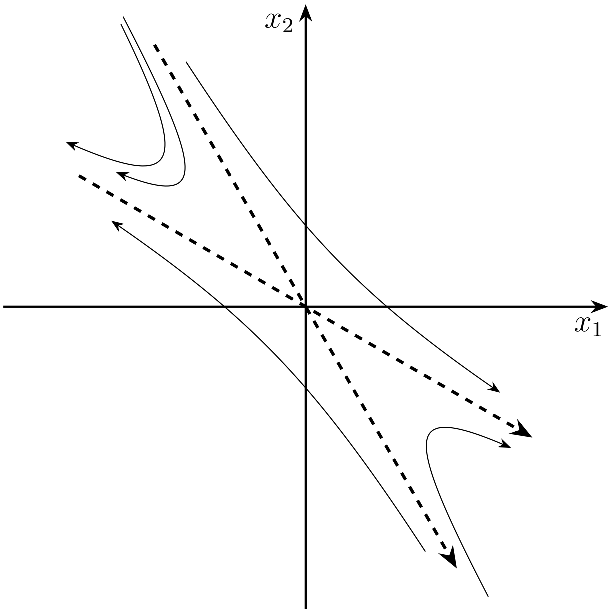


FIGURE 2.8. The new axes y are given by the dotted lines.

REMARK 2.6. In our examples the solutions were uniquely determined by their initial values. Through any point of phase space $X = \mathbb{R}^2$, there is a unique trajectory defining the evolution. This is a reasonable and necessary requirement to have a dynamical system. We will later establish conditions assuring this property.

EXAMPLE 2.3. Let $v(x) = 3x^{2/3}$. This system has the solution $x_0(t) \equiv 0$. However, for any $\tau \in \mathbb{R}$, $x(t) = (t - \tau)^3$ is also a solution, satisfying $x(\tau) = 0$ (see Figure 2.9). Hence, there is more than one trajectory passing through $x = 0$. Uniqueness of solutions is guaranteed whenever $v(x)$ satisfies the *Lipschitz condition*:

DEFINITION 2.1. Let U be an open, non-empty ball in $\mathbb{R}^n$. We say that the function v defined on U is *Lipschitz* (or that it satisfies the *Lipschitz condition*) on U whenever there exists a positive constant L such that

$$|v(x_1) - v(x_2)| \leq L|x_1 - x_2|.$$

L is called the Lipschitz constant of v (on U).

In the previous example, v fails to be Lipschitz at $x = 0$.

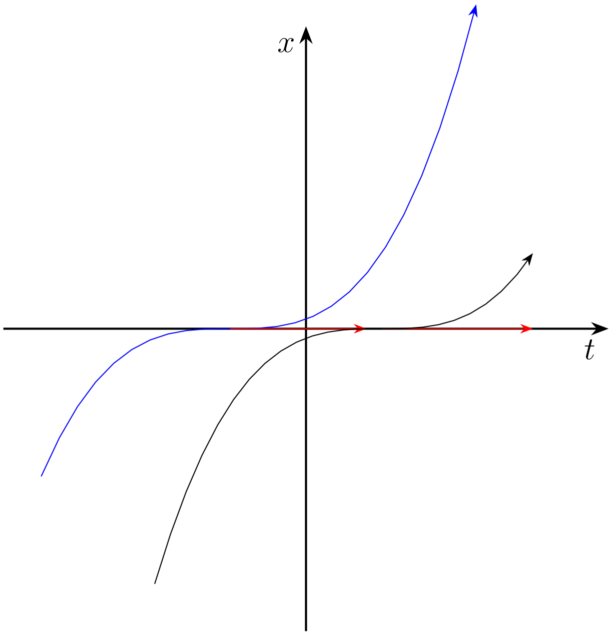


FIGURE 2.9. When Lipschitz condition is not fulfilled, we may have non-unique solutions.

2.2. Linear and Nonlinear

For the sake of dynamical systems theory, the autonomous linear case is usually addressed in earlier, more elementary, mathematical courses. The general problem is of the type $x \in \mathbb{R}^n$, A a real, $n \times n$, square matrix and

$$\dot{x} = Ax.$$

The case $n = 1$ is discussed in the first Calculus courses, and has the general solution $x(t) = e^{At}x_0$, as stated above. For arbitrary n there exists a general procedure that, in principle, exhausts all details of the problem.

1. Perform a linear change of coordinates rendering the matrix in its simplest form (call it B), this is called the **Jordan canonical form** [**Hal69**] of matrix A. In the simplest situation, B will be a diagonal matrix, but diagonalisation is not always possible for a general matrix A, see Appendix A.4 for the general case.
2. In the new coordinates y the solution has the general form
$$y(t) = e^{Bt}y(0).$$
3. The computation of the matrix e^{Bt} falls in one of two cases:
   - If B is diagonal (i.e., all non–diagonal elements are zero), then e^{Bt} is a diagonal matrix, with entries $e^{b_i t}$, b_i the corresponding diagonal element of B. The problem reduces to n independent copies of the situation for $n = 1$. Being A a real matrix, the b_i's may occur in complex-conjugate pairs.

Their treatment is formally the same as with real eigenvalues, only that complex exponentials occur, and pairs x_i, x_j of the original coordinate system will be coupled in the final solution together with trigonometric functions.

- If B is not diagonal, i.e., A has a non–trivial Jordan canonical form, then e^{Bt} will have, apart from (possibly complex) exponential factors, polynomials in t.

The nicest feature of linear systems is that the solutions are defined for all times, and for all initial conditions and moreover, that apart from the more or less hard work required to compute the Jordan canonical form, the solutions can in principle be completely described.

When the right hand side of a dynamical system is more complicated than just linear, these nice properties may fail partially or completely. The theory of dynamical systems deals with establishing what results are possible to achieve and establishing the validity range of such results (what we usually call *assumptions*). Along the way, many of the results generally valid for linear systems will be repeatedly recovered (under specific validity conditions) and frequently used. Moreover, some associated linear systems will help us understanding the qualitative behaviour of nonlinear systems.

2.3. Exercises

EXERCISE 2.1. Explain how a higher–order ODE can be transformed into a first order ODE (Section 2.1).

EXERCISE 2.2. The harmonic oscillator.

(a) Recast the equation $\ddot{x} + {w_0}^2 x = 0$ (Newton's equation for a harmonic oscillator) as a dynamical system. Solve the system for arbitrary initial conditions.

(b) Show that any harmonic oscillator may be recast as the *same* dynamical system through an adequate change of coordinates (possibly including a rescaling in time).

(c) Try similar steps for $\ddot{x} + \beta\dot{x} + {w_0}^2 x = 0$, the damped harmonic oscillator.

(d) Hooke's law for elastic forces reads $m\ddot{x} = -kx$, where m is the mass of a point particle, k the elastic constant (both positive) and x the elastic elongation. Recast Hooke's law using the previous steps. What is w_0?

EXERCISE 2.3. Solve the linear system $\dot{x} = Ax$, $x \in \mathbb{R}^2$, $x_0 = (3,4)$, for the following cases (cf Remark 2.5):

$$A_1 = \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix}; \qquad A_2 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}; \qquad A_3 = \begin{pmatrix} 0 & 2 \\ 0 & 1 \end{pmatrix}.$$

2.4. Fundamental Theory: Existence and Uniquenes of solutions to ODE:s

In this Section we want to investigate the connection between ordinary differential equations and dynamical systems.

First we observe that any smooth flow of a dynamical system in continuous time $\phi_t = \phi(\cdot, t)$, $t \in \mathbb{R}$ defines a differential equation which is satisfied by the trajectories $t \to \phi(y,t)$, $y \in \mathbb{R}^n$. Namely, we set

$$v(x) = \frac{d}{dt}\phi(x,t)\Big|_{t=0}$$

and get $\dot{x} = v(x)$; $x(0) = y$.

The main question is under what conditions a differential equation defines a dynamical system. We impose two conditions that are natural from the applications point of view:

1. For any initial value there must be a solution.
2. The solution must be unique (otherwise the evolution law would be not uniquely specified). Hence, situations as those in Figure 2.10 are forbidden.

Unicity of solutions is coupled to the concept of *determinism.*

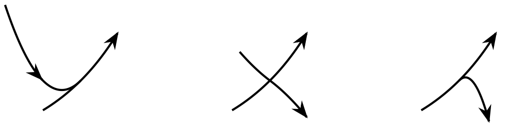


FIGURE 2.10. Unacceptable situations for a dynamical system.

The first condition is ensured by continuity. This is the content of Theorem 2.1. However, it is clear from Example 2.3 that a continuous vector field is not enough to ensure uniqueness of the solutions. For that, it will be necessary to demand an additional restriction.

THEOREM 2.1 (Peano [**Pea86**]). *Let* $v\colon \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ *be continuous in a neighbourhood* $U \times T$ *of* (x_0, t_0). *Then there is a time interval* $[t_0, t_0+c]$ *such that a solution for the equation* $\dot{x}(t) = v(x(t),t)$; $x(t_0) = x_0$ *exists for* $t \in [t_0, t_0+c]$.

We will prove instead a more general theorem covering both existence and uniqueness, in the more general setup of a non–autonomous case (for the autonomous case we just skip the dependence on time of the vector field). Before doing so we reformulate the (non-autonomous) differential equation

$$\dot{x}(t) = v(x(t),t); \quad x(t_0) = x_0. \tag{2.2}$$

By integrating with respect to t we get

$$x(t) = x_0 + \int_{t_0}^{t} v(x(s),s)\,ds. \tag{2.3}$$

So any solution of (2.2) satisfies (2.3). On the other hand if v is continuous we are allowed to differentiate (2.3) and therefore any solution of (2.3) satisfies (2.2). Hence, these two equations are equivalent whenever v is continuous. Unless otherwise stated, we will assume by default that the vector field v is continuous.

THEOREM 2.2 (Picard–Lindelöf). [**CL55**] *Let $U \in \mathbb{R}^n, T \in \mathbb{R}$. Let v be a continuous vector field satisfying the* Lipschitz condition*,*

$$|v(x_1,t) - v(x_2,t)| \leq L|x_1 - x_2|$$

on $U \times T$ (Lipschitz on x; for t continuity is enough). Then for any $(x_0,t_0) \in U \times T$ there are neighbourhoods $U_1 \times T_1 \in U \times T$ such that (2.2) *has exactly one solution $x(t)$ in $U_1 \times T_1$.*

PROOF. We are going to use the Banach Fixed Point Theorem[**Ban22**] (see Appendix A.3.1). For an arbitrary point (x_0,t_0) in the Lipschitz domain, we consider the map $A = A_{x_0,t_0} \colon C(T,U) \to C(T,U)$ defined by

$$A(x(t)) := x_0 + \int_{t_0}^{t} v(x(s),s)\,ds.$$

$C(T,U)$ denotes the space of continuous functions on $U \times T$. The map A transforms continuous functions into continuous functions on the same domain. By comparing with (2.3) we notice that any fixed point $A(x(t)) = x(t)$ of this map is a solution of (2.2) and moreover all solutions have this form. If we find a suitable complete metric space of functions where A is a contraction, then by Banach's fixed point theorem there exists a unique fixed point for A, and hence a unique solution to the differential equation.

We choose $(x_0,t_0) \in U_1 \times T_0 \subset U \times T$ such that $\operatorname{diam} U_1 = a$ and $\sup_{U_1 \times T_0} |v| < K$ for some a, K ($U_1 \times T_0$ must be chosen differently from $U \times T$ in case this set is not compact) . Now let $t_0 \in T_1 \subset T_0$ be such that $\operatorname{diam} T_1 = b < a/K$. Then any curve $y(t)$ with $y(t_0) \in U_1$ and $|\dot{y}(t)| \leq K$ satisfies

$$|y(t) - y(t_0)| \leq K|t - t_0| < Kb < a.$$

Hence $y(t) \in 2U_1$ for all $t \in T_1$ (see Figure 2.11).

We define X to be the space of all continuous functions $x \colon T_1 \to 2U_1$, $x(t_0) = x_0 \in U_1$ with the metric (distance)

$$d(x_1(t),x_2(t)) := \|x_1(t) - x_2(t)\|_\infty := \sup_{t \in T_1} |x_1(t) - x_2(t)|.$$

With this choice of distance, X is a complete metric space (see Appendix A.1). We need to prove that A maps X into itself and that it is a contraction.

We observe that $A(x(t))(t_0) = x_0 \in U_1$ and

$$|A(x(t)) - A(x(t_0))| \leq \left| \int_{t_0}^{t} |v|\,ds \right| \leq K|t - t_0| < a.$$

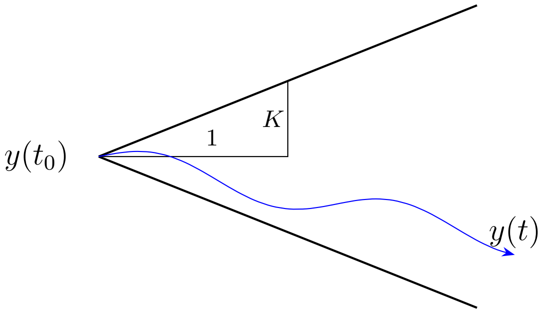


Figure 2.11. The trajectory $y(t)$ and its bounds.

This implies $A(x(t))(\tau) \in 2U_1$ for $\tau \in T_1$ and hence $AX \subset X$.

Moreover,

$$\begin{aligned} |A(x_1(t)) - A(x_2(t))| &\leq \left| \int_{t_0}^{t} |v(x_1(s), s) - v(x_2(s), s)| \, ds \right| \\ &\leq L \left| \int_{t_0}^{t} |x_1(s) - x_2(s)| \, ds \right| \\ &\leq L|t - t_0| \cdot \|x_1 - x_2\|_\infty \leq Lb\|x_1 - x_2\|_\infty. \end{aligned}$$

Since the right hand side is independent of t, the estimate holds also for $\|A(x_1(t)) - A(x_2(t))\|_\infty$. By making T_1 smaller if necessary we may assume that $Lb < 1$. Hence, A is a contraction on a complete metric space and the Banach fixed point theorem concludes the proof. □

Remark 2.7. The map $A = A_{x_0,t_0}$ depends continuously on x_0, t_0. By using the Banach fixed point theorem again we can show that the solutions depend Lipschitz-continuously on the initial values. We will see later that we actually have differentiable dependence.

The Banach fixed point theorem provides a recursive approximation to the exact solution.

Lemma 2.1. *Let $x_0(t) \equiv x_0$ and for $n \geq 0$ set $x_{n+1}(t) := A(x_n(t))$. Then*

$$\|x_{n+1}(t) - x_n(t)\|_\infty \leq \frac{KL^n b^{n+1}}{(n+1)!}.$$

Proof. We prove it by induction (see Appendix A.1.4 for a comment on the Induction Principle). The statement holds for $n = 0$ with the estimates of the previous Theorem. For a general n the statement implies

$$|x_{n+1}(t) - x_n(t)| \leq \frac{KL^n |t - t_0|^{n+1}}{(n+1)!}.$$

Then, using the same estimates of the Theorem and computing the integral in the last step, we have

$$\begin{aligned}
|x_{n+2}(t) - x_{n+1}(t)| &\leq \left| \int_{t_0}^{t} |v(x_{n+1}(s), s) - v(x_n(s), s)|\, ds \right| \\
&\leq L \left| \int_{t_0}^{t} |x_{n+1}(s) - x_n(s)|\, ds \right| \\
&\leq \frac{KL^{n+1}}{(n+1)!} \left| \int_{t_0}^{t} |s - t_0|^{n+1}\, ds \right| \\
&\leq \frac{KL^{n+1}}{(n+2)!} |t - t_0|^{n+2} \leq \frac{KL^{n+1} b^{n+2}}{(n+2)!}.
\end{aligned}$$

□

As an illustration, let us apply this Lemma to the class of linear systems $\dot{x} = Bx$, $x(0) = x_0$. We know already that the (unique) solution is a time-exponential function. Following the Lemma,

$$\begin{aligned}
x_0(t) &\equiv x_0 \\
x_1(t) &= x_0 + \int_0^t Bx_0\, ds = x_0 + tBx_0 \\
x_2(t) &= x_0 + \int_0^t B(x_0 + sBx_0)\, ds = x_0 + tBx_0 + \frac{t^2}{2} B^2 x_0 \\
&\cdots \\
x_n(t) &= x_0 + \int_0^t B \left(\left(\sum_{j=0}^{n-1} \frac{s^j}{j!} B^j \right) x_0 \right) ds \\
&= x_0 + \sum_{j=0}^{n-1} \left(\int_0^t \frac{s^j}{j!}\, ds \cdot B^{j+1} x_0 \right) = \left(\sum_{j=0}^{n} \frac{t^j}{j!} B^j \right) x_0.
\end{aligned}$$

Hence,

$$x(t) = \left(\sum_{j=0}^{\infty} \frac{t^j}{j!} B^j \right) x_0 = e^{tB} x_0.$$

2.5. Exercises

EXERCISE 2.4. Sketch the graphs of the integral curves of the vector field $v(x,t) = \frac{x-3t}{t+3x}$ (Section 2.1).

EXERCISE 2.5. Sketch the graphs of the integral curves of the vector field $v(x,t) = t - e^x$ (Section 2.1).

EXERCISE 2.6. Find defining ODE's for the families of curves

a) $x^2 + Cy^2 = 2y$

b) $y^2 + Cx = x^3$.

(Section 2.1.1).

EXERCISE 2.7. Let $v\colon \mathbb{R} \to \mathbb{R}$ be Lipschitz-continuous. Prove that any integral curve $x(t)$ of $\dot{x}(t) = v(x(t))$ is a monotone function. Hint: Show that to the sets $\{v > 0\}$, $\{v = 0\}$ and $\{v < 0\}$ there correspond different integral curves (Section 2.1.2).

EXERCISE 2.8. Let A be a $n \times n$ diagonal matrix with non–zero entries on the diagonal. Prove that the space of all solutions of $\dot{x} = Ax$ is an n–dimensional linear space (Section 2.1.2).

EXERCISE 2.9. Show that the solution $x(x_0, t)$ of a differential equation as given by Picard–Lindelöf theorem is Lipschitz continuous on x_0 (Section 2.4).

EXERCISE 2.10. Construct the first three approximations of the solutions of

a) $\dot{x}(t) = t - x^2$; $x(0) = 0$
b) $\dot{x}(t) = x^2 + 3t - 1$; $x(1) = 1$

(Section 2.4).

EXERCISE 2.11. Find the Taylor expansion of $\sin t$ with the help of the theorem of Picard–Lindelöf applied to

$$\ddot{x} = -x \quad x(0) = 0; \dot{x}(0) = 1.$$

(Section 2.4).

2.6. Dependence on Initial Conditions

Now that we established the existence and uniqueness of the solutions for a dynamical system with Lipschitz right hand side, let us further explore the properties of such a solution. In many, if not most, models for applications in Sciences and Engineering, the right hand side $v(x, t)$ of the equations is not only Lipschitz but also differentiable, or even C^∞ (infinitely differentiable). We state therefore the following

Assumption: For the function $v\colon U \times T \subset \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$, we assume that there exists a continuous derivative (matrix!)

$$D_x v\colon U \times T \to \mathbb{R}^{n\times n}(\mathbb{R}) \qquad (x, t) \to \left(\frac{\partial v_i}{\partial x_j}\right).$$

Let us show that under this assumption the solutions $x(x_0, t)$ to

$$\dot{x}(t) = v(x(t), t); \qquad x(t_0) = x_0$$

depend differentiably on x_0, i.e., that the derivative $D_{x_0} x(x_0, t)$ exists (see Figure 2.12). Let us gain some intuition around the problem. If this derivative exists, how would it change in time? Does it fulfill some differential equation? We proceed,

$$
\begin{aligned}
\frac{\partial}{\partial t}\left(D_{x_0}x(x_0,t)\right) &= D_{x_0}\left(\frac{\partial}{\partial t}x(x_0,t)\right)\\
&= D_{x_0}\left(v(x(x_0,t),t)\right)\\
&= D_x v(x,t)\cdot D_{x_0}x(x_0,t).
\end{aligned}
$$

with initial condition

$$D_{x_0}x(x_0,t_0) = \mathrm{Id} \quad \text{or} \quad x(x_0,t_0) = x_0.$$

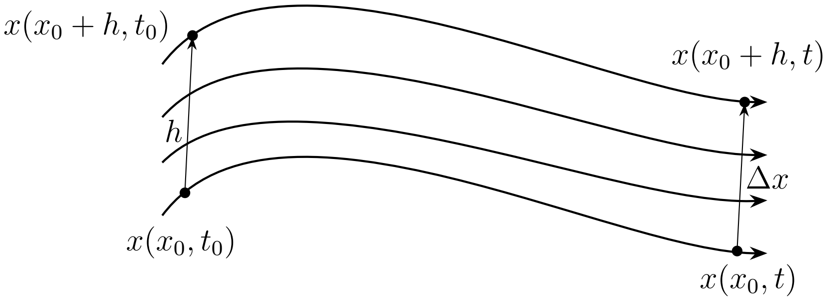


FIGURE 2.12. Modifications of the solution $x(t)$ when changing the initial condition x_0.

We can write this equation as the matrix differential equation

$$\dot{B}(t) = D_x v(x(t),t)\cdot B(t) \qquad B(t_0) = \mathrm{Id} \tag{2.4}$$

where $B(t)$ is the $n\times n$ matrix $D_{x_0}x(x_0,t)$. This is a linear equation and hence somewhat simpler to address. To make sense of this expression we must assure that such a matrix $B(t)$ exists, i.e., that $x(x_0,t)$ is actually continuously differentiable with respect to x_0.

THEOREM 2.3. [**CL55**] *Under the previous assumption the derivative $D_{x_0}x(x_0,t)$ exists and fulfills the* **variational equation** (2.4)

PROOF. Note first that we assume that v is differentiable, but we want to prove that also x is differentiable with respect to x_0. Fix U_1, T_1 as in Picard–Lindelöf Theorem. Let $x_0\in U_1, t_0\in T_1, x_0+h\in U_1$. We consider (see Figure 2.12)

$$\Delta x(x_0,t,h) := x(x_0+h,t) - x(x_0,t)$$

and define

$$A(t,h) := \int_0^1 D_x v\left(x(x_0,t) + s\Delta x(x_0,t,h),t\right)\, ds.$$

Hence, $A(t,h)$ is an $n\times n$ matrix since it is the integral of the $n\times n$ matrix D_xv. Moreover, $A(t,0)=D_xv(x(t),t)$. Since the solutions $x(\cdot,t)$ are differentiable functions of time, we can compute the derivative,

$$\begin{aligned}\frac{d}{dt}\left(\Delta x(x_0,t,h)\right) &= \frac{d}{dt}\left(x(x_0+h,t)-x(x_0,t)\right)\\ &= v(x(x_0+h,t),t)-v(x(x_0,t),t)\\ &= \int_0^1 \frac{d}{ds}v\left(x(x_0,t)+s\Delta x(x_0,t,h),t\right)\,ds\\ &= \int_0^1 D_xv\left(x(x_0,t)+s\Delta x(x_0,t,h),t\right)\,ds\cdot\Delta x(x_0,h,t)\\ &= A(t,h)\cdot\Delta(x_0,t,h).\end{aligned}$$

The first integral holds because of Leibniz' rule of differentiation under the integral sign, while the second integral holds since the dependence of v on s occurs via x. Hence,

$$\dot{\Delta}x(x_0,t,h)=A(t,h)\cdot\Delta(x_0,t,h)\quad \Delta x(x_0,t_0,h)=h$$

where $A(t,h)$ is a matrix (linear operator). Moreover, $A(t,h)$ is Lipschitz continuous in h. Hence, for fixed h there exists a unique solution to the matrix differential equation

$$\dot{M}(t,h)=A(t,h)M(t,h)\qquad M(t_0,h)=\mathrm{Id}$$

Multiplying by h we have that the vector $y(t,h)=M(t,h)\cdot h$ satisfies the equation

$$\dot{y}(t,h)=A(t,h)y(t,h);\quad y(t_0,h)=h.$$

Therefore

$$\Delta x(x_0,t,h)=M(t,h)\cdot h$$

since both functions are solutions to the same equation with the same initial condition. Since $M(t,h)$ is Lipschitz continuous in h, we have that

$$\lim_{|h|\to 0}\frac{\Delta x(x_0,t,h)}{|h|}=\lim_{|h|\to 0}\frac{M(t,h)\cdot h}{|h|}=M(t,0)$$

exists. This limit is $D_{x_0}x(x_0,t)$ and satisfies the the differential equation (2.4). □

Corollary 2.1. *If all the derivatives of the vector field up to order k exist and are continuous then the solutions are k–times differentiable in the initial conditions.*

Proof. We only have to consider the new equation

$$\begin{aligned}\dot{x}&=v(x,t)\\ \dot{z}&=D_xv(x,t)\cdot z\end{aligned}$$

which satisfies the conditions of the previous theorem. The complete argument uses induction: If v is k times differentiable, then D_xv is

$k-1$ times differentiable. But then so is $\dot{z}$. Hence z is $(k-1)+1=k$ times differentiable. □

Assumption: Unless otherwise stated we will assume in what follows that the vector field v is a C^k function, $k \geq 1$.

CHAPTER 3

Advanced Ordinary Differential Equations

At this point we know that under suitable conditions a dynamical system has a unique solution which is differentiable in time and in initial conditions. In this chapter we will start our analysis of local properties of the set of solutions[**GH86**]. We will see how many classes of solutions we may encounter, further we will realise that there are some regions of phase space where "nothing happens", i.e., the dynamics looks like a transport in time without significant features. Further, we will describe some topological (structural) features of phase space that are helpful to identify the special sets where the dynamics does have relevant, changing features. We end this chapter by studying properties of 2-dimensional dynamical systems.

3.1. The local flow

We start by noting that there is a way to unify the study of autonomous and non–autonomous situations.

Proposition 3.1. [**SNM96**] *Any dynamical system can be recast as an autonomous system on a larger phase space.*

To achieve this we can introduce time as a new coordinate s, where $s(t) = t$. Phase space is now described with coordinates (x, s). The new problem reads:

$$\begin{cases} \dot{x} &= v(x,s) \\ \dot{s} &= 1, \end{cases}$$

with initial conditions $x(t_0) = x_0$, $s(t_0) = t_0$, or

$$\dot{y} = \hat{v}(y)$$

where $\hat{v} = (v, 1) \in \mathbb{R}^n \times \mathbb{R} = \mathbb{R}^{n+1}$ and $y = (x, s) \in \mathbb{R}^{n+1}$. We will focus in the sequel on autonomous systems, without loss of generality. Explicit time–dependence will be restored whenever practical.

Remark 3.1. For autonomous systems, if $x(t)$ is a solution, so is $x(t + s)$:

$$\dot{x}(t+s) = v(x(t+s)); \qquad x((t_0 - s) + s) = x_0.$$

Hence, we can fix the initial condition at time $t_0 = 0$.

Theorem 3.1 (Extension theorem). [**AAA**$^{+}$**97**] *Let $x_i(t)$; $i = 1, 2$ be two solutions defined on open time intervals T_i. Assume that their initial values coincide $x_1(0) = x_2(0)$. Then $x_1 \equiv x_2$ on $T_1 \cap T_2$.*

Proof. Let $b = \sup\{s \geq 0 \,:\, x_1(t) = x_2(t); 0 \leq t < s\}$. The values of t where it makes sense to compare x_1 and x_2 belong to $T_1 \cap T_2$, which is open. Hence, $b \in \overline{T_1 \cap T_2}$. If $b \in T_1 \cap T_2$, then by continuity of the solutions $x_1(b) = x_2(b)$ (since both solutions exist for $t = b$). Then, by the Picard–Lindelöf theorem we can extend each solution in an unique way from the initial condition $x_i(b)$ for some time-interval $[0, \epsilon)$, $\epsilon > 0$. This means that $x_1(b + t) = x_2(b + t)$ for $t \in [0, \epsilon)$, thus contradicting the definition of b. Hence, the solutions coincide on all of $T_1 \cap T_2$. □

Remark 3.2. We can also define the solution on $T_1 \cup T_2$ as long as $T_1 \cap T_2 \neq \varnothing$ in a unique way:

$$x(t) = \begin{cases} x_1(t) & t \in T_1 \\ x_2(t) & t \in T_2 \end{cases}$$

The previous theorem guarantees that this will be the only solution on $T_1 \cup T_2$.

We may try to extend the solution as much as possible, by splicing together the orbits defined on intersecting domains. This leads to the notion of maximal integral curve. Let the solutions of a dynamical system be defined on some intervals T_λ; $\lambda \in \Lambda$, all containing the point x_0:

$$x_\lambda \colon T_\lambda \to \mathbb{R}^n \quad x_\lambda(0) = x_0$$

We note that x_λ itself may be constructed as before by gluing together (smaller) time intervals T_i (this might be necessary to get time $t = 0 \in T_\lambda$).

Definition 3.1 (Maximal Integral Curve). We define the maximal integral curve $x(x_0, t)$ on $T_{x_0} = \cup_{\lambda \in \Lambda} T_\lambda$ via $x(x_0, t)|_{T_\lambda} = x_\lambda$.

By the Picard–Lindelöf theorem we have that the maximal time interval T_{x_0} is open (this was perhaps noticeable already in Theorem 3.1: if b were an interior point of the maximal time interval, we could find an $\epsilon > 0$ such that we have a unique solution up to time $t + \epsilon$).

Definition 3.2 (Local Flow). The local flow is defined on $A = \{(x_0, t) \subset U \times T \,:\, t \in T_{x_0}\}$ by

$$\Phi(x_0, t) = \phi_t(x_0) = x(x_0, t)$$

where $\dot{x}(x_0, t) = v(x(x_0, t))$ and $x(x_0, 0) = x_0$.

We summarise these findings in the following

THEOREM 3.2 (Properties of Local Flow). [**Arn73**, p.51] *For the local flow,*

1. *A is open and the local flow is C^k.*
2. *$A \cap (\{x_0\} \times \mathbb{R}) = T_{x_0} \ni 0$.*
3. *$\Phi(\Phi(x_0,t),s) = \Phi(x_0,s+t)$ and $\Phi(x_0,0) = x_0$.*

PROOF. The only new point is that A is open. But by the theorem of Picard–Lindelöf and the dependence on initial conditions we can extend the solutions at any time and at any place into small time and space neighbourhoods. So the set A must be open. $\square$

REMARK 3.3. We note that there exist several notations indicating very similar or the same objects. For example, the expressions $\Phi(x_0,t)$, $\phi_t(x_0)$, $\phi(x_0,t)$ and $x(x_0,t)$ all denote a certain point of phase space, specifically that where the trajectory starting at x_0 arrives after a time t. Although the first expression is associated to the concept of flow and the last expression is associated to the concept of trajectory, the middle expressions, with ϕ, may denote both. We will follow this tradition as long as no ambiguity arises.

REMARK 3.4. The phase space $\mathbb{R}^n$ is partitioned into disjoint **orbits** $\{x(x_0,t) : t \in T_{x_0}\}$ (also denoted $\phi_t(x_0)$). If two orbits have a point in common, i.e., $x(x_0,r) = x(x_1,s) = y$ then they belong to the unique solution $x(y,t) = x(x_0,t+r) = x(x_1,t+s)$. In short: Distinct orbits never cross.

COROLLARY 3.1. *Let $x\colon T_{x_0} = (a,b) \to \mathbb{R}^n$ be a maximal solution with $b < \infty$ and $K \subset U \subset \mathbb{R}^n$ a compact set. Then there is a time $\tau \in (a,b)$ such that $x(x_0,t) \notin K$ for all $t \in (\tau,b)$. In other words, if a maximal integral curve stays in a compact region it is defined for all times.*

PROOF. Assume the statement is false, i.e., that the maximal solution remains within K for all times in (a,b). In such a case we will arrive to a contradiction showing that the solution could be extended beyond b.

Since A is open it contains the set $K \times (-\epsilon,\epsilon)$ for some $\epsilon > 0$ (since K is compact!) as a neighbourhood of $K \times \{0\}$ (any neighbourhood of a compact set contains an ϵ-neighbourhood – take the cover of K by balls $B(x,\epsilon) \subset A$ and choose a finite sub-cover). Let t be such that $b - t < \epsilon$. If $x(x_0,t) \in K$ then the solution could be extended up to time $t + \epsilon > b$, a contradiction. Hence, $x(x_0,t) \notin K$. $\square$

3.2. Exercises

EXERCISE 3.1. Solve the following equations with separable variables

a) $xt\,dt + (t+1)\,dx = 0$.
b) $(t^2-1)\dot{x} + 2tx^2 = 0$.

Use your previous knowledge from elementary Calculus courses.

Exercise 3.2. Solve the equation

$$\dot{x} = x + 2t - 3,$$

by applying a suitable linear transformation $z = at + bx$.

Exercise 3.3. Solve the equation

$$(t+2x)\,dt - t\,dx = 0,$$

by applying the transformation $x = \omega t$.

Exercise 3.4. Consider the parameter dependent equation

$$\dot{x} = x^2 + 2\mu t^{-1} \quad x(1) = -1.$$

We are looking for an approximation of the solution of the form

$$x(t) = w_0(t) + \mu w_1(t) + \mu^2 w_2(t).$$

Why does such an approximation exist? Find the functions $w_i(t)$ by plugging them into the equation and comparing the coefficients of the powers of μ.

Exercise 3.5. Let v be a C^k–vector field in $\mathbb{R}^2$. Assume that the circle $\mathbb{S}^1 = \{x \in \mathbb{R}^2 : \|x\| = 1\}$ is a periodic trajectory.

a) How long does a trajectory through a point x_0 with $\|x_0\| < 1$ exist?
b) Is there a stationary solution?

3.3. The Straightening–out Theorem

In this section we will see that the dynamics generated by $\dot{x} = v(x)$ close to a regular point is very simple. This means that, from the local point of view, only singular points are of interest.

Definition 3.3. A point $x \in \mathbb{R}^n$ is called **regular point** (for the vector field v) if $v(x) \neq 0$. Otherwise it is called **singular, stationary or equilibrium point**.

Theorem 3.3. [**AAA$^+$97**] *Let $v \in C^k$ and x_0 a regular point, i.e., $v(x_0) \neq 0$. Then there is a C^k–coordinate change $h : V \to \mathbb{R}^n$ in a neighbourhood V of x_0 with*

$$h_* v = e_n = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

Remark 3.5. The notation $h_* v$ indicates the map induced on the vector field v by the coordinate change h. Formally, if $\dot{x} = v$ and $z = h(x)$, we have that $\dot{z} = (dz/dx) \cdot \dot{x} = Dh\, v$, i.e., the vector field is transformed through the matrix $\{Dh\}_{ij} = \partial h_i / \partial x_j$.

Remark 3.6. This theorem can be interpreted as follows. After a change of coordinates, i.e., in the new coordinates, the vector field is very simple. Its solutions are "horizontal" straight lines. This means that in a small neighbourhood of x nothing interesting can happen (see Figure 3.1), the dynamics is just monotone evolution in time along parallel flow lines.

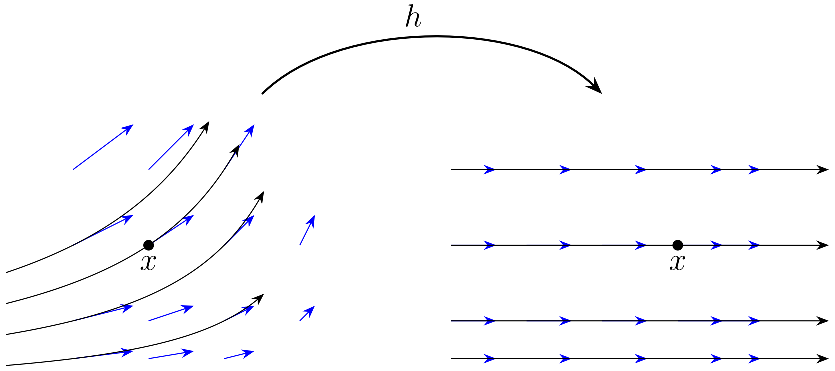


Figure 3.1. A change of coordinates h mapping the flow near a regular point to a standard flow.

Proof. The proof proceeds by showing that such a change of coordinates exists. Let[1] $v(x) = (v_1(x), \cdots, v_n(x))^T$. Without loss of generality, we may assume $v_n(x_0) \neq 0$ (some component has to be nonzero since x_0 is a regular point). Let $L = \{y = (y_1, \cdots, y_n)^T \in \mathbb{R}^n \,:\, y_n = 0\}$ and $S = (L + x_0) \cap U$ where U is chosen that $v_n|_U \neq 0$. This can be done since by continuity of v there is a neighbourhood of x_0 such that the vector field does not vanish, see Figure 3.2.

S is a portion of a hyperplane, passing through x_0 and perpendicular to the vector $(0, \cdots, 0, v_n(x_0))^T$. For sufficiently short time intervals $|t| < \epsilon$, any initial condition $x_0 + y \in S$ remains within U. The set of all trajectories of the dynamical system with initial condition $x(0) = x_0 + y \in S$ and extending for the whole time-interval $-\epsilon < t < \epsilon$, defines hence a small open subset $V \subset U$ containing S. The coordinates (y, t), with $y \in L$ and $-\epsilon < t < \epsilon$ describe a small cube B in $\mathbb{R}^n$ around the origin (with Cartesian, orthonormal coordinates). After having evolved for a time t, the trajectory starting at $x_0 + y$ arrives to the point $x(x_0 + y, t) \in V$. We give to this point the new coordinates

[1] Given a basis, we usually prefer to denote vectors by $n \times 1$ matrices, i.e., column matrices.

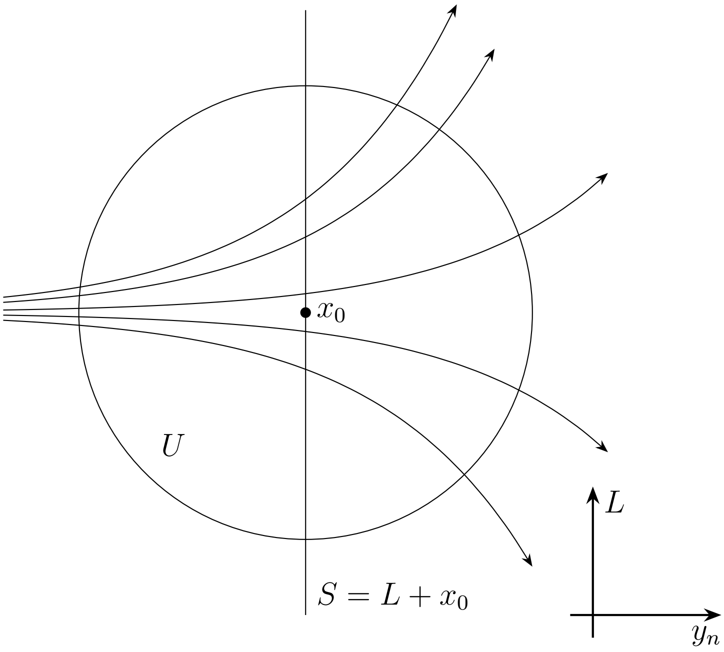


FIGURE 3.2. The procedure in the Straightening–out Theorem.

$(y,t)^T \in \mathbb{R}^n$ (we identify t with the last coordinate of $\mathbb{R}^n = L \times \mathbb{R}$). Hence, the change of coordinates $h : V \to B \subset \mathbb{R}^n$ reads:

$$\begin{pmatrix} y \\ t \end{pmatrix} = h(x(x_0 + y, t)) \in \mathbb{R}^n.$$

In other words, we use the trajectories as new coordinate lines. By the Theorem of Picard and Lindelöf, this map should be well defined and 1–to–1. Let us verify this by computing the determinant of its inverse map at some point in B. We use hence Dh^{-1}, the matrix of first derivatives of h^{-1} with respect to each coordinate. We have, for $x_0 + y \in S$,

$$D(h^{-1}(y,t)) = \left(\frac{\partial}{\partial y} x(x_0 + y, t), \dot{x}(x_0 + y, t)\right)$$

where the last column of this matrix is the original vector field v, and the first $n-1$ columns are the derivatives of the trajectories with respect to the initial condition. For $t = 0$ the trajectories lie at their initial condition, hence $h^{-1}(y, 0) = x_0 + y$ and the derivatives with respect to each y_i are either 0 or 1. Hence,

$$Dh^{-1}(y,0) = \begin{pmatrix} 1 & \cdots & 0 & v_1(x_0 + y) \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & v_{n-1}(x_0 + y) \\ 0 & \cdots & 0 & v_n(x_0 + y) \end{pmatrix}$$

with non–zero determinant at $y = 0$ (as expected) because $v_n(x_0) \neq 0$. The determinant is non–zero in all of B by continuity. Moreover, for $t \neq 0$ the last column of this matrix is still v. The vector field in the new coordinates can either be computed directly or transforming v with h. Since

$$v(x(x_0 + y, t)) = Dh^{-1}(y, t) \cdot h_* v,$$

the vector field in the new coordinates is just the unit Cartesian vector $h_* v = (0, \cdots, 0, 1)^T = e_n$ (the equation has unique solution since Dh^{-1} is non-singular). Hence, the transformation maps the vector field into the straight lines $e_n + y$ parallel to e_n through L. □

3.4. Differential inequalities

In this section we want to show how a vector field dominates differentiable functions. We can apply this result to prove the existence of trajectories in a given time interval.

THEOREM 3.4 (Comparison theorem). [**Arn73**, p.17] *Let v be a real-valued function fulfilling the assumptions of the theorem of Picard–Lindelöf and $x(t)$ be a solution of $\dot{x}(t) = v(x(t), t)$ for $t \in \mathbb{R}$. Let $y : \mathbb{R} \to \mathbb{R}$ be a differentiable function and $t_0 \in \mathbb{R}$ with*

$$\dot{y}(t) \leq v((y(t), t) \quad \textit{and} \quad y(t_0) \leq x(t_0).$$

Then

$$y(t) \leq x(t), \quad t \geq t_0.$$

PROOF. We will prove the situation in case the inequalities are strict. The general situation is only slightly more difficult.

Let

$$b = \sup\{s \,:\, y(t) \leq x(t), \quad t_0 \leq t \leq s\}.$$

By continuity $x(b) = y(b)$ if $b < \infty$. Hence,

$$\dot{y}(b) - \dot{x}(b) < v(y(b), b) - v(x(b), b) = 0.$$

This implies $y(b+t) < x(b+t)$ for small t since the derivative of $y - x$ at b is negative. This contradicts the definition of b. □

EXAMPLE 3.1. Let $v(x,t) = A(x,t) \cdot x$ with $A(x,t) \colon \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^{n \times n}$ be a continuous matrix function with operator norm (see Appendix A.5)

$$\|A(x,t)\| := \sup_{\|x\|=1} \|A(x,t) \cdot x\| < a.$$

We want to investigate the lifetime of the trajectories starting at $x(t_0) = x_0$. We have

$$\begin{aligned}\frac{d}{dt}\|x(x_0,t)\|^2 &= \frac{d}{dt} < x(x_0,t), x(x_0,t) >= 2 < x(x_0,t), \dot{x}(x_0,t) > \\ &= 2 < x(x_0,t), A(x(x_0,t),t)x(x_0,t) > \\ &\leq 2a < x(x_0,t), x(x_0,t) > \\ &= 2a\|x(x_0,t)\|^2.\end{aligned}$$

Now we consider the differential equation in $\mathbb{R}$

$$\dot{z} = 2az$$

with solution

$$z(t) = z(0)e^{2at}.$$

The previous theorem applied to $y(t) = \|x(x_0,t)\|$ tells us that

$$\|x(x_0,t)\|^2 \leq \|x_0\|^2 e^{2at}.$$

This means that the solutions are finite at any time and hence live forever by Corollary 3.1.

3.5. Types of flow–lines

In this section we will study the different ways a trajectory may look like. There are not too many possibilities.

THEOREM 3.5. [**Tes12**] *Let $x(t)$ be a maximal integral curve of the system $\dot{x} = v(x)$ with Lipschitz continuous right hand side, and let T be the maximal time interval the trajectory lives. Then one of the following holds (see Figure 3.3):*

1. *$x(t) \equiv const$ and $T = \mathbb{R}$, i.e., the trajectory is defined for all times,*
2. *$x(t)$ is* **regular** *(i.e., $\dot{x}(t) \neq 0$ for $t \in T$) and* **injective***, or*
3. *$x(t)$ is* **regular periodic***; i.e., $T = \mathbb{R}$ and there is a minimal positive τ with*

$$x(t) = x(t+\tau) \quad \textit{for all } t \in \mathbb{R}.$$

τ is called the **period.**

PROOF. Let $\dot{x}(s) = v(x(s)) = 0$ for some $s \in T$. Then $x(t) \equiv x(s)$ is the unique solution for all $t \in \mathbb{R}$ with this initial condition $x(s)$, by the theorem of Picard and Lindelöf. This is case 1.

Now let $\dot{x}(t) \neq 0$ for all $t \in T$. Then either $x(t)$ is injective (and we are in case 2) or there are $r < s \in T$ such that $x(r) = x(s)$. Let $\tau = s - r$. This yields $x(r) = x(r + \tau)$. We will see that the period is the same along the whole trajectory. Indeed, the integral curves

$$x_1(t) = x(t+r) \quad \text{and} \quad x_2(t) = x(t+r+\tau)$$

have the same initial condition $x_i(0) = x(r) = x(s)$. By uniqueness we get

$$x(t+r) = x_1(t) = x_2(t) = x(t+r+\tau) \quad \text{for all } t \in T.$$

Hence, $x(t) = x(t+\tau)$; $x(0) = x(\tau)$ ("set" $t = t+r$). Then $x(T) \subset x([0,\tau])$, since the values repeat afterwards. This is a compact set (as the image of a compact set under a continuous map), implying that $T = \mathbb{R}$ by corollary 3.1.

We are left to show that there is a least positive period. Let

$$G = \{\tau \in \mathbb{R}^+ \,:\, x(0) = x(\tau)\}.$$

$G \neq \mathbb{R}^+$, otherwise we are in case 1 since then the trajectory would be identically constant. Let us show that G has a least positive element. Since $x(t)$ is not constant, for some $t_0 > 0$ we have $x(t_0) \neq x(0)$. By continuity there exists $\epsilon > 0$ such that $x(t_0 + t) \neq x(0)$ for $|t| < \frac{\epsilon}{2}$. Hence, if $g \in G$, then $g \geq \epsilon$ (i.e., G has a positive infimum). In other words, the period cannot be less than ϵ, since we have produced an ϵ–interval where $x(t)$ is different from $x(0)$. Using the same argument, two different periods $g_1, g_2 \in G$ fulfill $|g_2 - g_1| \geq \epsilon$. Since τ is a period, we have $\epsilon \leq \tau$. The intersection $G \cap [\epsilon, \tau]$ is nonempty and finite. Hence, there exists a least element τ_0 for G. Moreover, the elements of G are of the form $\tau_0 \cdot \mathbb{N}$.

□

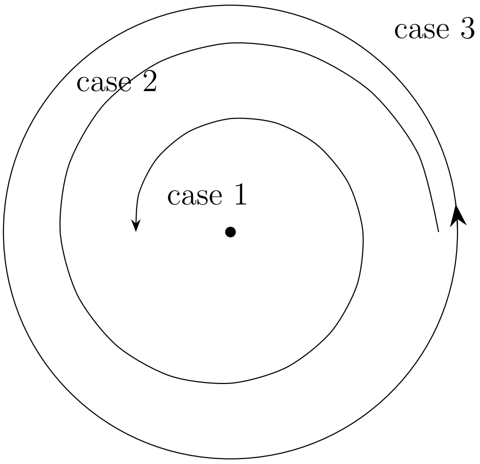


Figure 3.3. Types of flow lines

Remark 3.7. Since for a single orbit the set of periods (including positive and negative periods) is a proper closed additive subgroup of the reals, another proof of the Theorem can be based on the following algebraic result: Any proper closed subgroup of the reals is a lattice (see Theorem A.15), i.e., it is equal to a positive constant τ_0 (the least period) times $\mathbb{Z}$.

Example 3.2. We consider a special flow on

$$\mathbb{S}^1 \times \mathbb{S}^1 = \{(z_1, z_2), z_i \in \mathbb{C} : |z_i| = 1, i = 1, 2\} = \mathbb{T}^2$$

given by

$$\phi_t(z_1, z_2) = \left(e^{2\pi it} z_1, e^{2\pi iat} z_2\right).$$

We consider the integral curve through the point $(1, 1)$; i.e., the curve

$$\phi_t(1, 1) = \left(e^{2\pi it}, e^{2\pi iat}\right).$$

We have two different cases

a) $a \in \mathbb{Q}$; $a = p/q$. This implies for time $t = q$

$$\phi_q(1, 1) = \left(e^{2\pi iq}, e^{2\pi ip}\right) = (1, 1)$$

and we are in case 3 of Theorem 3.5 (see Figure 3.4).

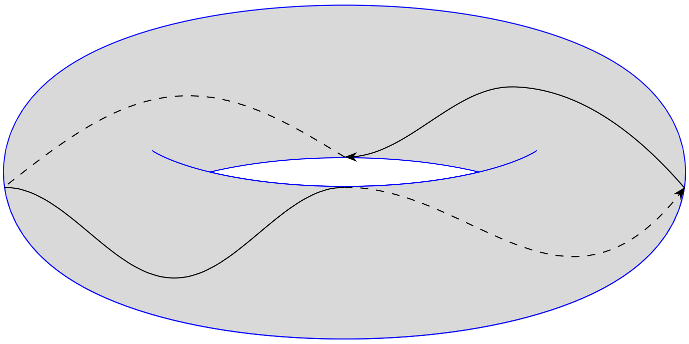


Figure 3.4. A periodic orbit on the surface of a torus. Dashed parts lie "behind" the torus.

b) $a \notin \mathbb{Q}$. Then the equation

$$\phi_t(1, 1) = (1, 1)$$

has the only solution $t = 0$, since $e^{2\pi it} = 1$ implies $t \in \mathbb{Z}$ and $e^{2\pi iat} = 1$ implies $ta = m \in \mathbb{Z}$ But then we would have $a = \frac{m}{t} \in \mathbb{Q}$, which is not the case. So we are in case 2 of Theorem 3.5 (see Figure 3.5). We see also that in this Example, case 1 of the Theorem is never attained.

In the case $a \notin \mathbb{Q}$ we can say even more. Let the first coordinate be arbitrary but fixed, i.e., $e^{2\pi ib}$. Then the trajectory intersects the circle $(e^{2\pi ib}, z_2)$ with $|z_2| = 1$, i.e., $\phi_t(1, 1) = \left(e^{2\pi ib}, z_2(t)\right)$ if and only if $t = b + k$; $k \in \mathbb{Z}$. In this case $\phi_{b+k}(1, 1) = \left(e^{2\pi ib}, e^{2\pi i(ab+ka)}\right)$. Hence, the trajectory of $(1, 1)$, regarded at fixed integer times starting at $t = b$, wanders around the circle $(e^{2\pi ib}, z_2)$ in some way.

Let now the second coordinate $z_2 = e^{2\pi ic}$ be given. Will the trajectory come any close to z_2 ? By Kronecker's theorem A.18 the sequence $ka \mod 1$ is dense in $[0, 1)$ and hence the points $\left(e^{2\pi ib}, e^{2\pi i(ab+ka)}\right)_{k\in\mathbb{N}}$ approximate the point $\left(e^{2\pi ib}, e^{2\pi ic}\right)$

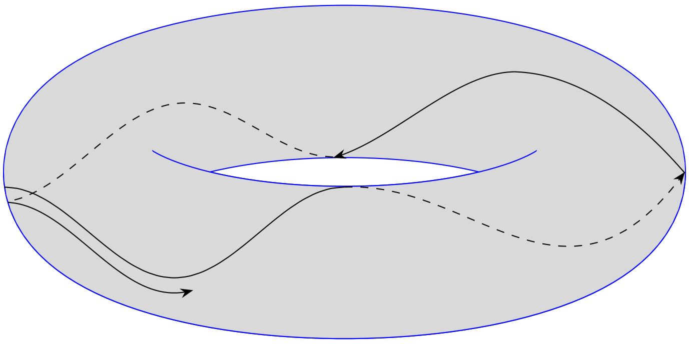


FIGURE 3.5. Part of a non–periodic trajectory.

and any arbitrary neighbourhood of z_2 will contain be points of the sequence. Hence, the orbit $\{\phi_t(1,1)\}$ is dense, provided a is irrational (recall that z_2 was given but arbitrary). The same holds for an arbitrary trajectory $\{\phi_t(z_1,z_2)\}$.

There is another method to investigate the flow of the previous Example. We fix the first circle at e.g., $b = 0$. We are left with a circle on the second coordinate, the set $\{(1,z_2) : |z_2| = 1\}$ of phase space. We are looking for the first intersection of the trajectory $\{\phi_t(1,z) : t > 0\}$ with the circle $\{(1,z_2) : |z_2| = 1\}$. This intersection occurs for $t = 1$ and subsequently for all integer times. This defines a map from the circle to itself denoted by $\hat{f} \colon \mathbb{S}^1 \to \mathbb{S}^1$ in the following way:

$$\hat{f}(z) =< (0,1), \phi_1(1,z) >= e^{2\pi i a} z,$$

Where $<,>$ is a scalar product, in this case the projection onto the second component of the image $\phi_1(1,z)$ of z. Let now $z = e^{2\pi i\alpha}$. Hence,

$$\hat{f}(e^{2\pi i\alpha}) = e^{2\pi i a} e^{2\pi i\alpha} = e^{2\pi i(\alpha+a)}.$$

This map is called the **Poincaré map**[**GH86, SNM96**] of the flow (with respect to the control section given by the circle). It contains all important information about this flow. It has the following properties:

1. $a \in \mathbb{Q}$: Every orbit is periodic and the phase space decomposes into an uncountable number of periodic (invariant) trajectories.
2. $a \notin \mathbb{Q}$: Every orbit is dense.

In a different notation, the Poincaré map corresponds to the map

$$f(\alpha) = \alpha + a \mod 1,$$

defined on the unit interval $[0,1]$.

3.6. More Topological Properties

We can now introduce more fundamental concepts from Dynamical Systems theory.[**SNM96, KH96**]

Definition 3.4. A non–empty set D in phase space is called invariant if for all $t \in \mathbb{R}$, $\phi_t(D) = D \equiv \phi_0(D)$. If this holds for positive times, the set is called *positively invariant.*

The most natural invariants are trajectories, of all three kinds described by the previous Theorem.

Definition 3.5. A system is called **(topologically) transitive** if there exists a dense orbit. In this case the phase space cannot be decomposed into closed invariant sets.

Proposition 3.2. [**KH96**, 1.4.2 p.29] *A system $f\colon X \to X$ on a compact metric space X is transitive if and only if for any pair of non–empty open sets U, V there is a number $N = N(U, V)$ such that*

$$\phi_N(U) \cap V \neq \varnothing.$$

Proof. First we assume f is transitive. Let U, V be given. Let $x_n = \phi_n(x_0)$ be a dense orbit. Then there is $n_1 < n_2 \in \mathbb{R}$ such that $\phi_{n_1}(x_0) \in U$ and $\phi_{n_2}(x_0) \in V$ and hence $\phi_{n_2}(x_0) \in \phi_{n_2-n_1}(U) \cap V$ (the intersection is non–empty). This proves the "only if" part of the statement.

Now we assume that part and we will prove that f is transitive (the "if" part). Let $\{x_l\}$ be a countable dense subset of X. Consider the countable family of open balls $(U_{\frac{1}{m}}(x_l))_{m,l}$ of radius $1/m$ around each element of $\{x_l\}$. We can number all these sets into the sequence (U_n). Let $V_1 = U_1$. By the assumption, there is a n_2 such that $\phi_{n_2}(U_2) \cap V_1 = V_2 \neq \varnothing$. Using V_2 and U_3 we can define in the same way a set V_3. We can continue this process using the induction principle and define the family $\{V\}$ of sets V_n, $n \in \mathbb{N}$ via

$$V_{k+1} = \phi_{n_{k+1}}(U_{k+1}) \cap V_k \neq \varnothing.$$

Then $V_{k+1} \subset V_k$ and there is a point $x \in \bigcap_n V_n$. The orbit of this point meets all the open sets V_n and hence is dense. □

Definition 3.6. A system is called **minimal** if all orbits are dense. In this case the phase space does not contain any proper closed invariant set.

The next theorem shows that every system contains a minimal subsystem. Its proof uses elements of set theory glimpsed at in the Appendix.

Theorem 3.6. [**KH96**, 3.3.6 p.130] *Let $f\colon X \to X$ be a continuous function governing a discrete dynamical system on a compact metric space. Then it contains a minimal invariant subset.*

Proof. Let $X \neq \varnothing$ be compact and metric. Then it is also a topological space with a countable basis of open sets. Also, X itself is compact and invariant. As a consequence of Definition 3.6, if X is not minimal, then it contains a proper, closed, non–empty invariant set K_0. In the same way, if K_0 is not minimal, then it must contain a proper closed invariant subset K_1. If this set is not minimal we can continue this process and find K_2. By transfinite induction we get a chain

$$K_0 \supset K_1 \supset K_2 \supset \cdots \supset K_\alpha \supset \cdots$$

of compact sets. The intuition of the proof goes by showing that such a chain cannot be continued "forever". It has to come to a stop on a minimal compact set.

For any ordinal α we find an open set U_n from the basis of the topology with $K_\alpha \supset U_n \supset K_{\alpha+1}$. Since we have a countable basis of the topology the above chain must stabilise at a **countable** ordinal α. But then $\bigcap_{j=0}^{\alpha} K_j$ is non–empty, compact, invariant and minimal since it does not contain any proper closed subset (it may consist of just one point). Actually, we do not need the countability of the basis to derive the stabilisation. □

Remark 3.8. Minimality implies transitivity. The above example of a **rotation** dynamical system on the Torus is minimal (transitive) if and only if a is irrational.

Definition 3.7. A system is called **mixing** if for any two open sets V, U there is a number $N = N(U, V) \in \mathbb{R}$ such that $\phi_n(U) \cap V \neq \varnothing$ for all $n > N$.

Remark 3.9. It is not hard to prove that mixing implies transitivity (use Proposition 3.2).

Example 3.3. The irrational rotation is not mixing. To analyse this statement it is sufficient to consider the induced map f on the unit interval obtained from the Poincaré map. Let $d = \|a\|_{\mathbb{S}^1} = \min\{a, 1-a\}$ and $U = V = (0, d/4)$. Then the property

$$f^n(U) \cap V = (na, na + d/4) \cap (0, d/4) \neq \varnothing$$

implies

$$f^{n+1}(U) \cap V = (na + a, na + a + d/4) \cap (0, d/4) = \varnothing.$$

Remark 3.10. We have seen that in this Example we could reduce our considerations from a flow to a map on a lower-dimensional space. This can be done in quite generality as we will see later. This is one reason why time–discrete systems are important. It is also possible to construct from any time–discrete model a continuous time model. Most of the results of this section for time–continuous systems can be restated for time–discrete systems without major differences.

We can now ask ourselves whether it is typical that all trajectories are regular. The next Theorem indicates that this depends on topological facts.

Theorem 3.7 (Theorem of the Hedgehog). [**EG79**] *Let $M = \mathbb{S}^2$, the surface of a sphere, and v a continuous vector field on the sphere. Then v has at least one singular point, i.e., there exists $x_0 \in \mathbb{S}^2$ with $v(x_0) = 0$.*

We defer the proof to Appendix A.2.3.2. Note that a consequence of this Theorem is that any differential equation on the sphere has at least one **stationary solution** (constant orbit, case 1 of Theorem 3.5).

Remark 3.11. The above theorem shows that global properties imply the existence of stationary points. We will see later that also local properties imply the existence of **stable** stationary trajectories.

3.6.1. Limit Sets. [**GH86**] In the theory of Dynamical Systems one is interested in asymptotic behaviour. Let us specify what we mean by this.

Definition 3.8. Let $x \in X$ a point and $\phi_t \colon X \to X$ be a flow. The ω–**limit set** of the trajectory through x is defined as the set of all accumulation points in the future:

$$\omega(x) = \omega_\phi(x) := \{y \in X \,:\, \exists t_n \to +\infty \text{ with } \phi_{t_n}(x) \to y\}\,.$$

The α–**limit set** of the trajectory through x is defined as the set of all accumulation points in the past:

$$\alpha(x) = \alpha_\phi(x) := \{y \in X \,:\, \exists t_n \to -\infty \text{ with } \phi_{t_n}(x) \to y\}\,.$$

We have,

Lemma 3.1. *If $y \in \{\phi_t(x)\}$ then $\omega(y) = \omega(x)$. Also $\omega_{\phi_t}(x) = \alpha_{\psi_t}(x)$ where $\psi_t = \phi_{-t}$.*

Proof. If $y \in \{\phi_t(x)\}$ then there is a time s such that $y = \phi_s(x)$. Since

$$\lim_{n\to+\infty} \phi_{t_n-s}(y) = \lim_{n\to+\infty} \phi_{t_n-s}\left(\phi_s(x)\right) = \lim_{n\to+\infty} \phi_{t_n}(x)$$

the first statement is proved. The second statement amounts to a interchange of names between positive and negative times, flows and limits. □

Let us identify with own names some special limit sets:

$$\mathfrak{O} := \bigcup_{x\in X} \omega(x) \quad \mathfrak{A} := \bigcup_{x\in X} \alpha(x) \quad L := \mathfrak{O} \cup \mathfrak{A}$$

We call $\mathfrak{O}$ ($\mathfrak{A}$) the ω–limit set (α–limit set) of the flow on X. The next statement is general for all continuous flows.

Theorem 3.8 (Properties of ω-limit). [**Ver90**, p.41] *Let X be compact and $\phi_t \colon X \to X$ be a continuous flow. Then for all $x \in X$ we have*

1. *$\omega(x) \neq \varnothing$,*
2. *$\omega(x) = \overline{\omega(x)}$, i.e., the ω–limit set is closed,*

3. *the ω–limit set is invariant, i.e., $\phi_t(\omega(x)) = \omega(x)$,*
4. *the ω–limit set is connected.*

Remark 3.12. As examples below will show, compactness is essential for the first and the last statement.

Proof. By compactness any sequence $\phi_{t_n}(x) \in X$ $(t_n \to +\infty)$ has an accumulation point. Such a point is by definition in the ω–limit set of x.

For the second statement we will show that the complement of $\omega(x)$ is an open set, i.e., that for each point y in this complement, a whole open neighbourhood $U_\epsilon(y)$ also belongs to the complement.

If a point $y \in X \setminus \omega(x)$, then no sequence $(\phi_{t_n}(x))_{t_n \to +\infty}$ can converge to y. Hence there is a number $\epsilon > 0$ such that

$$U_\epsilon(y) \cap \{\phi_t(x)\}_{t \geq 0} = \varnothing.$$

But then $\omega(x) \cap U_\epsilon(y) = \varnothing$.

Let us address the third statement. Let $y \in \omega(x)$, $\phi_{t_n}(x) \to y$ $(t_n \to +\infty)$ and $z \in \phi_t(y)$. In other words, we take a point y in the ω–limit set of x, a sequence in the trajectory of x tending to y and a point z in the trajectory of y. Note that the sequence $\phi_{t_n+t}(x)$ also has a limit belonging to $\omega(x)$. Then

$$\phi_{t_n+t}(x) = \phi_t\left(\phi_{t_n}(x)\right) \to \phi_t(y) = z.$$

Hence, $z \in \omega(x)$ This means that $\phi_t(\omega(x)) \subset \omega(x)$. Now we will show the reversed inclusion. If $w \in \omega(x)$ there is a sequence of times $s_n \to +\infty$ such that $\phi_{s_n}(x) \to w$. Hence, for some time s, we have

$$\phi_{s_n-s}(x) = \phi_{-s}\left(\phi_{s_n}(x)\right) \to \phi_{-s}(w) = u \in \omega(x),$$

since the left hand side of the previous line also tends to the limit set. This means that $w = \phi_s(u) \in \phi_s(\omega(x))$, so we proved that $\omega(x) \subset \phi_t(\omega(x))$, and therefore both sets are equal, i.e., the limit set is an invariant set.

Finally, let us assume that the ω–limit set of a point x is not connected. This means it can be "divided into two parts": there are open sets V_1, V_2 in X such that

$$\omega(x) \cap V_i \neq \varnothing, \quad \omega(x) \subset V_1 \cup V_2, \quad \overline{V_1} \cap \overline{V_2} = \varnothing.$$

This is, the closure of the open sets is disjoint, the limit set is completely contained in the union of both open sets and has nonempty intersection with both. If there is a time T such that $\phi_t(x) \in V_1$ for all $t > T$ then $V_2 \cap \omega(x) = \varnothing$ since all sub-sequences of points in V_1 that have limit, must have their limit within the closure $\overline{V_1}$. Hence, if the assumption holds, the trajectory has to "oscillate" between V_1 and V_2 for arbitrarily large times, it cannot "settle down" to one of the sets. In addition, since $\overline{V_1} \cap \overline{V_2} = \varnothing$, both sets are "separated", not even their closures have points in common, so there exists an $\epsilon > 0$ such that the distance between the sets (the infimum of distances between some point in one set and some point in the other set) $d(V_1, V_2) = \epsilon > 0$. However, since the trajectory is connected, there will be portions of it outside both sets and these portions will occur repeatedly for arbitrarily large times. Hence, there must be a sequence of moments $t_n \to +\infty$ such that $d(\phi_{t_n}(x), V_i) > \frac{\epsilon}{3}$. Since phase space is compact there is a convergent sub-sequence t_{n_k} such that

$$u = \lim_{k\to\infty} \phi_{t_{n_k}}(x) \notin V_1 \cup V_2.$$

Because of its definition, u belongs to $\omega(x)$ and the limit set must have points outside both open sets. This contradicts our assumption, hence the limit set is connected. □

Let us consider now some examples of non–compact phase spaces and realise that when the compactness hypothesis does not hold, then the statements of the previous Theorem need not hold either.

Example 3.4. The parallel vector field $v(x, y) = (1, 0)$ on the plane has straight parallel lines as trajectories. They only "accumulate at infinity". Therefore the ω–limit set of any point is empty.

Example 3.5. The ω–limit set of the flow in Figure 3.6 is not connected.

$$\omega(x) = L_1 \cup L_2 \qquad \alpha(x) = y$$

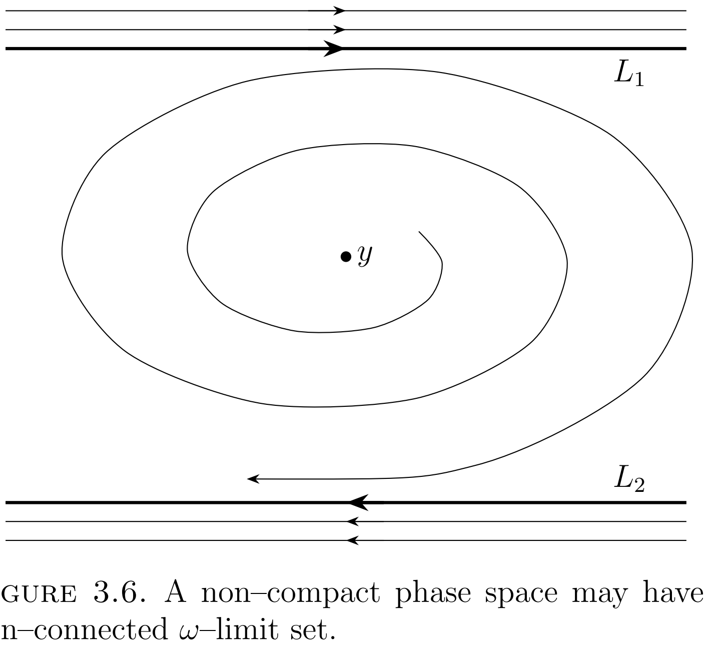


Figure 3.6. A non–compact phase space may have a non–connected ω–limit set.

Definition 3.9. A closed simply connected set D such that $\phi_t(D) \subset D$ for all positive times is called a ***trapping region***.

REMARK 3.13. *Simply connected* means that (a) Any two points of the set can be joined by an arc lying completely within the set and (b) Any closed loop in the set can be shrinked to a point.

REMARK 3.14. A sufficient condition for D being a trapping region is that the vector field is directed everywhere inward on the boundary of D.

If a dynamical system has a compact trapping region D, then there exists a neighbourhood U of D such that for all $x \in U$, $\omega(x) \in D$.

DEFINITION 3.10. A closed (positively) invariant set A such that there exists a neighbourhood V of A (V is open and $A \subset V$) such that for every $x \in V$ and $t \geq 0$, $\phi_t(x) \in V$ and also $\phi_t(x) \to A$ as $t \to \infty$ is called an *attracting set.*

DEFINITION 3.11. An *attractor* is a transitive attracting set.

REMARK 3.15. For a compact phase space, an attractor coincides with the ω–limit set. For non–compact phase spaces, the ω–limit set may be non–connected and there may exist many attractors. Each attractor however will coincide with some connected component of the ω–limit set.

3.7. Exercises

EXERCISE 3.6. Show that every bounded trapping region D such that the vector field points inwards on the boundary of D contains an attracting set.

EXERCISE 3.7. Consider the dynamical system defined on $\mathbb{R}^2$,

$$\begin{cases} \dot{y} &= -y \\ \dot{x} &= x(1-x)(1+x). \end{cases}$$

- List as many invariant sets you can.
- Show that there exist exactly three trajectories of type 1, i.e., $x(t) \equiv$ *const.*
- Find a trapping region. There are many choices.
- Show that the set $\{x \in [-1,1], y = 0\}$ is positively invariant, and moreover an attracting set.
- Find the ω–limit set of the system and check if it is connected (note that phase space is not compact).

3.8. Poincaré–Bendixson Theory

3.8.1. The Theorem of Poincaré–Bendixson. [**KH96**, p.452] This Section is devoted to one of the most beautiful theorems in Dynamical Systems. It is actually a topological theorem like the Theorem of the Hedgehog. It is valid on the sphere and other 2–d spaces but we will recognise later that it is not valid on the torus by recalling Example 3.2. To arrive to the Theorem we will need to prepare the terrain by building some foundational structures.

We consider an arbitrary Lipschitz vector field v on the sphere $\mathbb{S}^2$ with the assumption that the number of its singular points is finite:

$$\#\left\{x \in \mathbb{S}^2 \, : \, v(x) = 0\right\} < \infty.$$

This is not a big restriction since "typical" vector fields have only a finite number of singular points. It is automatically fulfilled when the vector field is real analytic.

Remark 3.16. A *(real) analytic* function is a (real-valued) function that is locally given by a convergent power series.

Lemma 3.2. *Let Σ be a curve transversal to the vector field v, i.e., the tangent vector (of the curve) $T_x\Sigma$ at a point $x \in \Sigma$ is not parallel to $v(x)$, for all $x \in \Sigma$. Consider now all forward intersections of the trajectory of x with Σ, with their time ordering. In other words, let $\{\phi_t(x)\}_{t\geq 0} \cap \Sigma = \{x_i\}$, $i \in \mathbb{N}$ where for $x_k = \phi_{t_k}(x)$ we have $t_k < t_{k+1}$. Then the sequence $\{x_i\}$ is monotone on Σ, i.e.,*

$$x_k \in [x_{k-1}, x_{k+1}].$$

Proof. We note on passing that transversality implies that we have a discrete set of times when the trajectory of x intersects Σ.

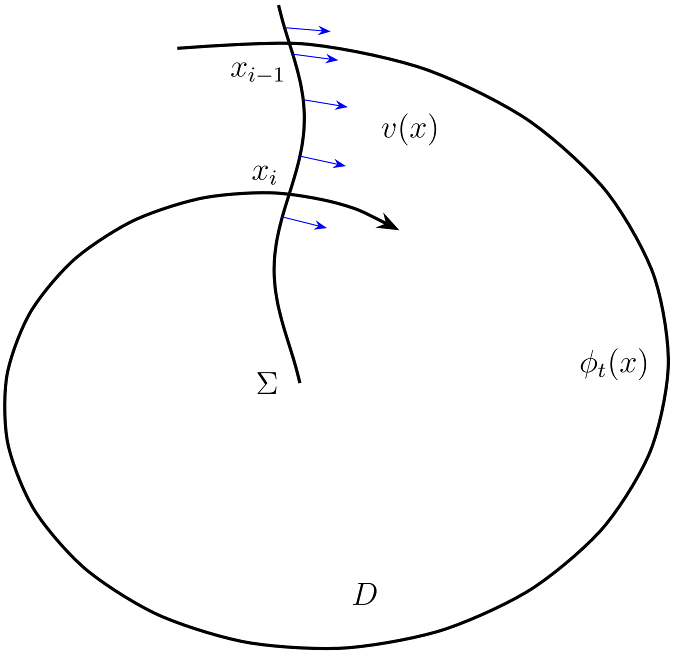


Figure 3.7. The curve Σ on the sphere, a trajectory $\phi_t(x)$ and some vectors from the vector field $v(x)$.

Assume the sequence of forward intersections meets at least two distinct points on Σ. Since the vector field is transverse to the curve Σ, the vectors $v(x)$, $x \in \Sigma$, they all "point" in the same direction relative to Σ, otherwise by continuity and the mean value theorem, some vector $v(x)$ would be tangent to Σ. Hence, the trajectory $\phi_t(x)$ crosses Σ always in the same direction (say "from left to right"). Let us consider two consecutive crossings, this is depicted in Figure 3.7.

By the Picard–Lindelöf Theorem, the trajectory cannot cross itself and therefore it cannot exit the region D which is bounded by the portion of the trajectory between crossings, i.e., $(\phi_t(x))_{t_{i-1} \leq t \leq t_i}$, and by the segment $[x_{i-1}, x_i] \subset \Sigma$ (the vector field on Σ points inwards). Therefore x_{i+1} must be "below" x_i along Σ and hence, $x_i \in [x_{i-1}, x_{i+1}]$.

Remark 3.17. We assumed in the picture that the vector field points inward. But actually this is not a restriction. If it points outwards we can consider the region $\mathbb{S}^2 \setminus D$ which then has the same picture as Figure 3.7.

□

Remark 3.18. Note that if the sequence of forward intersections $\{x_i\}$ consists of only one point, visited infinitely many times, the trajectory $\phi_t(x)$ is periodic. If the point is visited only once, $\phi_t(x)$ never returns to Σ for $t > 0$.

Corollary 3.2. *The number of points from the ω–limit set lying on Σ is at most one:*

$$\#\{x \,:\, \omega(x) \in \Sigma\} \leq 1.$$

Proof. To prove the Corollary we will show that if this number of points is not zero, then it is exactly one. Let $y \in \omega(x) \cap \Sigma$. Since the vector field is transversal to Σ we have $v(y) \neq 0$. By Theorem 3.3, the Straightening–out Theorem, there is a small neighbourhood U of y such that the vector field is horizontal (up to a coordinate transformation). Hence, any trajectory entering U must cross Σ. Now if $y = \lim_{n \to +\infty} \phi_{t_n}(x)$ there are points $x_k \in \{\phi_t(x)\}_{t \geq 0} \cap \Sigma$ close to y. These points build a sequence as that of Lemma 3.2. Therefore, any point $y \in \omega(x) \cap \Sigma$ is an accumulation point of the sequence $\{x_i\}$. By the previous Lemma, this sequence is monotone and hence it has at most one limit point. □

This limit point cannot be a singular point since the vector field is transverse to Σ.

Lemma 3.3. *If $\omega(x)$ does not contain a singular point, then $\omega(x)$ is a periodic trajectory. Moreover, for any y close to x we have*

$$\omega(x) = \omega(y).$$

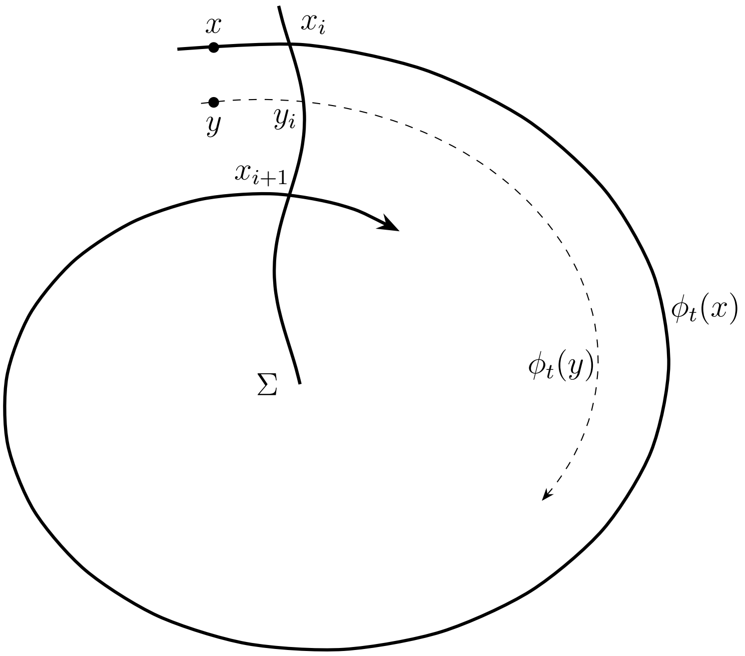


FIGURE 3.8. The points x, y, their trajectories, the curve Σ and the limit set $\omega(x)$.

PROOF. Let $z \in \omega(x)$ and $w \in \omega(z) \subset \omega(x)$. Since w is not a singular point, we can find a curve Σ through w and transversal to the vector field v. Let $\lim_{n\to+\infty} \phi_{t_n}(z) = w$ where t_n is chosen in such a way that $z_n = \phi_{t_n}(z) \in \Sigma$. But all the points $z_n \in \omega(x)$ by invariance. So Corollary 3.2 implies that

$$z_n = z_m = w \qquad \text{for all } n, m \in \mathbb{N}.$$

Hence the trajectory $\{\phi_t(z)\}$ is periodic by Theorem 3.5. Consider now another point $z' \in \omega(x)$, then $\{\phi_t(z')\}$ is also a periodic trajectory within $\omega(x)$. But the trajectory $\{\phi_t(x)\}$ cannot approach two different periodic orbits as $t \to +\infty$. Hence, $\omega(x)$ is a single periodic trajectory. This ends the proof of the first statement.

For the second statement, see Figure 3.8. Let y_i be the intersections $\{\phi_t(y)\} \cap \Sigma$. Since $\{\phi_t(y)\}$ does not intersect $\{\phi_t(x)\}$ we have $y_i \in [x_i, x_{i+1}]$. This implies

$$\lim_{i\to\infty} y_i = \lim_{i\to\infty} x_i$$

and hence the ω–limit sets of x and y coincide. □

Let us now deal with the case in which the ω–limit set contains singular points.

LEMMA 3.4. *Let $z_1 \neq z_2 \in \omega(x)$ and $v(z_1) = v(z_2) = 0$. Then there is at most one trajectory $\{\phi_t(y)\} \subset \omega(x)$ with*

$$\omega(y) = z_1 \qquad \alpha(y) = z_2.$$

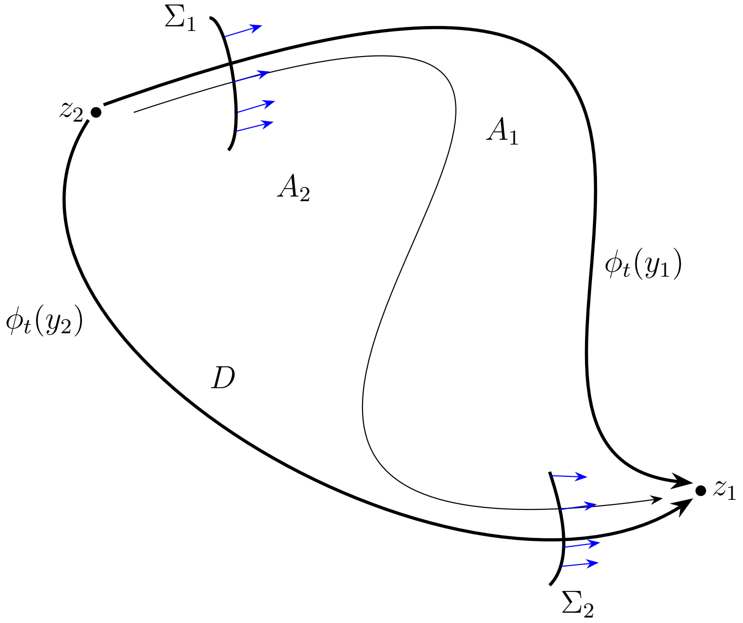


FIGURE 3.9. Hypothetical picture of an assumption contrary to Lemma 3.4.

PROOF. Let us assume the contrary, i.e., that we have two disjoint trajectories, based on the points y_1 and y_2, within $\omega(x)$. Refer to Figure 3.9 to visualise the different elements of the proof. We have $\{\phi_t(y_1)\} \cap \{\phi_t(y_2)\} = \varnothing$, $\{\phi_t(y_1)\} \cup \{\phi_t(y_2)\} \subset \omega(x)$ and $\alpha(y_i) = z_2$ and $\omega(y_i) = z_1$. Moreover, the vector field does not vanish on any point of $\{\phi_t(y_1)\}$ or $\{\phi_t(y_2)\}$. Therefore we always can find transversal curves Σ_1 and Σ_2 to the trajectories $\{\phi_t(y_1)\}$ and $\{\phi_t(y_2)\}$. The two trajectories $\{\phi_t(y_i)\}$; $i = 1, 2$ and the two points z_1, z_2 bound a region D. We may assume that $\{\phi_t(x)\} \subset D$ (otherwise we consider $\mathbb{S}^2 \setminus D$). The trajectory $\{\phi_t(x)\}$ must accumulate on the boundary of D. Moreover, the trajectory $\{\phi_t(x)\}$ must also intersect Σ_1 and Σ_2 transversally (since it comes arbitrarily close to the trajectories $\{\phi_t(y_i)\}$; $i = 1, 2$). Hence, the region D is divided into two parts $D = A_1 \cup A_2$ as in Figure 3.9. In such a case, the trajectory $\{\phi_t(x)\}$ enters A_1 and cannot leave it anymore. This contradicts the assumption that both trajectories $\{\phi_t(y_1)\}$ and $\{\phi_t(y_2)\}$ are in the ω–limit set of x (both parts A_1 and A_2 cannot be in the ω–limit set of x). □

Now we arrived finally at the theorem of Poincaré and Bendixson.

Theorem 3.9 (Poincaré–Bendixson). [**Ben01**] *Let v be a vector field on $\mathbb{S}^2$ with $\#\{x : v(x) = 0\} < \infty$. Then for $x \in \mathbb{S}^2$ we have one of the following possibilities:*

1. *$\omega(x)$ is a singular point*
2. *$\omega(x)$ is a periodic trajectory*
3. *$\omega(x)$ is the union of a finite number of singular points $x_1, \cdots, x_n$ and connecting regular trajectories $\{\phi_t(y_{ij})\}$ with*

$$\alpha(y_{ij}) = x_i \qquad \omega(y_{ij}) = x_j.$$

Moreover, for $i \neq j$ the curve $\{\phi_t(y_{ij})\}$ is unique.

Proof. If $\omega(x)$ does not contain a singular point we are in case 2 by Lemma 3.3.

If $\omega(x)$ does not contain any regular point it must be a single singular point (it is connected and the number of singular points is finite).

If there is a regular trajectory $\{\phi_t(y)\} \subset \omega(x)$ which is not periodic then $\omega(y)$ consists of a single point $\{z\}$. This is because in this case, for any transversal curve Σ we have by Lemma 3.2 that $\#\{\{\phi_t(y)\} \cap \Sigma\} \leq 1$. Moreover, z is a singular point, because if it were regular we could choose Σ such that $z \in \Sigma$. But this means that $\{\phi_t(y)\}$ intersects Σ infinitely often in the unique limit point. Therefore $\omega(y)$ is a periodic trajectory, which is a contradiction. The same holds for $\alpha(x)$. This is precisely case 3 and the proof is finished. □

Example 3.6. The three cases of the Poincaré–Bendixson Theorem are illustrated in Figures 3.10 and 3.11. Note the difference between the direction arrows in Figures 3.9 and 3.10.

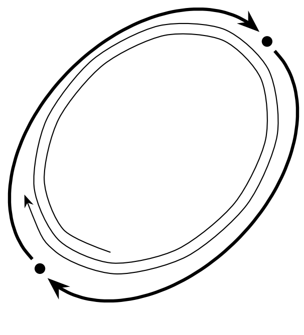


Figure 3.10. Case 3 of Poincaré–Bendixson Theorem.

Remark 3.19. A trajectory connecting two different singular points is called a **heteroclinic orbit** (see Figure 3.10). If both singular points coincide, the trajectory is called a **homoclinic orbit** (see Figure 3.12).

Remark 3.20. The proof of the Theorem builds on the construction of an open connected set, using a portion of a trajectory and the curve Σ, see Figure 3.7. Such construction is essentially a Jordan region on

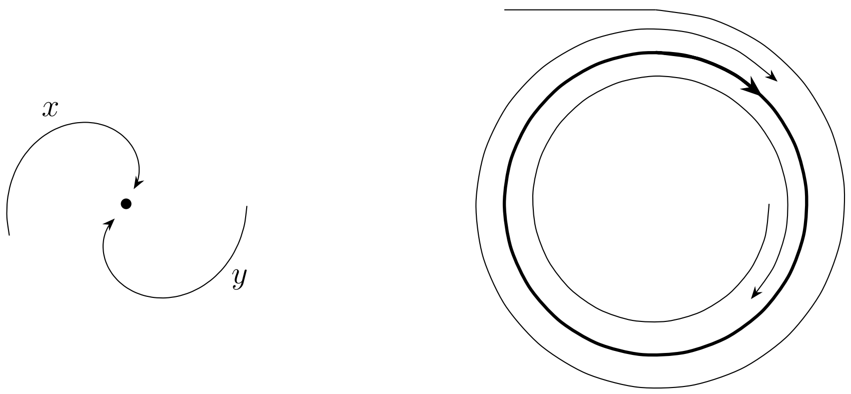


FIGURE 3.11. Cases 1 and 2 of Poincaré–Bendixson Theorem

the plane, hence, the Theorem holds as well within positively invariant Jordan regions of $\mathbb{R}^2$. On the other hand, in more complicated spaces such as the surface of a Torus, we noticed in Example 3.2 that the ω–limit for the case $a \notin \mathbb{Q}$ is the whole torus.

EXAMPLE 3.7. The following vector fields on the sphere $\mathbb{S}^2$ contain homoclinic and heteroclinic orbits and illustrate also Case 3 of Poincaré-Bendixson Theorem. The left picture has only one fixed point

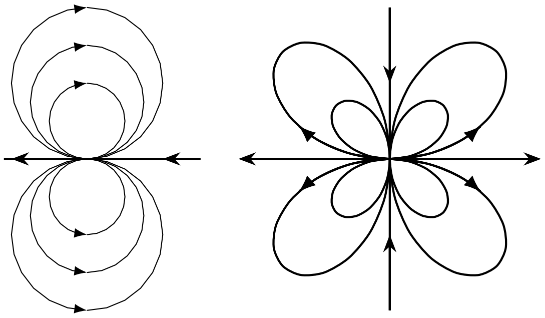


FIGURE 3.12. More examples of Case 3 of Poincaré-Bendixson Theorem.

(at the origin, which we can identify with the North pole) and the open orbit closes itself along a meridian. Through a stereographic projection the orbits shown correspond to parallel (horizontal) trajectories on a plane. The right picture is completed with a fixed point of saddle type at the South pole. In both cases it holds that $\alpha(x) = \omega(x)$. We will revisit these examples as portraits of planar vector fields in the Appendix (see Section A.2.3).

3.9. Summary

We have seen that locally only singular points are interesting, because of the Straightening–out Theorem 3.3. We will later in the book work on the qualitative features of the trajectories in the vicinity of singular points. We have also developed structures to deal with asymptotic behaviour (limit sets, trapping regions, etc.). Finally with the Poincaré–Bendixson Theorem 3.9 we realise that two-dimensional flows on relatively "simple" spaces have only simple asymptotic behaviour.

3.10. The Lorenz Equations

A full-featured dynamical system that nevertheless has a rather simple appearance, with only two (quadratic) nonlinear terms is given by the Lorenz equations[**Lor63**]:

$$\begin{aligned} \dot{x} &= \sigma(y - x) \\ \dot{y} &= rx - y - xz \\ \dot{z} &= xy - bz. \end{aligned}$$

This system originated when modeling the Bénard experiment, which studies the dynamics of a gas fluid in a small chamber (aspect ratio $b = 8/3$) with constant temperatures $T_0 > T_1$ at the lower and upper plates respectively, under the influence of gravity. x, y, z represent amplitudes in a simplified, truncated, model solution as well as the deviation from a linear vertical temperature profile. The fluid parameters are given by σ, the Prandtl number (often fixed at the value 10 for numerical illustrations) and r, the Rayleigh number. Usually, numerical and other experiments are done varying r in $(0, 30)$. At this moment in the book we should be able to identify singular points and realise that the number and location of these points change with r. Advanced students may succeed in finding a trapping region.

This model was conceived along the lines devised by Bénard for a laboratory experiment with typical sizes of the order of millimetres. Some of his observations occur in the model as well. It has been claimed that the model could be a metaphor for atmospheric behaviour, but it is too crude to describe the complicated atmospheric dynamics.

We mention these equations here because of their historical relevance for numerical explorations. The model was studied numerically in 1963 by Edward Lorenz, when computers were several orders of magnitude slower than today. The anecdote tells that two runs with initial conditions differing in the fourth significant figure, soon gave rise to completely different dynamics, thus awakening the attention of scientists towards *sensitivity to initial conditions*[**SNM96**], which we will discuss on a later Chapter.

3.11. Exercises

Exercise 3.8. Compute all singular points and (at least portions of) $\omega(x)$ for the system:

$$\begin{aligned} \dot{x} &= ax - y - x(x^2 + y^2) \\ \dot{y} &= x + ay - y(x^2 + y^2), \end{aligned}$$

for the cases $a > 0$ and $a < 0$.

Exercise 3.9. Compute all singular points and their dependence with r for the Lorenz equations.

Exercise 3.10. Find a trapping region for the Lorenz equations for the case $0 < r < 1$. Hint: Try with a region shaped as an ellipsoid. Choose the constants so that the vector field points inwards on the surface of the ellipsoid.

Exercise 3.11. Find a trapping region for the Lorenz equations that holds for any (non-negative) r. Hint: Try again with a region shaped as an ellipsoid. Now it has to be a much larger ellipsoid than in the previous exercise and its "centrum" will be shifted along the z-axis.

CHAPTER 4

Discrete Dynamical Systems

We investigate in this Chapter the important issue of discrete time dynamical systems. Apart from the natural connection between flows and maps, discrete time dynamical systems are relevant in many applications.

4.1. Flows versus Maps

There are several ways to obtain a map f from a flow ϕ_t. We list two important cases here:

4.1.1. The Time–one Map. [KH96, SNM96]

Given a flow $\phi_t(\cdot)\colon \mathbb{R}^d \to \mathbb{R}^d$, we define $f\colon \mathbb{R}^d \to \mathbb{R}^d$ by

$$f = \phi_1.$$

Thus, we obtain a new system, given by f^n; $n \in \mathbb{Z}$. This corresponds to a (time–)discretisation of the original system. Note that since this map is simply the flow of a continuous–time dynamical system recast at discrete time intervals, all properties of continuity and derivability of the flow are automatically inherited by the time–one map. Although time–one maps are very important and indeed useful, this approach has some limitations.

1. f is always **homotopic to the identity**, i.e., there is a continuous map $F : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ such that for each $t \in [0,1]$ the map $F(\cdot,t)\colon \mathbb{R}^d \to \mathbb{R}^d$ is a homeomorphism satisfying $F(\cdot,0) = id$ and $F(\cdot,1) = f$ (two objects are homotopic if they can be connected by a continuous family of objects of the same kind in the way described here)[1]. The map F can be chosen to be $\phi_t(\cdot)$. This means that the class of time-1-maps is rather restrictive. F.e. let $f\colon \mathbb{T}^2 \to \mathbb{T}^2$ be the "reflection" $(x,y) \to (y,x)$. This map exchanges the two circles $x = 0$ and $y = 0$. But these two circles cannot be deformed one into another by a continuous deformation. Hence this map cannot be the time–one map of a flow.
2. Let $f(x) \neq x$ but $f^n(x) = x$ for some $n > 0$. Then for all $t \in \mathbb{R}$ we have

$$f^n\left(\phi_t(x)\right) = \phi_{n+t}(x) = \phi_t\left(\phi_n(x)\right) = \phi_t\left(f^n(x)\right) = \phi_t(x).$$

[1] Whenever the flow ϕ_t is smoother than just C^0, f will inherit its smoothness and F can be actually "upgraded" to a diffeomorphism.

> This means all the points $\phi_t(x)$ are periodic. In particular we cannot have isolated periodic orbits!

There is also a need for studying discrete dynamical systems of a fundamentally different nature than just time–one maps.

4.1.2. Orientation Preserving Maps.

Definition 4.1. For any simple closed curve (loop) on a surface (2–d) in $\mathbb{R}^3$ consider three positively oriented orthogonal vectors at each point of the loop where v_1 is along the curve, v_2 on the surface and v_3 along the surface normal. The surface is called **orientable**, if when moving one turn along the loop the vectors v_2, v_3 return to their starting orientations (i.e., they are not the mirror image of the original pair). A well-known non-orientable surface is the Möbius band [**Emm80**].

Orientation is not confined to 2–d spaces, the choice of basis vectors in $\mathbb{R}^n$ defines an orientation in that space. Maps defined on orientable spaces may preserve the orientation i.e., the induced map on the tangent space has determinant one. In 2-d again, the image of two tangent orthogonal vectors is not the mirror image of the starting pair. In such a case they are called **orientation preserving**. Both the previous example and the next to come are standard examples of orientation preserving maps.

4.1.3. The Poincaré Map. [GH86, SNM96]

On a (small) transversal section Σ to a periodic trajectory γ we can define the following map:

$$f = f_\Sigma \colon U \subset \Sigma \to \Sigma,$$

where $f_\Sigma(x_0)$ is the first intersection point (for $t > 0$) of the trajectory $x(x_0, t)$, $x_0 \in \Sigma$ with Σ (see Figure 4.1). This map is also called the *First-return Map.* Σ is also called the *Poincaré section* or control section.

In order to have a Poincaré map the flow of the original system has to fulfill some basic requirements valid for each point $x \in \Sigma$:

1. Σ is orientable and transverse to the flow, i.e., there exists a normal vector $n(x)$ and $n(x) \cdot v(x) > 0$.
2. The flow has to return to Σ, i.e., there exists a positive $t_0(x)$ such that $\phi_{t_0(x)}(x) \in \Sigma$ (this guarantees that returns will occur for all real times).
3. $v(x) \neq 0$ on Σ (otherwise no return is possible).

By the differentiable dependence on initial conditions this map is well-defined and as smooth as the original flow. Moreover, the Poincaré map is invertible (suitably adjusting the domain), since the Picard–Lindelöf Theorem assures that each point in phase-space belongs to

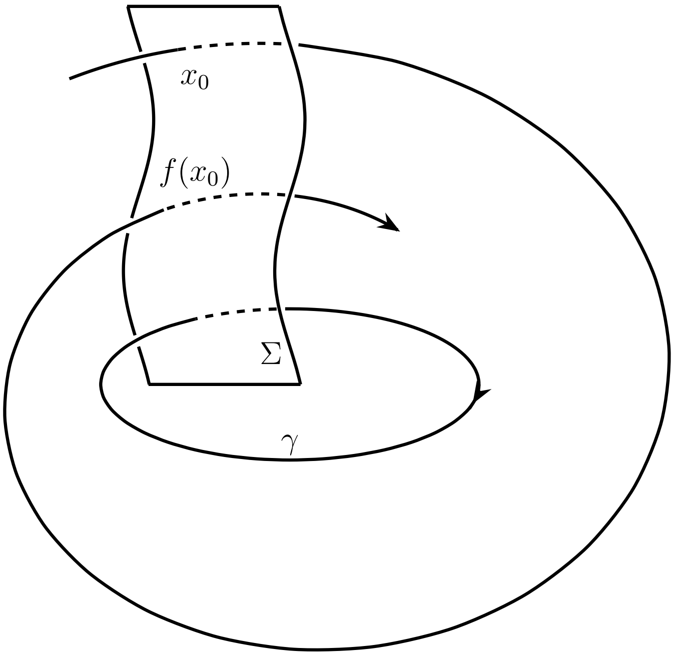


FIGURE 4.1. The Poincaré First-return Map.

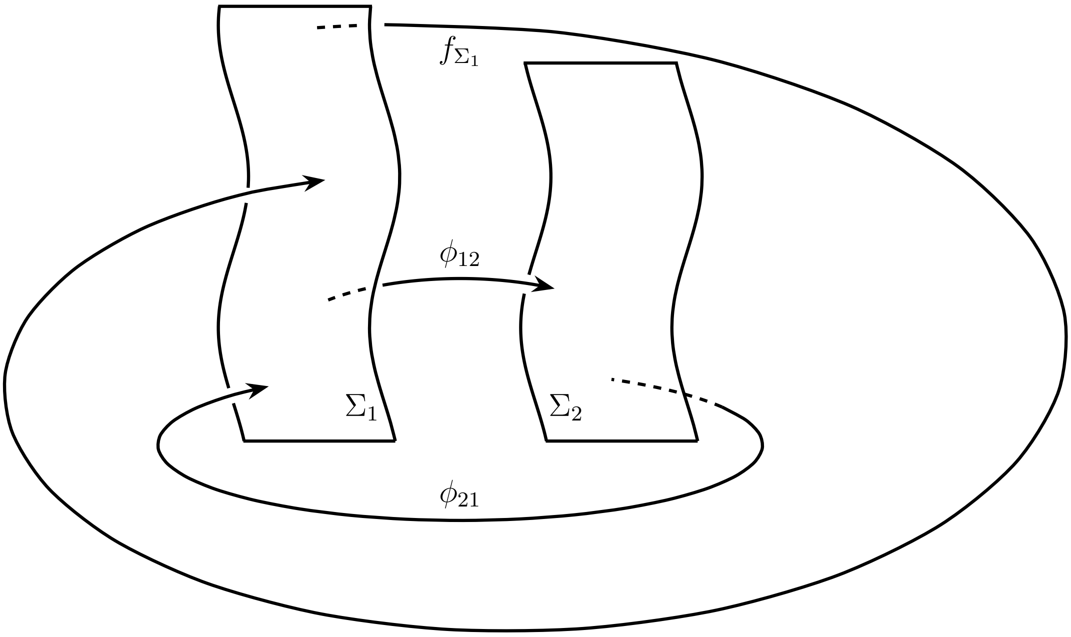


FIGURE 4.2. Two equivalent control sections connected by diffeomorphisms.

only one trajectory. We note that the dimension of Σ is smaller than that of the original phase space: $\dim \Sigma = d - 1$.

The new dynamical system generated by the Poincaré map f inherits many properties of the flow. It is widely used to understand the flow. Intuitively, the map factors out the "uninteresting" flow-direction where the dynamics of individual points consists only on transport from Σ back to Σ again (compare with the Straightening–out theorem).

Let us see how this map depends on the transversal section Σ. Assume we have two properly chosen transversal sections Σ_1 and Σ_2. Then we can define two maps $\phi_{12}\colon \Sigma_1 \to \Sigma_2$ and $\phi_{21}\colon \Sigma_2 \to \Sigma_1$, by sliding along a trajectory $x(x_0, t)$ for increasing times, where $x_0 \in \Sigma_1$ or $x_0 \in \Sigma_2$, respectively until the other transversal section is met. The maps are invertible with a proper choice of domains. Moreover, by the smooth dependence on initial conditions both maps are diffeomorphisms. Now we can compare the Poincaré maps $f_{\Sigma_1} = \phi_{21} \circ \phi_{12}$ and $f_{\Sigma_2} = \phi_{12} \circ \phi_{21}$. We have

$$f_{\Sigma_2} = \phi_{12} \circ f_{\Sigma_1} \circ \phi_{12}^{-1}.$$

Therefore the Poincaré map changes only by a smooth diffeomorphism, essentially a change of coordinates, and hence they all are (smoothly) equivalent (see Figure 4.2).

4.1.4. Suspensions. We can associate to a given map a flow which inherits many properties of the map. This flow is called the **suspension** of $f\colon \mathbb{R}^d \to \mathbb{R}^d$. We consider the larger space $(x, t) \in \mathbb{R}^d \times [0, 1]$ with the vector field $v(x, t) = (0, 1)$. then we identify the "top" and the "bottom" via

$$(x, 1) = (f(x), 0).$$

This gives a "torus-like" manifold with a smooth flow (the identification map is smooth) (see Figure 4.3). Recall that a manifold is a space locally equivalent to Euclidean space.

Recalling Example 3.2, the rotation f on the circle $\mathbb{R}/\mathbb{Z} = \mathbb{R} \mod 1$, mapping $x \to x + a \mod 1$, has as suspension the original linear flow on the torus.

4.1.5. Basic Topological Concepts for Maps. Many concepts defined in the previous chapters can be almost automatically transported from flows to maps, simply by replacing continuous time by discrete time when relevant. Among them, we can count the concepts of trajectory or orbit, invariant set, transitivity, minimality, α– and ω–limit sets, as well as other concepts discussed in the previous Chapter.

Some of these concepts, however, although they can be defined in almost the same way, have different properties in flows and maps. For example, fixed points both in maps and flows may be defined as invariant sets with only one element, i.e., $f^n(x) = x$ for all $n \in \mathbb{Z}$ or $\phi_t(x) = x$ for all $t \in \mathbb{R}$. However, a more functional definition for a map is just $f(x) = x$, whereas in flows they are defined as $\{x\colon v(x) = 0\}$.

To recall more differences, note that trajectories with initial condition x_0 for flows are invariant curves on phase space, while trajectories in maps are discrete sets of points. An invariant curve in a

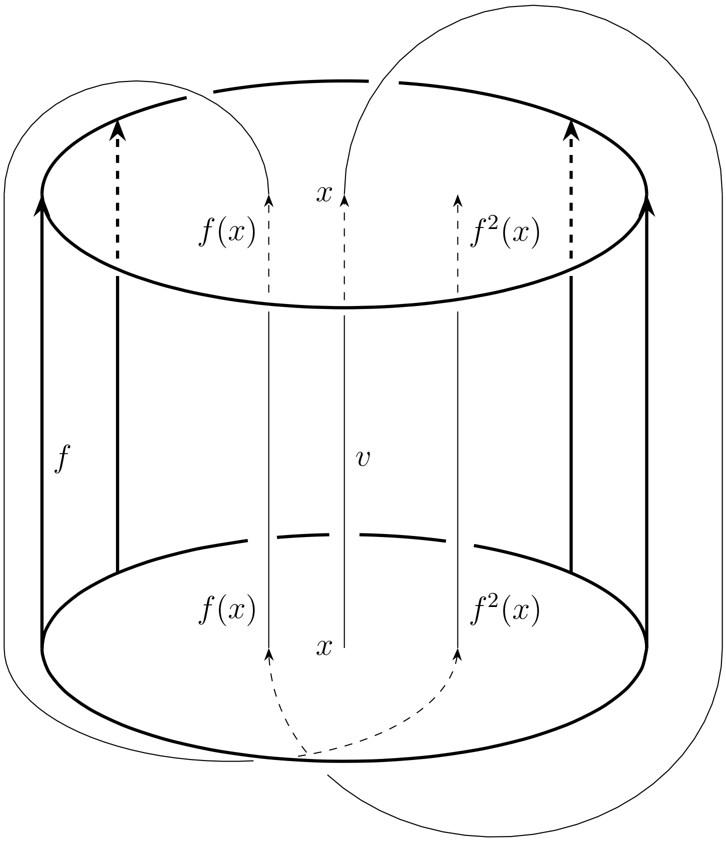


FIGURE 4.3. A suspension of f generates a flow.

map may contain infinitely many trajectories within it. Also, periodic orbits in flows are closed curves whereas periodic orbits in maps are discrete and finite sets of points (if N is the minimal positive integer such that $f^N(x) = x$, the periodic orbit consists of the set of points $\{x, f(x), \cdots, f^{N-1}(x)\}$).

Important subsets of phase space for a map f are $Fix(f) \subset Per(f)$, the set of fixed points and the set of periodic points (of any positive period $N \geq 1$) respectively.

4.2. One-dimensional Maps

We will start with discrete dynamical systems in one dimension. One of the main advantages over higher dimensional systems is that the domain, the real numbers, is an ordered set. More generally, phase space will be either $\mathbb{R}$, $[0,1]$ or $\mathbb{S}^1$ (the unit circle, which can be regarded as the interval $[0,1]$ where we identify the endpoints). First we deal with invertible systems and show that their behaviour is not too complicated.

4.2.1. Invertible Systems on $[0,1]$. [KH96]

An injective smooth system on the interval is given by a strictly monotone function $f\colon [0,1] \to [0,1]$. Since any graph of a continuous function defined on $[0,1]$ must intersect the diagonal any such map has at least one fixed point (see Theorem A.8; recall that on maps a fixed point is such that $x = f(x)$).

THEOREM 4.1. *Let f be an injective system. Then for any forward orbit $\{f^n(x)\}_{n\in\mathbb{N}}$ the ω-limit set is a fixed point.*

PROOF. The set of fixed points Fix(f) for f (i.e., the set of points where the graph of f intersects the diagonal (see Figure 4.4)) is non–empty (see a proof in Appendix A.3.3) and closed. Without loss of generality we may assume that f is strictly monotonically increasing.

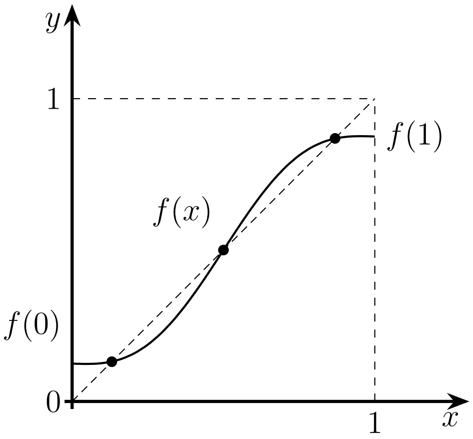


FIGURE 4.4. An injective map of the interval. Dots indicate fixed points

In case Fix$(f) = [0,1]$, i.e., $f = \mathrm{Id}$, we are done. If this is not the case, consider a sub-interval free from fixed points, i.e., let $x \in (a,b) \subset [0,1] \setminus Fix(f)$. Let the interval (a,b) be maximal. Then by maximality both a and b must be fixed points, i.e., $f(a) = a$ and $f(b) = b$. We have by monotonicity and continuity that

$$f((a,b)) = (f(a), f(b)) = (a,b) \subset [a,b].$$

Moreover we have for $x \in (a,b)$ that on the whole sub-interval (a,b) either $f(x) > x$ or $f(x) < x$. Otherwise we would have points $z, y \in (a,b)$ such that $f(z) > z$ and $f(y) < y$ (we may assume $z < y$ without loss of generality). In such a case, for some point $w \in (z,y)$ there would be an intersection of the graph of f with the diagonal (by continuity) and hence another fixed point in (a,b). Let the first case hold, i.e., $f(x) > x$. Then $b > f^n(x) > f^{n-1}(x)$ since $f^n(x) = f(f^{n-1}(x)) \in (a,b) = (f(a), f(b))$ for all n. By the convergence of bounded monotone sequences on a closed interval, there is a point

$$x_\infty = \lim_{n\to\infty} f^n(x) \in [a,b].$$

But we have

$$f(x_\infty) = f\left(\lim_{n\to\infty} f^n(x)\right) = \lim_{n\to\infty} f\left(f^n(x)\right) = \lim_{n\to\infty} f^{n+1}(x) = x_\infty.$$

Therefore x_∞ is a fixed point and must be equal to b, and the ω–limit set of the whole interval $(a, b]$ is the fixed point b. The case $f(x) < x$ gives $x_\infty = a$. In any case, all points that are not fixed points lie in some maximal interval and hence have a fixed point as ω–limit set. □

This theorem and many other related results can be found in the book by De Melo and van Strien [**MvS93**]

Remark 4.1. The set of fixed points might be complicated. There are differentiable functions whose fixed point set can be any compact subset of $[0, 1]$. On the other hand if the function f is real analytic the fixed points must be isolated and hence their number is finite.

4.2.2. Notions of Recurrence. [KH96]

We will see in this section that there is a well-developed theory of maps of the circle. This investigation started with Poincaré. He related arbitrary diffeomorphisms of the circle to the "linear" maps: the **rotations**. Let us start with some definitions.

Definition 4.2. A point x is called **positive (negative) recurrent** if $x \in \omega(x)$ $(x \in \alpha(x))$.

Remark 4.2. All periodic points are recurrent (positive and negative). But there can be more. Any trajectory of the irrational torus flow is recurrent $(\omega(x) = \alpha(x) = \mathbb{T}^2)$.

Proposition 4.1. *Any point in a minimal invariant subset is recurrent.*

Proof. Let $f|_Z \colon Z \to Z$ be minimal. This means that the orbit of any $z \in Z$ is dense in Z, i.e., $\alpha(z) \cup \omega(z) = Z \ni z$ and z is recurrent. □

Definition 4.3. A point x is called **non-wandering** if for each open neighbourhood U of x there is an $n \in \mathbb{N}$ such that

$$f^n(U) \cap U \neq \varnothing.$$

The set of non-wandering points is denoted by Ω.

Theorem 4.2. [**KH96**, 3.3.4 p.129]*]* Ω *is closed, invariant and contains the* ω*– and* α*–limit sets of all trajectories.*

Proof. Let $x_n \in \Omega$ and $x_n \to x$. Let U be an open neighbourhood of x. Then $x_n \in U$ for all large n. Let $x_n \in U$. since $x_n \in \Omega$ there is an $N \in \mathbb{N}$ such that $f^N(U) \cap U \neq \varnothing$. Hence $x \in \Omega$. This proves that Ω is closed since it contains the limits of its convergent sequences.

Let $f(x) \in U$. Then $x \in f^{-1}(U)$ and $f^{-1}(U)$ is open by continuity. If $x \in \Omega$ there is an N such that $f^N\left(f^{-1}(U)\right) \cap f^{-1}(U) \neq \varnothing$. Hence

$$\varnothing \neq f\left(f^N\left(f^{-1}(U)\right) \cap f^{-1}(U)\right) = f^N(U) \cap U$$

and $f(x) \in \Omega$.

Let $x = \lim_{k\to\infty} f^{n_k}(y) \in U$ (this means that x is in the ω–limit of some trajectory). Then for all large k we have $f^{n_k}(y) \in U$. Hence

$$f^{n_{k+1}}(y) \subset U \quad \text{and} \quad f^{n_{k+1}}(y) = f^{n_{k+1}-n_k}(f^{n_k}(y)) \subset f^{n_{k+1}-n_k}(U)$$

and for $N = n_{k+1} - n_k$ we have

$$U \cap f^N(U) \neq \varnothing$$

and $x \in \Omega$. □

Corollary 4.1. *Any recurrent point is contained in Ω since its α– and ω–limit are in Ω.*

Corollary 4.2. *If X is compact and f continuous then $\Omega \neq \varnothing$.*

Proof. This follows since the ω–limit sets are not empty by Theorem 3.8. □

The next statement is a simple consequence of Theorem 3.6 and Proposition 4.1.

Corollary 4.3. *The set $\mathfrak{R}$ of all recurrent points (positive and negative) is non-empty.*

Proof. Let $\mathfrak{M}$ denote the closure of the union of all minimal subsets of X. Then by Proposition 4.1 and Corollary 4.1 we have

$$Per(f) \subset \mathfrak{M} \subset \mathfrak{R} \subset \Omega.$$

□

Recall that $Fix(f) \subset Per(f)$ and it is non–empty (see Appendix A.3.3).

4.2.3. Circle Diffeomorphisms. Consider $\mathbb{R}$ as a phase space and also as an additive group. It has as such the proper subgroup $\mathbb{Z}$. The *quotient* between both groups is a new space $\mathbb{S}^1 = \mathbb{R}/\mathbb{Z}$. We understand this quotient as the identification of all elements of $\mathbb{R}$ that differ by an element of $\mathbb{Z}$. We end up thus with the interval $[0, 1]$ with the endpoints identified, since all other real numbers can be built by adding an integer to that interval. This space is equivalent to the unit circle $\mathbb{S}^1$ e.g., via the map $t \to e^{2\pi i t}$. The map $\pi\colon \mathbb{R} \to \mathbb{R}/\mathbb{Z} = \mathbb{S}^1$ given by $\pi(x) = x \mod 1$ is called a **factorisation**. We start by presenting a way of relating maps on different spaces.

4.2.3.1. *The Rotation Number.* [**Hal69, GH86**]

Definition 4.4. Let $f\colon \mathbb{S}^1 \to \mathbb{S}^1$ be a continuous map of the circle. We call a continuous function $F\colon \mathbb{R} \to \mathbb{R}$ a **lift** of f whenever the following relation holds: $f(x \mod 1) = F(x) \mod 1$ (this can be restated as $f(\pi(x)) = \pi(F(x))$. Without loss of generality we assume that f is orientation preserving (otherwise we consider f^2). $F(x+1)-F(x) = \deg f$ is called the **degree** of f.

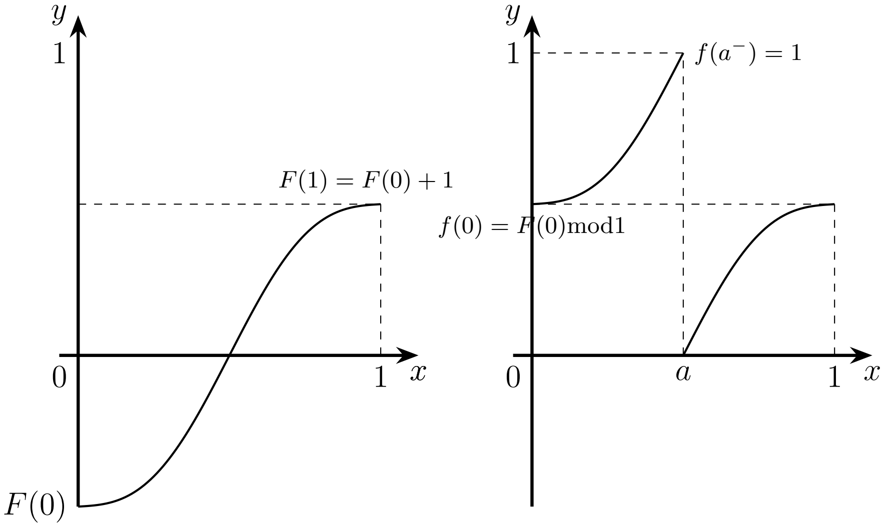


Figure 4.5. The graph of the lift F (left) and the graph of $f = F \mod 1$.

Figure 4.5 depicts a function f and its lift F. Before continuing to investigate circle diffeomorphisms we will prove some facts about lifts.

Lemma 4.1. *Let $f\colon \mathbb{S}^1 \to \mathbb{S}^1$ be a continuous map with lift F. Then* $\deg f \in \mathbb{Z}$ *and it does not depend on x nor on the choice of the lift.*

Proof. We have for $\pi\colon \mathbb{R} \to \mathbb{S}^1$ defined as above:

$$\pi\left(F(x+1)\right) = f\left(\pi(x+1)\right) = f\left(\pi(x)\right) = \pi\left(F(x)\right).$$

Since $\pi(a) = \pi(b) \iff a - b \in \mathbb{Z}$ we have $F(x+1) - F(x) \in \mathbb{Z}$. Because $F(x+1) - F(x)$ is a continuous function it must be constant, i.e., independent of x. Let G be another lift. Then

$$\pi\left(G(x)\right) = f(\pi(x)) = \pi\left(F(x)\right).$$

Since $G(x) - F(x)$ is continuous and integer valued it must be constant. In particular

$$G(x+1) - F(x+1) = G(x) - F(x).$$

□

Hence, there are as many different lifts of f as there are integers. In particular, we can always pick a lift that coincides with f at some point and by continuity at some open set in $(0,1)$.

Lemma 4.2. *The degree depends continuously on the map f. Hence it is locally constant.*

Proof. Let $g\colon \mathbb{S}^1 \to \mathbb{S}^1$ be close to f in the max distance:

$$\max_{x \in \mathbb{S}^1} |g(x) - f(x)| < \frac{1}{4}.$$

We consider two lifts F, G such that they coincide with f and g at some point in $[0, 1)$ and define

$$\phi(x) = G(x) - F(x).$$

Then $\|\phi\|_\infty < \frac{1}{4}$, since if this does not hold, then by continuity there exists x_0 such that $|G(x_0) - F(x_0)| = \frac{1}{4}$. But then $|g(\pi(x_0)) - f(\pi(x_0))| \geq \frac{1}{4}$ since the latter difference differs from the former in an integer.

For $x \in [0, 1)$ we have, then,

$$G(x+1) - \phi(x+1) = F(x+1) = F(x) + \deg f = G(x) + \deg f - \phi(x)$$

hence,

$$\begin{aligned} |G(x+1) - G(x) - \deg f| = |\deg g - \deg f| &= |\phi(x+1) - \phi(x)| \\ &\leq |G(x+1) - F(x+1)| + |G(x) - F(x)| \\ &< \frac{1}{4} + \frac{1}{4} = \frac{1}{2}. \end{aligned}$$

This proves the assertion. □

Lemma 4.3. *If f is an orientation preserving diffeomorphism, then*

$$F(x+1) - F(x) = \deg f = 1.$$

Proof. f is strictly monotonically increasing and therefore also F (they can be chosen to coincide on some open set in $[0, 1)$), so $\deg f > 0$. F maps the interval $[0, 1)$ continuously to an interval of width $\deg f$, since $F(1) - F(0) = \deg f$. Assume $\deg f \geq 2$. Pick two points y_1, y_2 on $F([0, 1))$ such that $y_2 = y_1 + 1$. These points correspond to two different points on $[0, 1)$, namely $y_i = F(x_i)$, $i = 1, 2$. But

$$0 = \pi(F(x_2) - \pi(F(x_1)) = f(\pi(x_2)) - f(\pi(x_1)) = f(x_2) - f(x_1),$$

and hence f is not bijective. □

Theorem 4.3. [**KH96**, 11.1.1 p.387] *Let $f\colon \mathbb{S}^1 \to \mathbb{S}^1$ be an orientation preserving diffeomorphism of the circle. Then there exists (independent of the lift)*

$$\tau(f) := \lim_{|n| \to \infty} \frac{F^n(x) - x}{n} \mod 1$$

for all $x \in \mathbb{S}^1$ and $\tau(f)$ is independent of x. This number is called the **rotation number**. *If f has a periodic orbit, the rotation number is rational.*

Proof. There is no loss of generality in taking n positive, nor in choosing the lift with $F(0) \in [0, 1]$. If f has a periodic orbit and the limit $\tau(f)$ exists independent of x, let us compute it using the periodic point y of period $m \geq 1$. Since the limit is assumed to exist, any infinite sub-sequence will have the same limit. We consider then

$n = km$, $k \to \infty$. Note first that since y is periodic, $F^m(y) = y + p$, for some integer $0 \le p \le m$. Hence,

$$F^{km}(y) = \underbrace{F^m(F^m(\cdots(F^m}_{k \text{ times}}(y))\cdots)) = y + kp$$

Computing the limit, we have

$$\lim_{km\to\infty} \frac{F^{km}(y) - y}{km} = \lim_{k\to\infty} \frac{F^{km}(y) - y}{km} = \lim_{k\to\infty} \frac{(y + kp) - y}{km} = \frac{p}{m}.$$

This proves that the last statement follows from the first.

Let $|x - y| < 1$. Then $|F(x) - F(y)| < 1$ (since f is injective and hence, $\deg f = 1$). Hence, $|F^n(x) - F^n(y)| < 1$ for all $n \in \mathbb{N}$ and

$$\left| \frac{F^n(x) - x}{n} - \frac{F^n(y) - y}{n} \right| \le \frac{1}{n} \left(|F^n(x) - F^n(y)| + |y - x|\right) \le \frac{2}{n}.$$

Hence if the rotation number exists for x, it takes the same value for any y such that $|x - y| < 1$. Moreover, let $y = y_0 + k$, $0 \le y_0 < 1$, $k \in \mathbb{N}$. Then

$$\begin{aligned} F(y) =& F(y_0 + k) \\ =& F(y_0) + (F(y_0 + 1) - F(y_0)) + \cdots \\ & + (F(y_0 + k) - F(y_0 + k - 1) = F(y_0) + k. \end{aligned}$$

Repeating this argument iterating on F, we obtain that $F^n(y) = F^n(y_0) + k$ and further that $F^n(y) - y = F^n(y_0 + k) - (y_0 + k) = F^n(y_0) + k - (y_0 + k) = F^n(y_0) - y_0$. Hence,

$$\left| \frac{F^n(x) - x}{n} - \frac{F^n(y) - y}{n} \right| \le \frac{1}{n} \left(|F^n(x_0) - F^n(y_0)| + |y_0 - x_0|\right) \le \frac{2}{n}.$$

for **all** $x, y \in \mathbb{R}$, proving that the rotation number is independent of x. It remains to prove that the limit exists. Let us attempt to calculate this limit for $x = 0$.

Since F is monotonically increasing and $\deg f = 1$, there exists an integer $0 \le r < n$ such that $r < F^n(0) < r + 1$. This shows that the sequence $\{F^k(0)/k\}$ is bounded. Repeating the argument m times, we have $mr < F^{nm}(0) < m(r + 1)$. Dividing these inequalities by n and nm, we obtain

$$\left| \frac{F^{nm}(0)}{nm} - \frac{F^n(0)}{n} \right| < \frac{2}{n}.$$

Exchanging the roles of m and n we obtain a similar relation. Thus,

$$\left| \frac{F^m(0)}{m} - \frac{F^n(0)}{n} \right| < \frac{2}{n} + \frac{2}{m}.$$

Hence, $\{F^k(0)/k\}$ is a Cauchy sequence and has therefore a limit. □

Theorem 4.4. [**KH96**, 11.1.3 p.388] *Let $h\colon \mathbb{S}^1 \to \mathbb{S}^1$ be an orientation preserving homeomorphism. Then*

$$\tau\left(h^{-1} \circ f \circ h\right) = \tau(f).$$

Proof. Let F, H be lifts of f, h, respectively. Then $H^{-1}FH$ is a lift of $h^{-1}fh$. Since f is an orientation preserving homeomorphism the map $F\colon \mathbb{R} \to \mathbb{R}$ is strictly monotone. We may assume (by choosing appropriate lifts) that $H(0) \in [0, 1)$. As in the previous proof this implies that

$$|H(x) - x| < 2, \quad x \in \mathbb{R}.$$

Setting $x = H^{-1}(y)$ we get also

$$\left|y - H^{-1}(y)\right| < 2, \quad y \in \mathbb{R}.$$

If $|x-y| < 2$ then for $z = (x+y)/2$ we have $|x-z| < 1$ and $|y-z| < 1$. Therefore, if $|x - y| < 2$ we have (from the previous Theorem)

$$\begin{aligned}|F^n(x) - F^n(y)| &= |F^n(x) - F^n(z) + F^n(z) - F^n(y)| \\ &\leq |F^n(x) - F^n(z)| + |F^n(y) - F^n(z)| \leq 2.\end{aligned}$$

Hence,

$$\begin{aligned}&\left|\left(H^{-1} \circ F \circ H(x)\right)^n - F^n(x)\right| = \left|H^{-1} \circ F^n \circ H(x) - F^n(x)\right| \\ &\leq \left|H^{-1}\left(F^n \circ H(x)\right) - F^n \circ H(x)\right| + |F^n \circ H(x) - F^n(x)| \\ &< 2 + 2 = 4.\end{aligned}$$

Therefore

$$\frac{\left|(H^{-1} \circ F \circ H(x))^n - F^n(x)\right|}{n} < \frac{4}{n}.$$

□

Theorem 4.5. [**KH96**, 11.1.4 p.389] *Let $f\colon \mathbb{S}^1 \to \mathbb{S}^1$ be an orientation preserving diffeomorphism of the circle. Then*

$$\tau(f) \in \mathbb{Q} \quad \iff \quad f \textit{ has a periodic orbit.}$$

Proof. One direction was proved in Theorem 4.3. Let then $\tau(f) = \frac{p}{q} \in \mathbb{Q}$. Then

$$\begin{aligned}\tau(f^m) &= \lim_{n\to\infty} \frac{F^{mn}(x) - x}{n} \mod 1 \\ &= m \lim_{n\to\infty} \frac{F^{mn}(x) - x}{nm} \mod 1 = m\tau(f) \mod 1.\end{aligned}$$

Hence,

$$\tau(f^q) = 0.$$

We will show now that assuming that f^q does not have fixed points leads to a contradiction. Let F^q be the lift of f^q with $F^q(0) \in [0, 1)$. If $F^q(x) - x \in \mathbb{Z}$ then x is a fixed point. Hence, if there are no fixed points, $0 < F^q(x) - x < 1$ (the relation holds for $x = 0$ and $F^q(x) - x$ is a continuous function that cannot take integer values). By continuity and periodicity, there exists a $\delta > 0$ such that

$$0 < \delta \leq F^q(x) - x \leq 1 - \delta.$$

Hence,

$$n\delta \leq \sum_{i=0}^{n-1} \left[F^q\left(F^{qi}(x)\right) - F^{qi}(x)\right] = F^{qn}(x) - x \leq n(1-\delta)$$

and the rotation number is $\delta \leq \tau(f^q) \leq 1-\delta$. The contradiction arised assuming f^q did not have fixed points. □

If $\tau \notin \mathbb{Q}$ then the orbits of f are displaced in the same way as the orbit of the rotation $R_{\tau(f)}(x) = x + \tau(f) \mod 1$.

Theorem 4.6. [**KH96**, 11.2.4. p.395] *Let the rotation number of an orientation preserving circle diffeomorphism f be irrational.*
Then for $n_1, n_2, m_1, m_2 \in \mathbb{Z}$ and $x \in \mathbb{R}$ we have

$$n_1\tau + m_1 < n_2\tau + m_2 \qquad \Longleftrightarrow \qquad F^{n_1}(x) + m_1 < F^{n_2}(x) + m_2.$$

Proof. The function

$$p(x) = F^{n_1}(x) + m_1 - F^{n_2}(x) - m_2$$

is continuous and never changes its sign since otherwise $p(x) = 0$ implies that $f^{n_1}(x) = f^{n_2}(x)$ (or $f^{n_1 - n_2}(x) = x$ since f is invertible) and x would be periodic, which is impossible since $\tau(f) \notin \mathbb{Q}$. Let us assume that $p(x) < 0$. The other case is similar.

We write $y = F^{n_2}(x)$. Then

$$F^{n_1-n_2}(y) - y < m_2 - m_1.$$

Let $n \in \mathbb{N}$. Define inductively $y_0 = 0$ and $y_{i+1} = F^{(n_1-n_2)}(y_i)$, $i = 0, \cdots, n-1$. Then

$$F^{n(n_1-n_2)}(0) = y_n = \sum_{i=0}^{n-1}(y_{i+1} - y_i) = \sum_{i=0}^{n-1}(F^{(n_1-n_2)}(y_i) - y_i) < n(m_2 - m_1).$$

Hence,

$$\tau = \lim_{n\to\infty} \frac{F^{n(n_1-n_2)}(0)}{n(n_1-n_2)} < \frac{m_2 - m_1}{n_1 - n_2}.$$

Similarly

$$F^{n_1}(x) + m_1 > F^{n_2}(x) + m_2 \qquad \Longrightarrow \qquad n_1\tau + m_1 > n_2\tau + m_2.$$

□

We are now ready to prove the following theorem[**KH96**]

Theorem 4.7 (Poincaré). [**KH96**, p.397] *Let f be an orientation preserving diffeomorphism of the circle with irrational rotation number. Then*

1. *f is transitive $\Longrightarrow$ f is topologically conjugate to $R_{\tau(f)}$,*

2. *f is not transitive $\Longrightarrow$ there is a surjective monotone continuous map $h\colon \mathbb{S}^1 \to \mathbb{S}^1$ with*

$$h \circ f = R_{\tau(f)} \circ h.$$

This is called a **semi-conjugacy**.

Proof. We start by proving the semi-conjugacy part and subsequently verify that if f is transitive then h is not only surjective but also injective. We fix a lift $F\colon \mathbb{R} \to \mathbb{R}$ of f, $x \in \mathbb{R}$ and write $\tau = \tau(f)$.

We define a map H on the set

$$B := \{F^n(x) + m \,:\, n, m \in \mathbb{Z}\}$$

by

$$H\left(F^n(x) + m\right) = n\tau + m.$$

The set B consists of all possible lifts of the orbit of x by f. On the set B we have already the property we are interested in, namely

$$H \circ F = R_\tau \circ H$$

since

$$\begin{aligned} H \circ F\left(F^n(x) + m\right) &= H\left(F^{n+1}(x) + m\right) = (n+1)\tau + m \\ &= R_\tau(n\tau + m) = R_\tau \circ H\left(F^n(x) + m\right). \end{aligned}$$

The first equality holds since $\deg f = 1$ and hence $F(x+1) - F(x) = 1$, from which it subsequently follows that $F(x+m) = F(x) + m$. First we will extend H to $\overline{B}$ and subsequently to all $\mathbb{R}$.

Lemma 4.4. *H has a continuous extension to $\overline{B}$.*

Proof of the Lemma. Note that by Theorem 4.6 H is monotone on B. For any $z \in \overline{B}$ there are sequences $x_n, y_n \in B$ having z as their limit and such that $x_n \nearrow z$ and $y_n \searrow z$. Since H is monotone both the left–hand–side and right–hand–side limits of H at z exist, independently of the choice of sequence, i.e., $\lim_{n\to\infty} H(x_n) = H(z_-)$ and $\lim_{n\to\infty} H(y_n) = H(z_+)$ exist. If these limits were different the set $\mathbb{R} \setminus H(B)$ contains the interval $(H(z_-), H(z_+))$. In the Appendix we will prove that for τ irrational the set

$$H(B) = \{n\tau + m \,:\, n, m \in \mathbb{Z}\}$$

is dense in $\mathbb{R}$ (Theorem A.18). Therefore $\mathbb{R} \setminus H(B)$ does not contain any interval, $H(z_-) = H(z_+) = H(z)$ and the limit $\lim_{x\to z} H(x) = H(z)$ exists. □

We are going to show now that we can extend H to the entire real line. Note on passing that H is continuous and monotone on $\overline{B}$ and surjective onto $\mathbb{R}$ (since $H(B)$ is dense).

Consider an interval $(a, b) \subset \mathbb{R} \setminus \overline{B}$. By surjectivity we have $H(a) = H(b)$ so we set for all $x \in (a, b)$

$$H(x) = H(a) = H(b).$$

Then $H\colon \mathbb{R} \to \mathbb{R}$ is monotone, surjective and continuous. Moreover

$$H \circ F = R_\tau \circ H$$

by monotonicity and continuity. For $z = F^n(x) + m \in B$ we get

$$H(z+1) = H\left(F^n(x) + m + 1\right) = n\tau + m + 1 = H(z) + 1.$$

This holds by continuity and monotonicity for all z. Hence H can be projected to a continuous function $h = H \mod 1\colon \mathbb{S}^1 \to \mathbb{S}^1$. This completes the proof of the semi-conjugacy part.

If f is transitive, i.e., there is a dense orbit $\left(f^n(x)\right)_n$, the set $\mathbb{R} \setminus B$ does not contain any interval. Moreover, $\overline{B} = \mathbb{R}$ and H is a bijection. □

Intuitively, one may expect that an orientation preserving diffeomorphism with irrational rotation number will behave as an irrational rotation. The above Theorem indicates that it is almost like this, such a map is semi-conjugate to a rotation. To assure full conjugacy, we need only a little more structure. This part of the theory was done 50 years after Poincaré by Denjoy in 1932.

THEOREM 4.8 (Denjoy). [**KH96**, p.401],[**GH86**] *Let $f\colon \mathbb{S}^1 \to \mathbb{S}^1$ be a diffeomorphism with irrational rotation number $\tau(f)$ and with derivative $f'(x)$ of bounded variation. Then f is transitive and hence, h is a conjugacy.*

REMARK 4.3. A function g is said to be of *bounded variation* on an interval $[a,b]$, if its total variation $V_a^b(g) = \sum_{k=0}^{n-1} |g(x_{k+1} - g(x_k)|$, for all $a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$, is finite. If g is e.g., twice differentiable then it is of bounded variation.

PROOF. For the proof we will need the following Lemma. It states that the asymptotic behaviour of **all** points on the circle is the same.

LEMMA 4.5. *For $x, y \in \mathbb{S}^1$ we have*

$$\omega_f(x) = \omega_f(y).$$

PROOF OF LEMMA 4.5. We want to show that $\omega(y) \subset \omega(x)$. By the invariance of the ω–limit set (Theorem 3.8) $y \in \omega(x)$ implies $\omega(y) \subset \omega(x)$. So let us assume that $y \in \mathbb{S}^1 \setminus \omega(x)$. Also by Theorem 3.8 the ω–limit set of x is compact. Hence its complement is open. But open sets on the circle are special: The set $\mathbb{S}^1 \setminus \omega(x)$ is the union of disjoint intervals

$$\mathbb{S}^1 \setminus \omega(x) = \bigcup_{n\in\mathbb{Z}} I_n \qquad I_n = (a_n, b_n) \quad a_n, b_n \in \omega(x).$$

The intervals (a_n, b_n) are the connected components of $\mathbb{S}^1 \setminus \omega(x)$. Therefore their endpoints must belong to $\omega(x)$. Since $\omega(x)$ is invariant its complement must also be invariant and hence, we have

$$f\left(\bigcup_{n\in\mathbb{Z}} I_n\right) = \bigcup_{n\in\mathbb{Z}} I_n.$$

Since f is a diffeomorphism the image of an interval (a_n, b_n) is also an interval J that (by invariance) cannot contain a point of $\omega(x)$ and whose endpoints are in $\omega(x)$. Therefore it must be one of the I_n:

$$f((a_n, b_n)) = (a_{m(n)}, b_{m(n)})$$

and inductively

$$f^k((a_n, b_n)) = (a_{m(n,k)}, b_{m(n,k)}).$$

The forward images of (a_n, b_n) cannot return to themselves, otherwise the map

$$f^k\colon (a_n, b_n) \to (a_n, b_n) = (a_{(m(n,k)}, b_{(m(n,k)})$$

would have a fixed point and then by Theorem 4.5 $\tau \in \mathbb{Q}$. Hence, $n \neq m(n,k)$ for all n and $k \neq 0$. Since the intervals (a_n, b_n) are all disjoint the orbit of y has at most one point in each interval I_n and hence $\omega(y) \cap \mathbb{S}^1 \setminus \omega(x) = \varnothing$. Hence, $\omega(y) \subset \omega(x)$. Exchanging the role of x and y we get $\omega(x) \subset \omega(y)$. □

Coming back to the Theorem, we note that the intervals I_n are called **wandering intervals**. They constitute the complement of the ω–limit set of any point on the circle. In particular $\Omega = \omega(x)$, i.e., the set of non-wandering points coincides with this universal ω–limit set ($\omega(x) \subset \Omega$ and $I_n \subset (\mathbb{S}^1 \setminus \Omega)$). We proceed with the proof by showing that assuming that f is not transitive *and* that f' is of bounded variation leads to a contradiction.

If f is not transitive there exists a nonempty interval $I_0 = (a_0, b_0) \subset \mathbb{S}^1 \setminus \omega(x)$. By Lemma 4.5, the orbit $\left(f^k(I_0)\right)_k$ of this interval consists of disjoint intervals and the map

$$f^k\colon I_0 \to f^k(I_0) = I_k$$

is a diffeomorphism between these two intervals.

Now we set

$$\phi(x) = \log|f'(x)|.$$

Then ϕ has bounded variation because f' has bounded variation and $|f'| \geq c > 0$ on the compact set $\mathbb{S}^1$ ($\log z$ has bounded derivative on each interval $[a, b]$ with $0 < a < b < \infty$). Let

$$V = \operatorname{var}(\phi) = \sup_{x_i\in\mathbb{S}^1, i=1,\cdots n} \sum_{k=1}^{n-1} |\phi(x_{k+1}) - \phi(x_k)|.$$

Let $I \subset \mathbb{S}^1$ be an interval such that all the intervals $I, f(I), \cdots, f^{n-1}(I)$ are pairwise disjoint for some $n \in \mathbb{N}$. If $x, y \in I$ then, letting $f^k(x) = x_{2k+2}$ and $f^k(y) = x_{2k+1}$ we have that

$$\begin{aligned} V &\geq \sum_{m=1}^{2n-1} |\phi(x_{m+1}) - \phi(x_m)| \geq \sum_{k\geq 0}^{n-1} |\phi(x_{2k+2}) - \phi(x_{2k+1})| \\ &= \sum_{k=0}^{n-1} \left|\phi(f^k(x)) - \phi(f^k(y))\right| \geq \left|\sum_{k=0}^{n-1} \phi(f^k(x)) - \sum_{k=0}^{n-1} \phi(f^k(y))\right| \\ &= \left|\log \prod_{k=0}^{n-1} |f'(f^k(x))| - \log \prod_{k=0}^{n-1} |f'(f^k(y))|\right| \\ &= \left|\log \left(\frac{(f^n)'(x)}{(f^n)'(y)}\right)\right|. \end{aligned}$$

Hence,

$$e^{-V} \leq \frac{\left|(f^n)'(x)\right|}{\left|(f^n)'(y)\right|} \leq e^V. \tag{4.1}$$

We will use the following norm on $\mathbb{S}^1 = \mathbb{R}/\mathbb{Z}$:

$$\|x\|_{\mathbb{S}^1} := \min\{x \mod 1, -x \mod 1\}.$$

We define inductively

$$n_1 := 1,$$

$$n_{k+1} := \min\{n > n_k \, : \, \|R^n_{\tau(f)}(0) = n\tau(f)\|_{\mathbb{S}^1} < \|R^{n_k}_{\tau(f)}(0)\|_{\mathbb{S}^1}\}.$$

We note that $n_k \nearrow \infty$ since $\tau(f) \notin \mathbb{Q}$, see Appendix, Theorem A.18. We need now another Lemma:

Lemma 4.6. *For any $x \in \mathbb{S}^1$ we have that the intervals*

$$(x, f^{-n_k}(x)), (f(x), f^{-n_k+1}(x)), \cdots, (f^{n_k-1}(x), f^{-1}(x))$$

are pairwise disjoint.

Proof. Assume that for some $0 < m < l < n_k$ we have

$$(f^m(x), f^{-n_k+m}(x)) \cap (f^l(x), f^{-n_k+l}(x)) \neq \varnothing.$$

Then by semi-conjugacy also

$$(R^m_{\tau(f)}(x), R^{n_k+n}_{\tau(f)}(x)) \cap (R^l_{\tau(f)}(x), R^{-n_k+l}_{\tau(f)}(x)) \neq \varnothing.$$

By translation-invariance of $\|\cdot\|_{\mathbb{S}^1}$ we derive

$$(R^m_{\tau(f)}(0), R^{-n_k+m}_{\tau(f)}(0)) \cap (R^l_{\tau(f)}(0), R^{-n_k+l}_{\tau(f)}(0)) \neq \varnothing.$$

This implies that $R^m_{\tau(f)}(0) \in (R^l_{\tau(f)}(0), R^{-n_k+l}_{\tau(f)}(0))$ or $R^{-n_k+m}_{\tau(f)}(0) \in (R^l_{\tau(f)}(0), R^{-n_k+l}_{\tau(f)}(0))$. Both these inclusions require an earlier closer return than at time n_k, what contradicts the definition of n_k. □

This lemma and the inequalities (4.1) imply for $y = f^{-n_k}(x)$

$$e^{-V} \leq \frac{|(f^{n_k})'(x)|}{|(f^{n_k})'(f^{-n_k}(x))|} \leq e^V.$$

On the other hand, since f^n is a diffeomorphism we have that for $0 < n \leq n_k$

$$\begin{aligned}|I_n| + |I_{-n}| &\geq \int_{I_0} |(f^n)'(x)|\, dx + \int_{I_0} |(f^{-n})'(x)|\, dx \\ &\geq \int_{I_0} \sqrt{|(f^n)'(x)| \cdot |(f^{-n})'(x)|}\, dx \\ &= \int_{I_0} \sqrt{\frac{|(f^n)'(x)|}{|(f^n)'(f^{-n}(x))|}}\, dx \\ &\geq |I_0|\sqrt{e^{-V}} = e^{-V/2}\, |I_0|\,.\end{aligned}$$

We used that $a + b \geq \sqrt{ab}$ for $a, b > 0$ and the derivative of the identity $f^n(f^{-n}(x)) = x$. Hence, for each $k \in \mathbb{N}$

$$\sum_{0 \leq n \leq n_k} |I_n| \geq \sum_{0 \leq n \leq n_k} e^{-V/2}\, |I_0| = (n_k + 1)e^{-V/2}|I_0| \to \infty$$

since $n_k \nearrow \infty$. This is a contradiction since the total length of the disjoint intervals I_n cannot exceed the length of $\mathbb{S}^1$. □

4.2.3.2. *The Denjoy Example.* [**KH96**] We are going to construct an example of an $C^{1+\delta}$–diffeomorphism that is not transitive. This is due to **Denjoy** and later to **Herman**. We fix a sequence

$$l_n = |J_n| = c_\delta(|n| + 1)^{-1/\delta}$$

where $c_\delta = \left(\sum_{n \in \mathbb{Z}}(|n| + 1)^{-1/\delta}\right)^{-1}$.

Given an irrational α, the intervals J_n are arranged in the same order on the circle as the orbit $n\alpha$ and we "blow up" a point of the orbit $n\alpha$ to an interval J_n and "squeeze" the complement of the intervals to keep total length equal to one. This way the complement Ω of the intervals is a compact perfect disconnected subset of the circle and hence a Cantor set (the orbit $n\alpha$ is dense, this implies that for any two points of the circle there is an n such that $n\alpha \in (x, y)$; hence, there is an interval I_n between the two points and Ω is totally disconnected). For details about the Cantor set we refer to Appendix A.2.1 and Chapter 6.

A standard Denjoy map f is defined by mapping the interval J_n diffeomorphically to J_{n+1}, with derivative 1 at the endpoints (see figure 4.6).

We can choose the map $f\colon J_n \to J_{n+1}$ to have Hölder continuous derivative of exponent δ. For $n < 0$ and $(x, y) = J_n$ we have $(f(x), f(y)) = J_{n+1}$ and

$$|f(x) - f(y)| = |J_{n+1}| \leq \sup_{z \in J_n} |f'(z)| \cdot |J_n|\,.$$

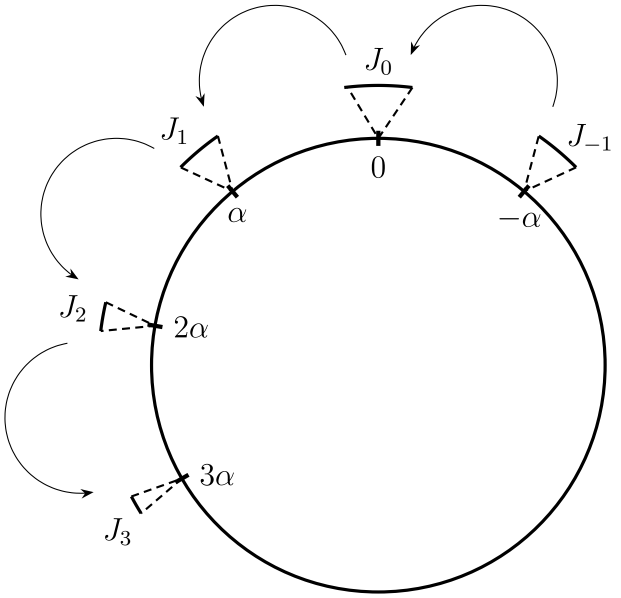


FIGURE 4.6. A Denjoy map.

This gives the following bound on the derivative (we need to make the derivative only slightly larger than the ratio of the length of the two intervals). Hence, we can find a diffeomorphism $f\colon J_n \to J_{n+1}$ whose at any point $z \in J_n$ is at most

$$\begin{aligned} 1 \le |f'(z)| &\le \left(1 + \frac{1}{|n|}\right) \frac{|J_{n+1}|}{|J_n|} = \left(1 + \frac{1}{|n|}\right) \frac{c_\delta(|n+1|+1)^{-1/\delta}}{c_\delta(|n|+1)^{-1/\delta}} \\ &\le \left(1 + \frac{1}{|n|}\right) \left(1 - \frac{1}{n}\right)^{-1/\delta}. \end{aligned}$$

Hence we have for $x, y \in J_n$ that

$$\begin{aligned} |f'(x) - f'(y)| &\le \left| 1 - \left(1 + \frac{1}{|n|}\right) \left(1 - \frac{1}{n}\right)^{-1/\delta} \right| \\ &\le \left| 1 - \left(1 + \frac{1}{n}\right) \left(1 - \frac{2}{\delta \cdot n}\right) \right| \\ &\le D \frac{1}{|n|+1} = D\,|J_n|^\delta \le D \cdot |x-y|^\delta. \end{aligned}$$

Similarly we can estimate the derivative for $n > 0$. This defines a map in the class $C^{1+\delta}$. Moreover the intervals J_n are wandering. Hence, f is not transitive.

4.3. Exercises

Exercise 4.1. Compute the solutions of the system

$$\begin{cases} x_{n+1} &= ax_n + b, \quad n \geq 0 \\ x_0 &= X \end{cases}$$

where a and b are constants.

Exercise 4.2. Compute the solution of the system

$$\begin{cases} x_{n+1} &= \beta(x_n)^\alpha, \quad n \geq 0 \\ x_0 &= X \end{cases}$$

where β and X are positive numbers (Hint: Recast the problem as in the previous exercise by a change of variables performed taking logarithms).

Exercise 4.3. Consider the function

$$f(x) = \frac{ax}{1+bx}$$

where $a > 1$ and the discrete dynamical system

$$\begin{cases} x_{n+1} &= f(x_n), \quad n \geq 0 \\ x_0 &= X \end{cases}$$

Through a change of variables of type:
$x = \phi(z) = z/(1+\gamma z)$ (choose some adequate γ), the system becomes linear in z. Solve the system.

Exercise 4.4. The linear map, $x_{n+1} = rx_n$ was advanced centuries ago by Malthus to model population grow.

1. Find the general solution for $x_0 > 0$.
2. For which values of r the solution presents exponential growth?
3. For which values of r the solution presents exponential decrease?
4. Is exponential growth biologically sensible? Under which limits?

A sometimes serious, sometimes less relevant flaw of this model is that in population problems x is a (non-negative) integer, not just any arbitrary real number.

Exercise 4.5. The logistic map ([**MSS73**]), $x_{n+1} = f(x_n)$, where $f(x) = rx(1-x)$, $r \in (0, 4]$, $x \in [0, 1]$, has received an extraordinary attention in the past.

1. Explain how this map "corrects" the problem of exponential growth in Malthus model.
2. The model assumes a maximal population K, the *carrying capacity* (hidden in this formulation). Show that the non-integer "flaw" of Malthus model reflects here as x taking rational values in $[0, 1]$ with denominator K.

3. Find the location of fixed point(s) of the map as a function of r.
4. For what values of r can we use Banach Fixed Point Theorem to compute the fixed points?
5. Repeat the two last points for the first iterated map $f \circ f$, i.e., $x_{n+1} = f(f(x_n))$.

Exercise 4.6. Consider the system

$$\begin{cases} \dot{x} &= -y + x(1-x^2-y^2) \\ \dot{y} &= x + y(1-x^2-y^2) \end{cases}$$

This system may be easily analysed using polar coordinates. Perform the coordinate change and verify that the set $\Sigma = \{r > 0, \phi = 0\}$ is a good Poincaré Section. Compute Poincaré's "First-return map" by picking an initial condition on Σ and calculating its time-evolution up to $t = 2\pi$. Compute the fixed points of the map.

Exercise 4.7. Show that the system $y'' + (y^2 + (y')^2 - 2) + y = 0$ has a periodic, non-constant solution (Hint: Polar coordinates).

Exercise 4.8. Show that the map $A_\alpha\colon \mathbb{T}^2 \to \mathbb{T}^2$ given by

$$A_\alpha(x,y) = (x+\alpha, x+y) \mod 1$$

is topologically transitive if $\alpha \notin \mathbb{Q}$.
Is this map mixing if $\alpha \notin \mathbb{Q}$ (cf. Example 3.3, Chapter 3) ?
What happens for $\alpha \in \mathbb{Q}$?

Exercise 4.9. Let Φ_t be the billiard flow in the unit square, i.e., the trajectories flow linearly inside the square and are reflected by the elastic law ("incoming angle equals outgoing angle"). Let us fix an initial angle and consider the Poincaré map with respect to one of the sides of the square. Show that this map is equivalent to a rotation.
Hint: You can calculate the map directly in elementary trigonometry. Another way is to reflect the square to get 4 copies which constitute a square of side–length 2. By reflection the trajectory is a straight line through the boundary.

Exercise 4.10. Let X be a compact, perfect (i.e., no isolated points) metric space. Show that when $f\colon X \to X$ is a transitive homeomorphism (recall: *homeomorphism*: a continuous map that is invertible; *transitive*: there is a point $x \in X$ such that the set $\{f_n(x) : n \in \mathbb{Z}\}$ is dense in X), then there is a point $y \in X$ such that already its forward orbit is dense, i.e.,

$$\overline{\{f_n(y) : n \in \mathbb{N}\}} = X.$$

Exercise 4.11. Let $f\colon [0,1) \to [0,1)$ be given by

$$f(x) = 10x \mod 1.$$

How does the set $\alpha(x)$ look like ($\alpha(x)$ is defined here as the set of accumulation points of all pre-images)? What is the set $\alpha(0)$?

Exercise 4.12. Let $f\colon [0,1) \to [0,1)$ defined by

$$f(x) = 2x \mod 1.$$

Is there a point $x \in [0,1)$ such that $\omega(x)$ is countable infinite?
Hint: Use the binary expansion of a real number. Try to identify the repeating patterns in this expansion

Exercise 4.13. Consider the system $x_{n+1} = 4x_n^3 - 3x_n$. Show that periodic points are dense in $[-1,1]$ *Hint:* Let $x = \cos\theta$ and consider the map $\theta \to 3\theta$ on the unit circle.

CHAPTER 5

Stability of Dynamical Systems

5.1. Stability

We are now interested in understanding how a dynamical system changes when it is "slightly" modified. This is a very important question since in experimental situations nothing is known with absolute exactness. We are always "approximating" in one way or the other. Either in the choice of our initial conditions, that can never be exactly matched any closer than the accuracy threshold that we can achieve, or in the choice of the model. For example, parameters that are assumed to be fixed, such as temperature, will not be exactly the same in two runs of the same experiment, some variation, even if it is of one part in a million, will be unavoidable. Therefore, we need to know how these facts influence the behaviour of the system. This complex set of questions is called **stability**.

We may think of two kinds of **stability**.

1. How do nearby initial conditions for a given system behave asymptotically?
2. What happens if one modifies (perturbs) the dynamical system?

The first question is known as *dependence on initial conditions*, namely, whether arbitrarily close initial conditions will separate as $t \to \infty$ or not. The second question is part of the topic of structural stability, considering what is the fate of a given initial condition under the influence of "arbitrarily close" dynamical systems. This last problem may be subsequently divided into a *local* and a *global* part, depending on whether the modifications occur for a small region near a singular point or for large portions of phase space. We will be more specific below.

Focusing on the first question, let p be a singular point, i.e., $v(p) = 0$ and $\dot{x} = v(x)$. Then $x(t) \equiv p$ is an integral curve of the system. If x_0 is close to p, what happens to the trajectory $x(x_0, t)$ for $t \to \infty$? Does it get closer to p, away from p or is it indifferent?

Example 5.1. Let $U\colon \mathbb{R}^n \to \mathbb{R}$ be a potential (energy) (see Figure 5.1). The **gradient field**

$$\dot{x}(t) = -\nabla U(x(t))$$

points into the direction of maximal decrease. Hence, the trajectories flow "downhill".

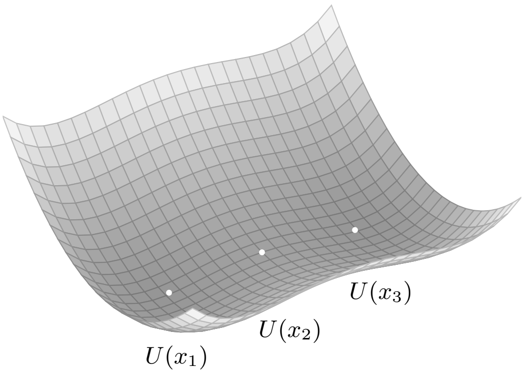


FIGURE 5.1. Dynamics under the influence of a potential.

DEFINITION 5.1 (Lyapunov Stability). [**GH86, SNM96**] A singular point x_0 is called **(Lyapunov) stable** if for any neighbourhood $V(x_0)$ there exists an open set $U(x_0) \subset V(x_0)$ such that for all $y \in U(x)$ and for all $t > 0$, the trajectory $x(y,t) \in V(x_0)$. If in addition $\lim_{t\to\infty} x(y,t) = x_0$, then the singular point is called **asymptotically stable**. Singular points that are not stable are called **unstable**.

Question: When is a singular point asymptotically stable?

In the case of a gradient flow as in Figure 5.1, the points x_1 and x_3 are stable because any nearby trajectory flows to these points. There are at the local minima of $U(x)$ and there is no possibility to flow downhill further on. However, the point x_2 is not asymptotically stable. It is called a *saddle point* and there are many ways to continue to flow downhill for nearby trajectories. The projected vector field looks like in Figure 5.2 (left) for x_1 and x_3, while for x_2 it is depicted on Figure 5.2 (right).

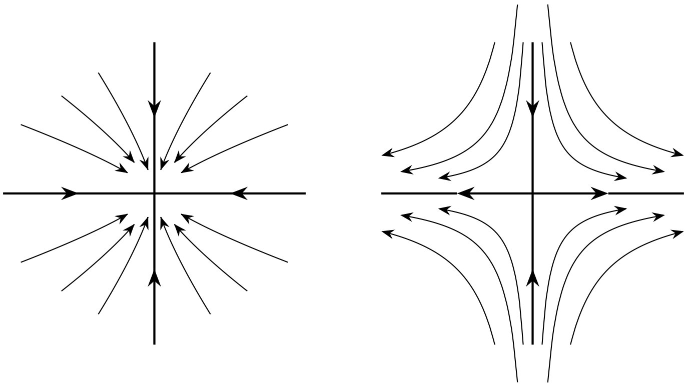


FIGURE 5.2. Local flow. Left: Near a stable fixed point. Right: Near a saddle point.

Let us now generalise this idea. We want to find a function which mimics the role of a potential in such a way that we can decide whether a point is stable or not. Such a function is called a **Lyapunov function**.

5.2. Lyapunov Functions

5.2.1. Inner Lyapunov Stability.

Definition 5.2. Let x_0 be a singular point of the dynamical system $\dot{x} = v$ and $D(x_0)$ a neighbourhood of x_0. A C^1 function $H\colon D(x_0) \to \mathbb{R}$ is called a **Lyapunov function** if $H(x) \geq 0$, $H(x) = 0 \iff x = x_0$ and

$$\nabla H \cdot v \leq 0 \quad \text{on } D.$$

Remark 5.1. H simulates a potential for v. The inequality in the definition means

$$\frac{d}{dt} H(x(x_0, t)) = \sum_{i=1}^{n} \frac{\partial H}{\partial x_i} \frac{dx_i}{dt} = \nabla H \cdot v \leq 0$$

i.e., the function H is non-increasing and in case of strict inequality in $D \setminus \{x_0\}$, it is strictly monotonically decreasing along trajectories.

Theorem 5.1 (First Lyapunov Stability Theorem). [**GH86, SNM96, AAA$^+$97**] *(a) If for a singular point x_0 there exists a Lyapunov function, then this point is (Lyapunov) stable. (b) If in addition $\nabla H \cdot v < 0$ on $D \setminus \{x_0\}$, then x_0 is asymptotically stable.*

Proof. (a) For any neighbourhood $V(x_0)$, consider the compact ball B of radius $r > 0$ centred at x_0, with r sufficiently small such that $B \subset V(x_0) \cap D$. Let $2\epsilon > 0$ be the minimum value of $H(x)$ on the surface ∂B of B (this surface is a compact set and hence $H(x)$ attains a minimum), and $U(x_0) = \{x \in B \,:\, H(x) < \epsilon\}$. $V(x_0)$ and $U(x_0)$ are the neighbourhoods in the definition of stable fixed point. For $y_0 \in U(x_0)$, $x(y_0, t) \in U(x_0)$ for all $t > 0$ since $H(x(y_0, t))$ is non-increasing along trajectories. Therefore, the trajectory exists forever and x_0 is stable.

(b) It remains to show that $\lim_{t\to\infty} x(y_0, t) = x_0$. Since H is in this case monotonically decreasing along trajectories, $\lim_{t\to\infty} H(x(y_0, t))$ exists. We will show that this limit is zero. Assume that this limit is some value $a > 0$ instead. Then $H(x(y_0, t)) > a > 0$ for all $t > 0$. Let $\delta > 0$ be so small that $H(x) < a$ if $|x - x_0| < \delta$. This implies $|x(y_0, t) - x_0| \geq \delta$ for all $t > 0$. Let $W = U(x_0) \cap \{x \,:\, |x - x_0| \geq \delta\}$. Then we have, for some negative b,

$$\nabla H \cdot v < b < 0$$

in the region W. This implies that

$$\frac{d}{dt} H(x(y_0, t)) < b < 0$$

and after a *finite* amount of time $H(x(y_0,t))=0$. This is a contradiction since x_0 is the only point where the Lyapunov function vanishes and this point does not belong to W. Hence,

$$\lim_{t\to\infty} H(x(y_0,t))=0$$

what implies (by continuity) that $\lim_{t\to\infty} x(y_0,t)=x_0$. □

5.2.2. Gradient Systems. [**GH86**] Gradient systems are dynamical systems where the vector field is the gradient of some suitable function U. This gradient is called the *potential* (recall Example 5.1). Let x_0 be a minimum of the potential. Then $\nabla U(x_0)=0$ and $\nabla U>0$ in an *excluded* neighbourhood $D\setminus\{x_0\}$ of x_0. Let us consider U itself as a Lyapunov function. Hence,

$$\nabla U\cdot v=-|\nabla U|^2<0$$

on $V\setminus\{x_0\}$. Hence U is a Lyapunov function and x_0 is asymptotically stable.

Example 5.2. Let $\dot{x}=Ax$ and $A=\operatorname{diag}(\lambda_1,\cdots,\lambda_n)$ is a diagonal (real–valued) matrix. We set

$$H(x)=x\cdot x=\sum x_i^2$$

then $\nabla H=2(x_1,\cdots,x_n)$ and

$$\nabla H\cdot v=2x\cdot Ax=2\sum \lambda_i x_i^2$$

Hence the solution $x\equiv 0$ is asymptotically stable if all $\lambda_i<0$ (in this example we may say “if and only if”).

Theorem 5.2 (Second Lyapunov Stability Theorem). [**AAA$^+$97**, p.24] *If p is a singular point and the Jacobi matrix $A=Dv(p)$ has only negative eigenvalues then p is asymptotically stable.*

Proof. After shifting the origin of coordinates to the singular point p, we write $v(x)=A(x)\cdot x$ with $A\colon\mathbb{R}^n\to GL_n(\mathbb{R})$ and $A(0)=A$. We prove the theorem by producing an adequate Lyapunov function. Since $A(0)$ is negative definite, we choose the same Lyapunov function as in the previous Example. Then

$$\nabla H\cdot v=2x\cdot A(x)x=2x\cdot A(0)x+2x\cdot o(x)x<0$$

in a neighbourhood of the origin because of the continuity of $A(x)$. □

Remark 5.2. The result is valid also when the Jacobi matrix has complex eigenvalues, provided that their real parts are negative.

Remark 5.3. The previous Theorem is in the flavour of structural stability. The non–linear system $v(x)$ is regarded as a perturbation of the linear system $A(0)x$. It will be important to establish under what conditions this is possible.

EXAMPLE 5.3.

$$\dot{x} = ax - y + kx(x^2 + y^2)$$

$$\dot{y} = x - ay + ky(x^2 + y^2)$$

with $a^2 < 1$, $k < 0$. The linearisation at the fixed point $(0, 0)$ is

$$Dv(0) = \begin{pmatrix} a & -1 \\ 1 & -a \end{pmatrix}$$

and has eigenvalues $\lambda_{1,2} = \pm i\sqrt{1 - a^2}$. In this case, the linear system alone is not enough to decide stability since the eigenvalues do not have negative real parts. However, the nonlinear part of this system is such that there exists a Lyapunov function:

$$H(x, y) = x^2 - 2axy + y^2 = (x - ay)^2 + (1 - a^2)y^2$$

Then $H > 0 \iff (x, y) \neq (0, 0)$ and $\nabla H = 2(x - ay, y - ax)$. Hence,

$$\begin{aligned} \frac{1}{2}\nabla H \cdot v =& (x - ay)(ax - y) + (x - ay)kx(x^2 + y^2) \\ & + (y - ax)(x - ay) + (y - ax)ky(x^2 + y^2) \\ =& k(x^2 + y^2)(x^2 - 2axy + y^2) \\ =& k(x^2 + y^2)H(x, y) < 0 \end{aligned}$$

for $(x, y) \neq (0, 0)$, and we have asymptotic stability.

5.2.3. Lyapunov Functions and Trapping Regions. [GH86, SNM96]

Recall the definition of Trapping Region in Definition 3.6.1 and Remark 3.14. Consider a dynamical system having an asymptotically stable fixed point x_0 and a Lyapunov function that is strictly decreasing in some domain $D(x_0)$. Take some other point $y \in D(x_0)$. Then $a = H(y)$ is a positive number.

PROPOSITION 5.1. *The set $x : H(x) = a$ is the boundary of a trapping region around x_0.*

PROOF. Since H is positive definite, $H(x) = a$ denotes the boundary of an open neighbourhood of the point x_0. Since x_0 is asymptotically stable H is strictly decreasing and hence $\nabla H \cdot v < 0$ on the surface of the ellipsoid (and everywhere outside x_0). □

REMARK 5.4. Trapping regions are however broader than what the above proposition suggests, since they can be defined without requiring that the enclosed region should have as only singularity a stable fixed point. To have a trapping region is sufficient to find a closed surface homeomorphic to a sphere where the vector field points inwards, regardless of which kind of dynamics one has on the inside.

5.2.4. Asymptotic Stability for Maps. Definition 5.1 of stability and asymptotic stability can be directly translated to maps, the only difference being the use of "discrete time" instead along with a "substitute" for taking derivatives.

The concept of Lyapunov function and the first Lyapunov Theorem can be restated for maps as follows. Consider a C^1 map $x_{n+1} = F(x_n)$ with a fixed point x_0. By a shift of the origin of coordinates, let us locate this origin at the fixed point, so that $x_0 = 0$, or in other words $F(0) = 0$.

Definition 5.3. Let 0 be a singular point and D a neighbourhood of 0. A C^1 function $H\colon D \to \mathbb{R}$ is called a **Lyapunov function** if $H(x) \geq 0$, $H(x) = 0 \iff x = 0$ and

$$\Delta H(x) = H(F(x)) - H(x) \leq 0 \quad \text{on } D.$$

Remark 5.5. Again, the function H is non-increasing along trajectories and in case of strict inequality in $D \setminus \{x_0\}$, it is strictly monotonically decreasing along trajectories.

Theorem 5.3. *(a) If for the singular point $x = 0$ there exists a Lyapunov function, then this point is stable. (b) If in addition $\Delta H(x) < 0$ on $D \setminus \{0\}$, then 0 is asymptotically stable.*

Proof. The proof is completely analogous to that of Theorem 5.1, only that the quantity to be monitored along trajectories is now $\Delta H(x)$. □

The analogue of Theorem 5.2 is valid as well. By considering time–one maps, it is clear that its valid formulation has to be such that stability for flows automatically translates into stability of time–one maps and vice-versa. Consider the linear case to build up some intuition. A linear map generates a flow $\phi_t(x_0) = e^{At}x_0$. The time–one map is: $x_{n+1} = e^A x_n$. The origin is a fixed point of the linear vector field Ax and of the time–one map. Theorem 5.2 is satisfied if and only if the eigenvalues of e^A have absolute value smaller than unity. Hence, we state:

Theorem 5.4 (Lyapunov Stability Theorem for Maps). *If p is a singular point of the C^1 map $x_{n+1} = F(x_n)$ and the Jacobi (linearisation) matrix $A = DF(p)$ has only eigenvalues of absolute value smaller than unity, then p is asymptotically stable.*

5.3. Exercises

Exercise 5.1.

Use the Lyapunov function method to show that the origin is a stable fixed point for the system

$$\begin{aligned} \dot{x} &= -2xy - x^3 \\ \dot{y} &= x^2 - y^3 \end{aligned}$$

Exercise 5.2. Compute a Lyapunov function for the systems:
(a) $\ddot{x} + \omega^2(x + x^3) = 0$.
(b) $\ddot{x} + \alpha\dot{x} + \omega^2(x + x^3) = 0$.
Show a trapping region for any of these systems.

Exercise 5.3. Consider the following dynamical system.

$$\begin{aligned}\dot{x} &= ax - 2y + x^2\\ \dot{y} &= x + y + xy.\end{aligned}$$

For which value(s) of the parameter a is the solution $(x(t), y(t)) \equiv (0, 0)$ asymptotically stable ?

5.4. Structural Stability

[**GH86, KH96, AAA$^+$97**] In this Section we want to investigate the second issue of stability, namely what happens when we modify the dynamical system (the vector field). In particular, we might ask whether it makes sense to linearise a system (this may be seen as a drastic modification, i.e., replacing a given $v(x)$ by its first–order Taylor expansion, $v(x_0) + a(x - x_0)$). A linearisation can be rephrased as the "best fitting" linear map. Since linear systems are completely understood, it would be very interesting to profit of them to understand non–linear systems. We would like to carry over the linear information to the non–linear system. In this section we investigate when this is possible. Let us start with an example.

Example 5.4. Recall the presentation in Chapter 2 of linear systems on $\mathbb{R}$. Let two linear systems $\dot{x} = ax$ be given, with constants $a = \lambda$ and $a = \gamma$. We have seen before that the solutions of each system, $x_i(t) = x_i(0)e^{a_i t}$, are similar iff both a_i have the same sign. We will call such systems *similar* or **conjugate**, inspired in the considerations of Chapter 4 where we said that the orientation preserving diffeomorphisms f (with irrational rotation number τ) and R_τ (the rotation by an angle τ) were conjugate if there exists a bijective continuous function h with continuous inverse, such that $f \circ h = h \circ R_\tau$. A conjugacy may be defined between more general objects: maps, vector fields, etc., may be conjugate. Let us see how conjugacy works for these linear systems.

We will show that the systems $\dot{x} = \lambda x$ and $\dot{y} = \gamma y$ are conjugate by the change of coordinates

$$h(x) = \begin{cases} \operatorname{sg}(x)|x|^{\gamma/\lambda}, & x \neq 0 \\ 0, & x = 0 \end{cases}.$$

For $x \neq 0$, consider the dynamics of $y = h(x)$. Outside the origin h is differentiable and hence

$$\dot{y} = \frac{dh}{dx}\dot{x} = \frac{\gamma}{\lambda}|x|^{\frac{\gamma}{\lambda}-1}\lambda x = \gamma h(x) = \gamma y.$$

The dynamics coincides also at the origin, which is a fixed point both for y and x. The coordinate transformation h is not differentiable at the origin for $0 < \gamma/\lambda, \lambda \neq \gamma$ but it is continuous, injective and with continuous inverse. How do trajectories map by h?

$$h(x_0 e^{\lambda t}) = \mathrm{sg}(x_0)|x_0|^{\frac{\gamma}{\lambda}} e^{\gamma t} = y_0 e^{\gamma t}.$$

Hence h maps trajectories on trajectories.

This example motivates the following definition.

Definition 5.4. Two vector fields v, w are **topologically equivalent** (we write $v \sim w(h)$) if there is a continuous invertible map (a homeomorphism) $h \colon \mathbb{R}^n \to \mathbb{R}^n$ (called a **conjugacy**) which maps trajectories into trajectories:

$$h(x_v(x_0, t)) = x_w(h(x_0), t).$$

Theorem 5.5. [**GH86**, p.38] *Let h be a conjugacy between v and w then*

1. *If x_0 is a singular point so is $h(x_0)$,*
2. *If $x_v(x_0, t)$ is periodic so is $x_w(h(x_0), t)$.*

Proof. Since $x_v(x_0, t) \equiv x_0$ for all $t \in \mathbb{R}$ we have $x_w(h(x_0), t) \equiv h(x_0)$ for all $t \in \mathbb{R}$. This proves the first statement. For the second we only have to notice that if τ is a period of $x_v(x_0, t)$ then

$$x_w(h(x_0), t) = h(x_v(x_0, t)) = h(x_v(x_0, t + \tau)) = x_w(h(x_0), t + \tau)$$

and therefore the trajectory $x_w(h(x_0), t)$ is periodic with the same period. □

Remark 5.6. Different kinds of equivalences between systems may be defined, which are suitable for different circumstances. For example we could speak of linear equivalence (h is a linear map), etc.

Example 5.5. In this Example we will show topological equivalence between systems whose trajectories appear to be quite different at a first glance. Let $v(x, y) = (x, y)$ and $w(x, y) = (x + y, -x + y)$. These vector fields have phase portraits as in Figure 5.3.

The corresponding trajectories are

$$x_v((x_0, y_0), t) = e^t (x_0, y_0)$$

and

$$x_w((x_0, y_0), t) = e^t (x_0 \cos t + y_0 \sin t, -x_0 \sin t + y_0 \cos t).$$

We note that each regular trajectory has a unique point of intersection with the unit circle $\mathbb{S}^1$. Hence, if $(x_0, y_0) \neq (0, 0)$ there is a unique time $t(x_0, y_0)$ such that $x_v((x_0, y_0), t(x_0, y_0)) \in \mathbb{S}^1$. Now we define a map $h \colon \mathbb{R}^2 \to \mathbb{R}^2$ by $h(0, 0) = (0, 0)$ and

$$h(x_0, y_0) = x_w\left(x_v((x_0, y_0), t(x_0, y_0)), -t(x_0, y_0)\right).$$

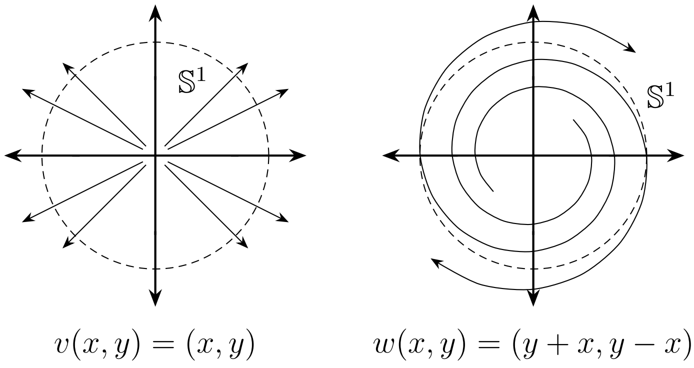


FIGURE 5.3. Two exponentially growing vector fields.

This means we flow to the unit circle along the trajectories of v and then flow "backwards" on the corresponding trajectory of w (see Figure 5.4). This "flight time" (and the whole map h) can be computed explicitly for each initial condition: $2t = -\log{(x_0^2 + y_0^2)}$.

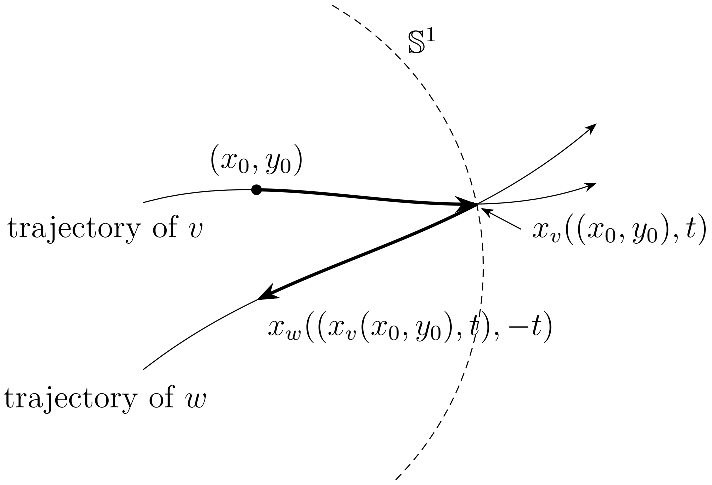


FIGURE 5.4. A point (x_0, y_0) and its image by h.

This map is differentiable everywhere except at $(0,0)$ by the differentiable dependence on initial conditions. It is continuous at $(0,0)$ ($t(x_0,y_0) \to +\infty$ as $(x_0,y_0) \to (0,0)$ and $x_w((x,y),-t) \to (0,0)$ as $t \to +\infty$) and invertible with a continuous inverse. hence it is a coordinate transformation. We want to show now that it maps trajectories on trajectories. For any given time t let $t_0 = t(x_0,y_0)$ and $t_1 = t(x_v((x_0,y_0),t))$. Then $t = t_0 - t_1$ and with $p = (x_0, y_0)$

$$\begin{aligned} h(x_v(p,t)) &= x_w(x_v(x_v(p,t),t_1),-t_1) = x_w(x_v(p,t+t_1),-t_1) \\ &= x_w(x_w(x_v(p,t+t_1),-t-t_1),t) = x_w(x_w(x_v(p,t_0),-t_0),t) \\ &= x_w(h(p),t). \end{aligned}$$

The first equality is the definition of h, the second equality holds since moving a time t and subsequently a time t_1 is the same as moving in one sweep for the total time $t + t_1 = t_0$. The third equality follows since moving a time t_1 and subsequently a time $-t_1$ corresponds to not moving at all. The fourth equality uses that $t + t_1 = t_0$ and the last equation follows again from the definition of h. This shows that the two vector fields are conjugate.

Remark 5.7. We need to give a definition of neighbourhood for the space of vector fields or simply of *nearby vector field* as well. To avoid complications, we consider vector fields defined on $\mathbb{R}^n$ and simplify some details that may be chosen differently. The C^r open ball of radius $\epsilon > 0$ around a vector field v consists of all vector fields w, such that for any compact set K of phase space, $\sup_{x \in K} ||D^k v(x) - D^k w(x)|| < \epsilon$, for $k = 0, \cdots, r$ (the vector fields as well as their first r derivatives lie close in the sup–norm). For structural stability we may accept at this moment $r = 0$, but soon we will use $r = 1$ (see below).

Definition 5.5. The vector field v is called **structurally stable** if any vector field w within some (sufficiently small) open ball around v is topologically equivalent to v.

Remark 5.8. Topological equivalence is an equivalence relation, i.e., we have

- $v \sim v(Id)$ (reflexive)
- $v \sim w(h)$ implies $w \sim v(h^{-1})$ (symmetry).
- $v_1 \sim v_2(h_1)$ and $v_2 \sim v_3(h_2)$ implies $v_1 \sim v_3(h_2 \circ h_1)$ (transitivity).

Therefore, all vector fields in a neighbourhood of a structurally stable field are structurally stable.

Example 5.6. The linear flow on the torus is not structurally stable because periodic (rational a) and non–periodic (irrational a) flows can be arbitrarily close to each other.

Definition 5.6. v, w are **locally topologically equivalent** at x, y if the conjugacy h is defined only in a neighbourhood of x and $h(x) = y$.

Remark 5.9. The concept of nearby vector field can be understood locally as well. Two vector fields are locally C^r–close at x_0 if there exists a compact neighbourhood of x_0 where they are close in the sense of Remark 5.7.

Remark 5.10. If x, y are regular points for v, w respectively then the vector fields are locally topologically equivalent at x, y. This follows from the Straightening–out Theorem since both vector fields are conjugate to the "horizontal" vector field.

This Remark indicates that the only interesting points for local topological equivalence are the singular ones. Away from singular points, the local dynamics in small regions of phase space is essentially the horizontal vector field dynamics. The next Theorem deals with the case of singular points. Note that here we have to proceed beyond continuity and consider C^1 vector fields.

5.5. Hyperbolicity – The Hartman-Grobman Theorem

[**GH86**] In this Section we will put together some ideas from inner stability and from structural stability.

Definition 5.7. A **hyperbolic singular point** (or hyperbolic fixed point) for a C^1 vector field is a singular point x at which the linearisation $Dv(x)$ has no purely imaginary eigenvalues.

Definition 5.8. A **hyperbolic singular point** (or hyperbolic fixed point) for a C^1 map F is a singular point x at which the linearisation $DF(x)$ has no eigenvalues of modulus one (on the complex unit circle).

It is interesting to notice that we have two different (but related) notions of hyperbolicity depending on whether we deal with flows or maps. The same occurred with the notions of Lyapunov stability and asymptotic stability. This is a recurrent feature in the theory of dynamical systems. Many "flow"–theorems have a map companion.

For hyperbolic fixed points of flows, the question of asymptotic stability has a simple translation. Using Theorem 5.2, we give the name *stable* (dropping the adverb *asymptotically*) to the hyperbolic fixed points where all eigenvalues of the linearisation have negative real parts. Other names frequent in the literature are:

node For fixed points whose linearisation eigenvalues have either all positive real parts (*unstable node*) or all negative real parts (*stable node*).

focus For nodes with complex eigenvalues.

saddle For hyperbolic fixed points with both stable and unstable linearisation eigenvalues.

center For (non hyperbolic) fixed points where all eigenvalues lie on the imaginary axis.

These ideas can be translated word by word to maps (now using Theorem 5.4):

stable Hyperbolic fixed points with linearisation eigenvalues of absolute value smaller than one.

node For fixed points whose linearisation eigenvalues have absolute value larger than one (*unstable node*) or smaller than one (*stable node*).

focus For nodes with complex eigenvalues.

saddle For hyperbolic fixed points with both stable and unstable linearisation eigenvalues.

center For (non hyperbolic) fixed points where all eigenvalues have absolute value unity.

THEOREM 5.6 (Hartman–Grobman). [**Gro59, Har60**]
Let x_0 be a hyperbolic singular point of a vector field $\dot{x} = v(x)$ (respectively map $x_{n+1} = v(x_n)$), i.e., the linearisation $A = Dv(x_0)$ satisfies Definition 5.7 (respectively 5.8) above. Then $v(x)$ and the linear field (map) $A \cdot (x - x_0)$ are locally topologically equivalent in a neighbourhood of the singular point.

PROOF. Without loss of generality we may assume that $x_0 = 0$, otherwise we apply a translation by the vector x_0. We will only sketch the basic ideas of the proof. Let us proceed in two steps. First, we notice that out of two flows we can produce corresponding time–one maps. Then we prove the Theorem for maps. Finally we realise that it holds also for flows, since the map–proof shows that the time–one maps of a hyperbolic flow and its linearisation are conjugate.

Let $x_v(\cdot, 1)\colon \mathbb{R}^n \to \mathbb{R}^n$ be the time–one map associated to the vector field v. By the considerations about the dependence on initial conditions discussed in Chapter 2, we know that the derivative of the time–one map fulfills the variational equation 2.4 (recall the comment in the introduction to time–one maps in Section 4.1.1). This is a linear equation and has the solution $z(t) = e^{Dv(0)t} z(0)$. Hence, we can write

$$x_v(x, 1) = Ax + \phi(x)$$

where ϕ is small in norm (and also zero at the origin, together with its derivative) and $A = e^{Dv(0)}$. By assumption, A has no eigenvalues on the unit circle (since $Dv(0)$ has no eigenvalues with zero real part). Let ϕ, ψ be two bounded, small (with sup norm less than $\epsilon > 0$), Lipschitz continuous maps of $\mathbb{R}^n$ (vanishing at the origin together with their first derivative) and A be hyperbolic (no eigenvalues on the unit circle). Then we will show that the two maps $A + \phi$ and $A + \psi$ are "nicely" conjugate, i.e., there is a unique map $H = \operatorname{Id} + u$, u bounded, small and Lipschitz continuous, such that

$$(\operatorname{Id} + u) \circ (A + \phi) = (A + \psi) \circ (\operatorname{Id} + u)$$

or

$$Au(x) - u(Ax + \phi(x)) = \phi(x) - \psi(x + u(x)).$$

Note that the following part of the proof holds irrespective of A arising from a time–one map. First we remark that since A has no eigenvalues of modulus 1, we can decompose

$$\mathbb{R}^n = E^- \oplus E^+ \quad A(E^-) = E^- \quad A(E^+) = E^+$$

where the invariant subspaces E^- and E^+ correspond to the eigenvalues of modulus larger than 1 (E^+) or smaller than 1 (E^-), respectively.

Then

$$\|A^-\| \leq a < 1 \quad \| \left(A^+\right)^{-1} \| \leq a < 1 \tag{5.1}$$

where $A^{-,+} = A|_{E^{-,+}}$. We can now decompose all equations into their $-$part and $+$part (projections onto E^- and E^+).

Let us prove first that the operator

$$\mathfrak{L}(u) = Au - u(A+\phi)\colon C_b^0(\mathbb{R}^n) \to C_b^0(\mathbb{R}^n)$$

is a linear invertible operator with norm of the inverse at most $\frac{1}{1-a}$. To do this, we will show that for any function g there exists a unique v such that $\mathfrak{L}v = g$, and then we will estimate the norm of $\mathfrak{L}^{-1}$. We start by decomposing the equation via E^- and E^+:

$$\begin{aligned} A^+v^+ - v^+ \circ (A+\phi) &= g^+ \\ A^-v^- - v^- \circ (A+\phi) &= g^- \end{aligned}$$

Defining the auxiliary operators $G_-(v^-) = A^-v^- \circ (A+\phi)^{-1}$ and $G_+(v^+) = (A^+)^{-1}v^+ \circ (A+\phi)$ we can rewrite the equations above after some manipulation as:

$$\begin{aligned} (1-G_-)v^- &= -g^- \circ (A+\phi)^{-1} \\ (1-G_+)v^+ &= (A^+)^{-1}g^+ \end{aligned}$$

The good thing is that $||G_\pm|| \leq a$, which guarantees (Theorem A.14) that the operators $(1-G_\pm)^{-1}$ exist, and hence there is a unique solution $v = (v^+, v^-)$ to the above equation. Moreover, $||\mathfrak{L}^{-1}g||/||g|| = ||v||/||g|| \leq \frac{1}{1-a}$.

Knowing what sort of operator $\mathfrak{L}$ is, we can address the original issue, namely to show that the operator equation $\mathfrak{L}(u) = \phi - \psi \circ (I+u)$ has a unique solution u. We consider instead the equivalent equation

$$u = \mathfrak{L}^{-1}[\phi - \psi \circ (I+u)] \equiv F(u).$$

We will show that F is a contraction and hence the above equation has a unique solution. Indeed

$$||F(u) - F(v)|| = ||\mathfrak{L}^{-1}[\psi \circ (I+u) - \psi \circ (I+v)]|| \leq \frac{\epsilon}{1-a}||u-v||$$

since ψ (and ϕ) have Lipschitz constant smaller than ϵ. Restricting the domain of ψ (and ϕ) we can make ϵ as small as necessary in order to obtain $\frac{\epsilon}{1-a} < 1$.

Hence, there is a unique solution for $\mathfrak{L}(u) = \phi - \psi \circ (I+u)$. The map $\operatorname{Id} + u$ is invertible because we can also solve

$$(\operatorname{Id} + v)(A+\psi) = (A+\phi)(\operatorname{Id} + v)$$

in a similar way, which gives the inverse to $\operatorname{Id} + u$.

Now we consider

$$A + \phi = x_v(\cdot, 1) \qquad A + \psi = x_w(\cdot, 1).$$

Then $\mathrm{Id}+u = H$ conjugates these two time–one maps. We derive the final map by "averaging" it from time 0 till time 1:

$$h = \int_0^1 x_w(-t,\cdot) \circ (\mathrm{Id}+u) \circ x_v(t,\cdot)\, dt.$$

To finish the proof we check the following lines saying that we really have a conjugacy of the flow with the linear flow $w = L = Dv(0)$.

$$\begin{aligned}
&x_L(h(x_v(x,s)),-s) = x_L\left(\int_0^1 x_L(H(x_v(x_v(x,s),t)),-t)\,dt,-s\right)\\
&= x_L\left(\int_0^1 x_L(H(x_v(x,s+t)),-t)\,dt,-s\right)\\
&= \int_0^1 x_L\left(H(x_v(x,t+s)),-(t+s)\right)\,dt\\
&= \int_{s-1}^s x_L\left(H(x_v(x,u+1)),-u-1\right)\,du\\
&= \int_0^s x_L(\cdot,-u) \circ [x_L(\cdot,-1) \circ H \circ x_V(\cdot,1)] \circ x_v(x,u)\,du\\
&\qquad + \int_{s-1}^0 x_L(\cdot,-u-1) \circ H \circ x_v(x,u+1)\,du\\
&= \int_0^s x_L(\cdot,-u) \circ H \circ x_v(x,u)\,du + \int_{s-1}^0 x_L(\cdot,-u-1) \circ H \circ x_v(x,u+1)\,du\\
&= \int_0^s x_L(\cdot,-z) \circ H \circ x_v(x,z)\,dz + \int_s^1 x_L(\cdot,-z) \circ H \circ x_v(x,z)\,dz = h
\end{aligned}$$

□

Next we want to investigate the local structural stability of vector fields. Since we just proved that we can reduce to the linear case if we have a hyperbolic singularity we will start with linear fields.

Theorem 5.7. [**Arn73**, p.143] *Two linear hyperbolic vector fields are conjugate if and only if the number of eigenvalues with negative real parts (and hence also those with positive real parts) coincide.*

Proof. We sketch the proof. The one–dimensional case was proven in Example 5.4, e.g.,

$$\dot{x} = \lambda x \qquad \dot{y} = \gamma y \quad \lambda, \gamma < 0$$

are conjugate via $h(x) = \mathrm{s}g(x)x^{\gamma/\lambda}$. By the invariance of subspaces associated to different eigenvalues (a basic result of Linear Algebra), the result holds for a general matrix with all eigenvalues distinct. It remains to show that a matrix with an eigenvalue $\lambda > 0$ of multiplicity $k > 1$ generates a flow that is conjugate to that of the identity matrix of dimension k. Such proof was given by Ladis in 1973, using similar ideas to those inspiring Example 5.5, namely considering a

$(k-1)$–dimensional sphere centred in the origin. All trajectories outside the origin in both vector fields intersect the corresponding sphere only once. Hence, a conjugacy between both systems exist and it can be constructed computing the intersection time, as it was done in Example 5.5. □

This Theorem is the hyperbolic version of a slightly more general Theorem by Ladis:

THEOREM 5.8 (Ladis). [**Lad73**] *Two linear vector fields on $\mathbb{R}^n$ are conjugate if and only if the number of eigenvalues with negative real parts (resp positive real parts) coincide and the blocks with purely imaginary eigenvalues coincide (up to a positive proportionality constant).*

Since the imaginary–eigenvalues block defines an invariant subspace of phase space, if those blocks coincide and the orthogonal complement satisfies the previous Theorem, then Ladis Theorem follows immediately.

REMARK 5.11. We recall that in general, the conjugacy h is not differentiable at $x = 0$.

COROLLARY 5.1. *If p is a hyperbolic singular point. Then v is locally structurally stable at p.*

PROOF. If two vector fields are close and one has a hyperbolic singular point so has the other (and it is close to the first one). Hence, both vector fields are conjugate to their respective linearisations by Hartman-Grobman Theorem. Moreover, both linearisations have the same index (the index does not change under small perturbations). Hence, they are conjugate by Ladis Theorem. □

Note that here we have used the notion of structural stability in a local way, namely we use "locally conjugate" instead of just "conjugate".

REMARK 5.12. The Theorems of Hartman–Grobman and Ladis help to give a complete local understanding of the dynamical behaviour near hyperbolic fixed points, even for a parametric family of vector fields or maps. Indeed, for hyperbolic fixed points, the linearisation matrix suffices to establish if the point is stable, unstable or saddle. Hartman–Grobman theorem suffices to establish the local behaviour of phase–space trajectories, since trajectories for linear systems can be completely computed. Finally, "nearby" dynamical systems obtained by continuously varying some system parameter(s), will, by Ladis Theorem, have conjugate dynamics, as long as all linearisation eigenvalues remain hyperbolic (assuming of course that the dependency with parameters is such that nearby systems have a corresponding fixed point). Hence, knowing the qualitative dynamics for one member of the family via Hartman–Grobman theorem is enough to understand the qualitative dynamics for all nearby hyperbolic systems.

Moreover, the Straightening-out Theorem gives a qualitative understanding of the dynamics far away from singular points. Hence, a significant class of dynamical systems may be qualitatively understood just by adequately using these three theorems.

5.6. Exercises

Exercise 5.4. Let $\dot{x} = A \cdot x$ and $\dot{y} = B \cdot y$ be C^1–conjugate (i.e., the conjugating map is differentiable and the inverse also). Show that then the matrices A, B have proportional eigenvalues.

Exercise 5.5. Let $f \colon U(p) \to \mathbb{R}$ be a C^1–function, $\dot{x} = v(x)$ and $v(p) = 0$. Assume that $f(p) = 0$, $\frac{d}{dt} f(x(x_0, t)) > 0$ for $x_0 \in U(p) \setminus \{p\}$. Moreover there is a sequence $x_n \to p$ such that $f(x_n) > 0$. Then $x(t) \equiv p$ is unstable.

Exercise 5.6. Let $v = -\nabla f$, $f \in C^2(\mathbb{R}^n, \mathbb{R})$. Then $p \in \mathbb{R}^n$ is a hyperbolic singular point if and only if $df(p) = 0$ and $d^2 f(p)$ is a non–degenerate bilinear form.

Exercise 5.7. Use the Poisson integration formula for the circle to show that if an electrostatic potential on a circle is constant then the potential is constant everywhere inside the circle.

Exercise 5.8. (a) Compute the fixed points for the system $x_{n+1} = f(x_n)$, where $f(x) = \mu x(1 - x)$. Study their stability as a function of $\mu \in [0, 4]$ (consider yourself satisfied with a few typical (μ-values).
(b) Repeat for the period-2 points, i.e., the additional fixed points of $f \circ f$.
(c) Try to show that there exists a period three orbit, by studying the fixed points of $f \circ f \circ f$ near $\mu = 1 + 2\sqrt{2}$. For which values of μ the fixed point is stable?

Exercise 5.9. (a) For the system of the previous Exercise, compute for which values of μ the fixed point outside the origin has negative Jacobian.
(b) What is the value of the Jacobian when the system $x_{n+1} = f(f(x_n))$ develops two new fixed points? Compute the stability of these points.

Exercise 5.10. Analyse the stability for the fixed point of the systems $\dot{x} = x^2$ och $\dot{x} = x^3$. Note that linear stability analysis is not enough.

Exercise 5.11. Compute fixed points and stability for
(a) $\ddot{x} + \sin x = 0$.
(b) $\ddot{x} + \epsilon \dot{x} - x + x^3 = 0$.
(c) $\dot{x} = x^2 - x$, $\dot{y} = x + y$.

CHAPTER 6

Center Manifold Theory and Local Bifurcations

6.1. Introduction

We have investigated several kinds of asymptotic behaviour. Many of the systems considered so far were structurally stable, i.e., small perturbations on the vector field or map did not change the system drastically.

However, this is not always the case. It is known that many systems present qualitatively different behaviour when certain changes are operated on them. Consider for example the Lorenz equations. For $0 < r < 1$ there exists only one singular point, while for $r > 1$ there are three such points: the "old" stable fixed point becomes unstable, while a couple of stable, symmetry related fixed points shows up. A drastic transition has occurred for $r = 1$. One of our goals is to be able to address such changes.

The simplest way to study transitions in the qualitative dynamics is to consider parametric families of systems which go from one regime to another while changing the parameter. At some special parameter values the qualitative (asymptotic) dynamics changes drastically. We call such value a **bifurcation point** and the general study of such changes is called **bifurcation theory**.

The word *manifold* appearing on the title of this Chapter refers to a space where each point has a neighbourhood equivalent to Euclidean space. Certain manifolds of phase space are dynamically interesting for being invariants of the flow and will be discussed in detail in Section 6.3.

There is an interest from applications in studying bifurcation problems, since the parameters of a differential equation or map usually represent a given choice of experimental conditions, e.g., constant temperature T_0 or constant amplitude β of an external electromagnetic field in coupled laser devices, the reproduction rate of a population, etc. To some extent, these constant conditions may be altered from experiment to experiment until eventually the asymptotic behaviour of the system is deeply modified.

Apart from the already mentioned Lorenz equations, we will discuss two other examples in this Section, focusing on the effective changes. Consider first the logistic map

$$x \longrightarrow ax(1-x)$$

on the unit interval. This map has been widely studied along the years, starting perhaps with Volterra's population models of the 1920's, up to our days. It is not completely understood but many results are known in detail.

Consider the family of logistic maps for $a \in [0, 4]$. We have seen that for small values of the parameter a (namely for $a < 1$) the singular point $x = 0$ is attracting (Exercise 5.8). By checking the locus of the fixed point and the derivative of the map, we can establish its stability for most parameter values using Theorem 5.6, Hartman–Grobman Theorem. For $a = 1$, the origin is no longer linearly stable. For $a > 1$ and $a - 1$ sufficiently small, a second fixed point appears and the stability of $x = 0$ changes (we could say that stability is "transferred" to the second fixed point when it appears, like in a relay-race while "moving" a away from zero). Taking larger values of a we reach a situation where a periodic orbit of period–2 appears (an invariant set of the map consisting of exactly two points, that map onto each other) branching out of the second fixed point, which still exists (recall Exercise 5.9). Again, the fixed point becomes unstable, while the period–2 inherits its stability. Taking even larger values for a, a period–4 branches out from the period–2 and we subsequently run through a sequence of period–doublings, where the stability is transferred consequently from one new invariant set to the next that branches out: first the period–2, then period–4, period–8, $\cdots$ become attracting periodic orbits. Analytic calculations become very involved pretty soon. The values of a where these transitions take place accumulate monotonically towards a point near $a = 3.57$. At that value, we arrive at a regime where the attractor has infinitely many points. Increasing this parameter further one sees again periodic attracting orbits accumulating on chaotic sets. At $a = 1 + 2\sqrt{2}$ the attractor is a period 3 orbit (Exercise 5.8 again), which subsequently undergoes a "period–doubling" cascade through periods $3 \cdot 2^n$. When $a = 4$ the system develops *complete chaos* meaning that it contains (unstable) periodic orbits of all periods and it is sensitive with respect to initial conditions (in this case, the system is conjugate to the doubling map). In this way, the logistic map exhibits all kind of bifurcations, it develops *"from order to chaos"*.

The second example will be treated in detail in Chapter 8. Consider a hyperbolic saddle fixed point p of a map whose stable and unstable manifolds coincide (see Figure 6.1, left). For the time being and until Section 6.3, consider the stable manifold of a fixed point p as the invariant set of initial conditions such that $\phi_t(x_0) \to p$ as $t \to \infty$, and the unstable manifold as the corresponding set for $t \to -\infty$. Intuitively, we may think that the stable and unstable manifolds of a fixed point each have a "life of its own", i.e., that small changes in the equations (e.g., small changes in some parameters present in the equations) alter these manifolds differently or independently. Therefore, the fact that

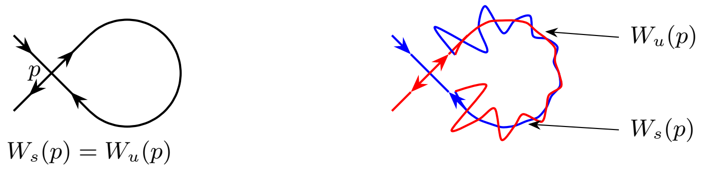


Figure 6.1. A separatrix splitting.

they coincide is highly unlikely for a random choice (of parameters) and it puts specific constraints on the problem. However, it does occur in interesting problems or in some approximation to them.

If we perturb the system slightly, it can be possible to arrange a **separatrix splitting** that creates a transversal homoclinic point (see Figure 6.1, right). The two curves no longer coincide, and eventually they may cross transversally. As we will further discuss in Chapter 8, this situation forces the existence of highly complicated dynamics, since one crossing of the manifolds forces the existence of infinitely many other crossings.

The Lorenz transition and the period–doubling cascade of the logistic map are examples of **local bifurcations**, namely those where the dynamical changes are confined to a local region of phase space close to the associated fixed point. Separatrix splittings are an example of **global bifurcations**, where the changes on phase space are extended (global).

In general, bifurcations are too complicated to be understood in full. However, much insight is available without the need of a highly elaborated mathematical structure, starting by the study of the change of stability of fixed points while varying some system parameters.

In this Chapter, we will focus on techniques to simplify the study of local bifurcations (Centre Manifold Theory, inspired in [**Car81**]), leaving for the next Chapter the specific computation of the simplest local bifurcations. We will finally address global bifurcations and *chaos* in Chapter 8.

6.2. The Bifurcation Programme for Fixed Points of Flows

The setup for our study is a family of dynamical systems $\dot{x} = f(x, \mu)$ defined on $\mathbb{R}^n$ (or on a compact ball of $\mathbb{R}^n$ when necessary) and depending on only one real parameter μ. We will assume that f is of class C^k in both x and μ, for k large enough. This means that whenever we need to take one or many derivatives of f, such computation will make sense.

Let us recall some results from Chapter 5, especially Remark 5.12. Far away from fixed points, the Straightening–out Theorem assures that the dynamics will be simple: Up to a coordinate transformation, flow lines are parallel "straight" lines. It is in the local vicinity of fixed points where most of the structure becomes visible. The Theorems of Ladis (Theorem 5.8) and Hartman–Grobman (Theorem 5.6) help us to address the problem. Suppose we have a hyperbolic fixed point for some value of the parameter: $f(x_0, \mu_0) = 0$. Hyperbolicity implies that no eigenvalue of the Jacobian at the fixed point has zero real part, in particular, no eigenvalue is zero and $\det D_x f(x_0, \mu_0) \neq 0$. Hence, by the Implicit Function Theorem (Dini's Theorem) of Calculus, there exists a unique curve of fixed points $x(\mu)$, with $x(\mu_0) = x_0$, at least for a small parameter interval around μ_0. Since we assumed C^k differentiability, the curve will be smooth and also the Jacobian $D_x f(x(\mu), \mu)$ will be a smooth matrix function of μ and hence still hyperbolic for a sufficiently small interval around μ_0 (as long as no real part of an eigenvalue changes sign). Hartman–Grobman Theorem assures that the local dynamics along the curve $x(\mu)$ is equivalent to the linearised dynamics. Ladis Theorem assures that all linearised dynamics are equivalent, so locally, the dynamics is one and the same (except for a coordinate change) along the curve, as long as all invoked theorems hold.

The dynamical changes can only arise if the hyperbolicity condition breaks down. We still would like to describe the problem in such a way that the dynamical changes are separated from other modifications that can be compensated for with e.g., a diffeomorphism or a change of coordinates. It is reasonable to ask whether there is a coordinate choice making some sort of splitting of phase space into an unaffected part and a bifurcating part. We will first address the dynamical situation for one dynamical system with a non-hyperbolic fixed point and subsequently extend the method to study parametric families of dynamical systems. The situation is described by the following Theorem, which we state without proof:

Theorem 6.1 (Sositaisvili). [**Š73**] *Suppose that a differential equation in $\mathbb{R}^n$ with C^2 right hand side having a singular point at the origin can be written in the following way*

$$\begin{aligned}
\dot{x} &= Ax + g_1(x, y, z) \\
\dot{y} &= B_+ y + g_2(x, y, z) \\
\dot{z} &= B_- z + g_3(x, y, z),
\end{aligned}$$

where A, B_+ and B_- are the (Jordan) blocks of the linearisation matrix corresponding to eigenvalues having zero, positive and negative real parts respectively and the functions g_1, g_2 and g_3 are zero at the origin along with all their first derivatives. Then, in a neighbourhood of the singular point 0, the equation under consideration is locally topologically equivalent to the following system:

$$\begin{aligned}
\dot{x} &= A\,x + f(x) \\
\dot{v} &= B_+(x)\,v \\
\dot{w} &= B_-(x)\,w
\end{aligned}$$

(the equivalence is in fact C^k) where $f(0) = f'(0) = 0$ and $B_\pm(0)$ are the original blocks $B_\pm$.

REMARK 6.1. The intuition behind this Theorem is as follows. The vw-part may be seen as a hyperbolic system parameterised by x and hence locally conjugate to its linear part. The crucial part of the Theorem is the change of coordinates, separating vw from x, where the substantial dynamical changes take place.

Instead of proving Sositaisvili's Theorem, we will get a closer look at this separation and its consequences, proving other, more detailed, theorems, inspired in [**Car81**].

6.3. Centre Manifold Theorems

First: What is a manifold?

DEFINITION 6.1. A n-manifold is a set of points (or a topological Space, see Appendix, Definition A.7 for details) where each point has a neighbourhood homeomorphic to an open set of $\mathbb{R}^n$. For example, a straight line of $\mathbb{R}^n$ is a 1-manifold.

REMARK 6.2. Dynamically interesting manifolds are those that are invariants of the flow, i.e., where $x \in M \Rightarrow \phi_t(x) \in M$. The graph of a smooth function $y = h(x)$, $y \in \mathbb{R}^p$, $x \in \mathbb{R}^q$ defines an **invariant manifold** of a dynamical system in $\mathbb{R}^{p+q}$ whenever the dynamics of an initial condition $(x_0, h(x_0))$ remains within the graph throughout the dynamics, i.e., if $\phi_t(x_0, h(x_0)) = (x(t), h(x(t)))$.

REMARK 6.3. For a flow of a Lipschitz vector field, all trajectories are invariant manifolds because of the unicity of solutions in Picard Theorem. In any case, for both flows and maps, the stable manifold U_s of a fixed point p is the invariant manifold where $phi_t(x) \to p$ for $t \to \infty$, and the unstable manifold is the corresponding invariant manifold $\phi_t(x) \to p$ for $t \to -\infty$.

Let us begin the analysis by considering motivating examples.

6.3.1. Dynamics on an Invariant Manifold. Consider the following system defined in $\mathbb{R}^2$:

$$\begin{aligned} \dot{x} &= ax^3 \\ \dot{y} &= -y \end{aligned}$$

The coordinate axes $I_1 = \{y = 0\}$ and $I_2 = \{x = 0\}$ are invariant sets for the dynamics, since any initial condition on e.g., I_1 is of the type $(x_0, y_0) = (x, 0)$, which plugged into the equations yields a solution in which y is identically zero. Hence, orbits with initial condition on I_1 do not leave I_1. The same holds for I_2. Both sets are examples of invariant manifolds. They are of course invariant sets and they are also manifolds (in this example each set is equivalent to $\mathbb{R}^1$). The set I_2 can be recognised as belonging to the stable invariant set (stable manifold) of the fixed point at the origin (the nature of I_1 depends on the value of a).

Proposition 6.1. *The system*

$$\begin{aligned}\dot{x} &= y\\ \dot{y} &= 0\\ \dot{z} &= -z + x^2 + y^2\end{aligned}$$

has an invariant manifold $z = h(x, y)$ *such that* $h(0, 0) = 0$ *and* $\nabla h(0, 0) = 0$.

Proof. The first two equations decouple from the third and have the solutions $x(t) = x_0 + y_0 t$, $y(t) = y_0$. We look for a function $z(t) = h(x_0 + y_0 t, y_0)$ satisfying the equation

$$\frac{\mathrm{d}h}{\mathrm{d}t} = -h + x^2 + y^2.$$

We seek for a solution that does not lie along the stable manifold of the fixed point (the z axis) but one that is tangent to the xy-plane at the origin (the non-hyperbolic local directions near the origin). Since the stable manifold near the origin behaves as $z(t) = e^{-t}$, we impose to h the additional condition

$$\lim_{t \to -\infty} h(x_0 + y_0 t, y_0) e^t = 0.$$

After multiplying the differential equation by e^t and performing partial integration we obtain $h(x, y) = (x - y)^2 + 2y^2$. The reader can verify that it is invariant for the dynamics and tangent to the xy-plane at the origin. □

Remark 6.4. The manifold $h(x, y)$ is an example of a **centre manifold**, an invariant manifold tangent to the non-hyperbolic linearised subspace at the fixed point.

Consider the coordinate transformation $(x, y, z) \to (x, y, w)$, where $w = z - h(x, y)$. The dynamics in the new coordinates reads

$$\begin{aligned}\dot{x} &= y\\ \dot{y} &= 0\\ \dot{w} &= -w.\end{aligned}$$

to be compared with the statement in Theorem 6.1 (Sositaisvili). Indeed, the separation provided by Sositaisvili's Theorem requires the identification of a manifold, depending in this example on the xy-coordinates associated to a non-hyperbolic linearisation, where all the local dynamical complexity takes place. The other directions of phase space, transversal to this manifold, behave locally in an essentially linear fashion. The formalisation of this idea is given by the following three theorems.

6.3.2. Centre Manifold Theorems. We will state and prove the theorems for the case of a fixed point at the origin with empty unstable linearisation subspace. The general case is a more or less straightforward generalisation.

THEOREM 6.2. [**Car81**] *Let f, g be C^2 functions vanishing at the origin along with their derivatives, the $n \times n$ matrix A have eigenvalues with zero real part and the $m \times m$ matrix B have eigenvalues with negative real parts. The system*

$$\begin{aligned} \dot{x} &= Ax + f(x,y) \\ \dot{y} &= By + g(x,y) \end{aligned} \tag{6.1}$$

has a local C^2 invariant (centre) manifold $y = h(x)$, $|x| < \delta$, with $h(0) = 0$, $Dh(0) = 0$. The flow on the centre manifold is governed by the equation

$$\dot{x} = Ax + f(x, h(x)). \tag{6.2}$$

PROOF. The last statement follows from the first, since if we have an invariant manifold $y = h(x)$, then on the manifold the first equation 6.1 becomes eq.(6.2) while the second equation 6.1 implicitly defines the manifold as

$$\frac{dh}{dx}\dot{x} = \frac{dh}{dx}(Ax + f(x, h(x))) = Bh(x) + g(x, h(x))$$

The proof proceeds by recasting this second equation as a fixed point problem.

First we replace our problem by a locally equivalent problem vanishing outside a compact region in x. Let $\phi(x)$ be a C^∞ function with $\phi(x) = 1$ for $|x| \leq 1$ and $\phi(x) = 0$ for $|x| \geq 2$. Further, for $\epsilon > 0$ consider $F(x,y) = f(x\phi(\frac{x}{\epsilon}), y)$ and $G(x,y) = g(x\phi(\frac{x}{\epsilon}), y)$. We will prove the theorem for the system

$$\begin{aligned} \dot{x} &= Ax + F(x,y) \\ \dot{y} &= By + G(x,y), \end{aligned}$$

for small enough ϵ. This system coincides with eq.(6.1) in a small neighbourhood of the origin.

For $q > 0$ consider the space X of functions $h : \mathbb{R}^n \to R^m$ with Lipschitz constant $p > 0$ such that $h(0) = 0$ and $|h(x)| \leq q$. X is a complete metric space, using the sup norm (see Proposition A.1 in the Appendix). We will analyse our system when y is some function from X.

Let $x(t, x_0, h)$ be the unique solution of

$$\dot{x} = Ax + F(x, h(x)), \qquad x(0, x_0, h) = x_0,$$

for $h \in X$. This solution exists for all $t \in \mathbb{R}$ because of the bounds on F and h. We can consider now an operator $T : X \to X$ defined by

$$T(h(x_0)) = \int_{-\infty}^{0} e^{-Bs} G(x(s, x_0, h), h(x(s, x_0, h))ds$$

If h is a fixed point of T then $y = h(x)$ is an invariant manifold of the system and we are almost finished. We will proceed now by showing that T is a contraction on X for small p, q and ϵ. We will estimate bounds for all the ingredients entering the definition of T and finally verify that it is possible to have the Lipschitz constant of T smaller than unity.

Since F and G are C^2 functions vanishing at the origin along with their derivatives, we have that for some continuous function $k(\epsilon)$ with $k(0) = 0$,

$$\begin{aligned}
&|F(x, y)| + |G(x, y)| \leq \epsilon k(\epsilon) \\
&|F(x, y) - F(x', y')| \leq k(\epsilon)\left(|x - x'| + |y - y'|\right) \\
&|G(x, y) - G(x', y')| \leq k(\epsilon)\left(|x - x'| + |y - y'|\right)
\end{aligned}$$

for any $x, x' \in \mathbb{R}^n$ and $|y|, |y'| < \epsilon$. Also, there exist positive constants β, C such that for $s \leq 0$ and any y, $|e^{-Bs}y| \leq Ce^{\beta s}|y|$ and finally for $r > 0$ and some constant $M(r)$ we have that $|e^{As}x| \leq M(r)e^{r|s|}|x|$. For $q < \epsilon$ any function h we consider will fit these bounds and we can proceed with our estimates of the ingredients in the integral defining T. We start with

$$|T(h(x_0))| \leq \frac{C}{\beta}\epsilon k(\epsilon).$$

From eq.(6.3.2), the solution $x(t, x_0, h)$ reads

$$x(t, x_0, h) = e^{At}\left(x_0 + \int_0^t e^{-As} F(x(s, x_0, h), h)\, ds\right)$$

hence, for $t \leq 0$ (this is what we need, considering the definition of T),

$$\begin{aligned}
|x(t, x_0, h) - x(t, x_1, h)| &\leq M(r)e^{-rt}|x_0 - x_1| \\
&\quad + (1 + p)M(r)k(\epsilon)\int_t^0 e^{r(s-t)}|x(s, x_0, h) - x(s, x_1, h)|\, ds
\end{aligned}$$

(we use the bounds on F, h and $e^{At}x$ described above). Hence,

$$\begin{aligned}
e^{rt}|x(t, x_0, h) - x(t, x_1, h)| &\leq M(r)|x_0 - x_1| \\
&\quad + (1 + p)M(r)k(\epsilon)\int_t^0 e^{rs}|x(s, x_0, h) - x(s, x_1, h)|\, ds
\end{aligned}$$

Using Gronwall's inequality (see Appendix A.1.3) with the functions $\psi(t) = 1$ and $\phi(t) = e^{rt}|x(t, x_0, h) - x(t, x_1, h)|$, we have

$$e^{rt}|x(t, x_0, h) - x(t, x_1, h)| \leq M(r)|x_0 - x_1|e^{(1+p)M(r)k(\epsilon)|t|}$$

and finally, letting $\gamma = r + (1+p)M(r)k(\epsilon)$,

$$|x(t,x_0,h) - x(t,x_1,h)| \leq M(r)|x_0 - x_1|e^{-\gamma t}.$$

The next step is to use the above preparation to estimate the size of $|T(h(x_0)) - T(h(x_1))|$. Using the definition of T and the bounds for G and $|e^{-Bs}y|$ we obtain

$$|T(h(x_0)) - T(h(x_1))| \leq \frac{Ck(\epsilon)M(r)(1+p)}{\beta - \gamma}|x_1 - x_0|$$

provided that ϵ and r are small enough so that $\beta > \gamma$. Note that there is no restriction here, since for a given small r, ϵ can be chosen independently so that $k(\epsilon)$ is sufficiently small (even after multiplication by $(1+p)M(r)$) and hence γ can be made as small as needed. This estimate is required to verify that $T(h)$ is also a Lipschitz function with Lipschitz constant smaller than ϵ, which means that T is an operator mapping elements of X to other elements of X.

Now we need another estimate for x in order to address the final computation. Again, for $t \leq 0$,

$$\begin{aligned}
&|x(t,x_0,h_1) - x(t,x_0,h_0)| \\
&\leq M(r)\int_t^0 e^{r(s-t)}|F(x(s,x_0,h_1),h_1) - F(x(s,x_0,h_0),h_0)|\,ds \\
&\leq M(r)k(\epsilon)\int_t^0 e^{r(s-t)}\left(|x(s,x_0,h_1) - x(s,x_0,h_0)| + ||h_1 - h_0||\right)\,ds
\end{aligned}$$

Letting now $\phi(t) = e^{rt}|x(t,x_0,h_1) - x(t,x_0,h_0)|$, we write

$$\phi(t) \leq \frac{M(r)k(\epsilon)}{r}||h_1 - h_0|| + M(r)k(\epsilon)(1+p)\int_t^0 \phi(s)\,ds$$

(the factor $(1+p)$ may not seem necessary, but it renders the estimate similar to the previous one) and using Gronwall's inequality again, we obtain (same γ as in the previous estimate)

$$|x(t,x_0,h_1) - x(t,x_0,h_0)| \leq \frac{M(r)k(\epsilon)}{r}||h_1 - h_0||e^{-\gamma t}.$$

Now we are ready to compute the Lipschitz constant of T:

$$\begin{aligned}
&|T(h_1(x_0)) - T(h_0(x_0))| = \\
&\quad = \int_{-\infty}^0 e^{-Bs}(G(x(s,x_0,h_1),h_1) - G(x(s,x_0,h_0),h_0))\,ds \\
&\quad \leq Ck(\epsilon)\int_{-\infty}^0 e^{\beta s}\left(\frac{M(r)k(\epsilon)}{r}e^{-\gamma s}||h_1 - h_0|| + ||h_1 - h_0||\right)\,ds \\
&\quad \leq Ck(\epsilon)\left(\frac{1}{\beta} + \frac{M(r)k(\epsilon)}{r(\beta - \gamma)}\right)||h_1 - h_0||.
\end{aligned}$$

This constant can be made smaller than unity by taking r and ϵ sufficiently small. Hence, the unique fixed point of T is a Lipschitz centre manifold. Moreover, the operator T also is a contraction restricted to Lipschitz differentiable functions, hence the fixed point is C^1. To prove that h is also C^2, note that both F and G are C^2 and hence the trajectories are C^2 regarded as functions of their initial condition. Therefore, T is a contraction operator also when restricted to C^2 functions $h(x)$. □

The method of proof of the previous Theorem, namely that of recasting the problem as a fixed point equation is used in the following two Theorems as well. Although the underlying idea is relatively simple, the estimates to compute the Lipschitz constant can get very involved.

Theorem 6.3. [**Car81**] *(a) If the zero solution of eq.(6.2) is stable (asymptotically stable, unstable) then the zero solution of eq.(6.1) is stable (asymptotically stable, unstable).*

(b) Suppose the zero solution of eq.(6.1) is stable. Then for x_0, y_0 sufficiently small there exists a solution $u(t)$ of eq.(6.2) such that $x(t) = u(t) + O(e^{-\gamma t})$ and $y(t) = h(u(t)) + O(e^{-\gamma t})$ for $t \to \infty$, where γ is a positive constant depending only on B.

This Theorem completes the picture given by Sositaisvili's Theorem, since the dynamics outside the Centre manifold converges to the manifold exponentially in time.

We skip the proof of the above Theorem and focus directly on the next one, where a method of approximating the Centre Manifold is presented.

Theorem 6.4 (Approximation to the Centre Manifold). [**Car81**] *Suppose that $\phi(0) = 0$, $D\phi(0) = 0$ and that (cf. eq.(6.1))*

$$D\phi(x)\,(Ax + f(x, \phi(x))) - B\phi(x) - g(x, \phi(x)) = O(|x|^q)$$

as $|x| \to 0$ with $q > 1$. Then as $|x| \to 0$ the centre manifold satisfies

$$|h(x) - \phi(x)| = O(|x|^q).$$

Proof. Since we are looking for an approximation near the origin, let θ be a C^1 function with compact support (i.e., vanishing outside a ball) that coincides with ϕ for $|x|$ sufficiently small. We build with this function the approximant

$$N_\theta(x) = D\theta(x)\,(Ax + F(x, \theta(x))) - B\theta(x) - G(x, \theta(x)),$$

where F and G are as in Theorem 6.2. Note that $N_\theta(x) = O(|x|^q)$ near the origin, while for the center manifold we have $N_h = 0$. Consider now the mapping $S : Y \to X$, defined as $S(z) = T(z + \theta) - \theta$. Here $T : X \to X$ is the operator of Theorem 6.2 and the function $z(x)$ is taken from a closed subset Y of X defined as

$$Y = \{z \in X : |z| \leq K|x|^q,\ x \in \mathbb{R}^n\}$$

If we manage to specify the constant K in such a way that $S(Y) \subset Y$, then S is a contraction mapping and has a unique fixed point. In such a case, $z+\theta = h$, is the unique fixed point of T and consequently, within the support of θ we have, $|h(x)-\phi(x)| = |h(x)-\theta(x)| \leq |z(x)| \leq K|x|^q$ and the Theorem is proved.

Let us begin by rewriting S using the definition of T. For $z \in Y$ let $x(t, x_0)$ be the solution of $\dot{x} = Ax + F(x, z(x)+\theta(x))$ with the condition $x(0, x_0) = x_0$. Then,

$$T(z+\theta)(x_0) = \int_{-\infty}^{0} e^{-Bs} G(x(s,x_0), z(x(s,x_0)) + \theta(x(s,x_0)))ds$$

First, we recast $\theta(x_0)$ as,

$$\begin{aligned}\theta(x_0) &= \int_{-\infty}^{0} \frac{d}{ds}\left(e^{-Bs}\theta(x(s,x_0))\right) ds \\ &= \int_{-\infty}^{0} e^{-Bs}\left(\frac{d}{ds}\theta(x(s,x_0)) - B\theta(x(s,x_0))\right) ds\end{aligned}$$

The expression in brackets can be computed as (we do not write the dependency on s, x_0 to lighten the notation):

$$\begin{aligned}B\theta(x) - \frac{d}{ds}\theta(x) &= B\theta(x) - \theta'(x)\left(Ax + F(x, z(x)+\theta(x))\right) \\ &= -N_\theta(x) - G(x,\theta(x)) + \theta'(x)\left(F(x,\theta(x)) - F(x, z(x)+\theta(x))\right)\end{aligned}$$

Now we collect the above expressions into

$$(Sz)(x_0) = T(z+\theta)(x_0) - \theta(x_0) = \int_{-\infty}^{0} e^{-Bs} Q(x(s,x_0), z(x(s,x_0)))ds$$

where

$$Q(x,z) = G(x,\theta+z) - G(x,\theta) - N_\theta(x) + \theta'(x)(F(x,\theta) - F(x, z+\theta)).$$

Next step is to estimate the size of different quantities as in Theorem 6.2. We use indeed the same Lipschitz constants for F and G. We have, for some constant $R > 0$,

$$\begin{aligned}|Q(x,z)| &\leq |Q(x,0)| + |Q(x,z) - Q(x,0)| = |N_\theta(x)| + k(\epsilon)|z(x)| \\ &\leq (R + Kk(\epsilon))|x|^q.\end{aligned}$$

Following Theorem 6.2 again, the computations starting in eq.(6.3.2) and Gronwall's inequality, we have, for $t \leq 0$ and $\gamma = r + 2M(r)k(\epsilon)$,

$$|x(t,x_0)| \leq M(r)|x_0|e^{-\gamma t}.$$

With the two previous estimates and the bound $|e^{-Bs}y| \leq e^{\beta s}|y|$ for $s \leq 0$ (also from Theorem 6.2), we can estimate

$$|(Sz)(x_0)| \leq C\int_{-\infty}^{0} e^{\beta s}(R + Kk(\epsilon))M(r)^q|x_0|^q e^{-q\gamma s}.$$

Hence,

$$|(Sz)(x_0)| \leq \left(C(R+Kk(\epsilon))\frac{M(r)^q}{\beta - q\gamma}\right)|x_0|^q = C_1|x_0|^q.$$

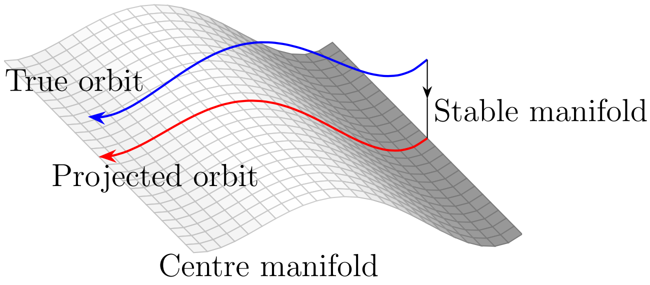


Figure 6.2. Schematic view of a true orbit and its "companion", the projected orbit on the centre manifold.

Taking ϵ and r sufficiently small we can assure $\beta - q\gamma > 0$ and further, since $k(\epsilon)$ goes to zero with ϵ independently of the other bounds, we can have $C_1 \leq K$, and the Theorem is proved. □

Remark 6.5. The last three theorems have a fundamental importance to understand the qualitative dynamics near a non-hyperbolic fixed point with stable local hyperbolic subspace. The dynamics of initial conditions sufficiently close to the fixed point "mimic" the dynamics on the Centre Manifold ("differences" among both dynamics decrease exponentially). On this manifold, the dynamics is described by a reduced equation (in a lower dimensional space), and it can be approximated by a polynomial vector field. Hence, these three theorems accomplish the programme of Sositaisvili's Theorem, providing additionally an approximation scheme for the computation of $h(x)$.

Remark 6.6. The Centre Manifold need not be unique. Indeed, Theorem 6.2 starts by replacing the original problem with a truncation. Within the truncation, the fixed point theorem is used and solutions are unique, but the arbitrariness of the truncation allows us to construct different Centre Manifolds for the same situation. However, if $h_1(x)$, $h_2(x)$ are two Centre Manifolds, because of Theorem 6.4 we have that $|h_1(x) - h_2(x)| = O(|x|^q)$ as $x \to 0$ for any $q > 1$. If f and g are C^k, $k > 2$, the Centre Manifold is also C^k, however if f and g are analytic, the Centre Manifold need not be analytic.

Example 6.1. Let us illustrate the ideas behind the centre manifold for the case with no unstable directions in the linearisation. In Figure 6.2 a true orbit of the system is depicted in full line. The vertical lines represent the product of this orbit with the (1-d) stable manifold. The blue dotted curve represents the projection of the true orbit along the stable manifold onto the centre manifold. The true orbit and its projection approach each other exponentially fast.

6.4. Centre Manifold Theorems for Maps

There exist three parallel theorems for maps with non-hyperbolic fixed points. The theorems are essentially a direct translation of their differential equation versions. We state the theorems without proof.

Consider the system

$$\begin{aligned} x_{n+1} &= Ax_n + f(x_n, y_n) \\ y_{n+1} &= By_n + g(x_n, y_n), \end{aligned}$$

where A and B are square matrices, the eigenvalues of A have modulus one and those of B modulus smaller than one, and f and g are C^2 functions vanishing at the origin together with their first derivative. Further, for functions $\phi(x)$ such that $\phi(0) = 0$, $\phi'(0) = 0$ we define

$$(M\phi)(x) = \phi(Ax + f(x, \phi(x))) - B\phi(x) - g(x, \phi(x)).$$

For a Centre Manifold h as in the next Theorem we have $Mh = 0$.

THEOREM 6.5. [**Car81**] *There exists a C^2 function $y = h(x)$ such that $h(0) = 0$, $h'(0) = 0$ and for $|x| \leq \epsilon$ h satisfies*

$$\begin{aligned} x_{n+1} &= Ax_n + f(x_n, h(x_n)) \\ h(x_{n+1}) &= Bh(x_n) + g(x_n, h(x_n)). \end{aligned}$$

THEOREM 6.6. [**Car81**] *(a) If the zero solution of $u_{n+1} = Au_n + f(u_n, h(u_n))$ is stable (asymptotically stable, unstable), then the zero solution of eq.(6.4) is stable (asymptotically stable, unstable).*
(b) If the zero solution of eq.(6.4) is stable, then for a solution (x_n, y_n) of eq.(6.4) with initial condition (x_0, y_0) sufficiently small, there exists a solution u_n of $u_{n+1} = Au_n + f(u_n, h(u_n))$ such that $|x_n - u_n| \leq K\beta^n$ and $|y_n - h(x_n)| \leq K\beta^n$ for all n, $K > 0$ and $0 < \beta < 1$.

THEOREM 6.7. [**Car81**] *If $\phi \in C^1$ and $(M\phi)(x) = O(|x|^q)$ as $x \to 0$ for some $q > 1$ then $|h(x) - \phi(x)| = O(|x|^q)$ as $x \to 0$.*

6.5. Computation of Centre Manifold Approximations [GH86, SNM96]

We will produce a chain of systematic approximations to the Centre Manifold. To fix ideas, let us assume that f and g can be derivated arbitrarily many times (a situation frequently assumed in applications) and expanded in power series. Theorems 6.4 and 6.7 suggest a procedure.

6.5.1. Centre Manifold Approximation for Flows. For each integer $k \geq 1$, let us call $\phi_k(x)$ the approximation to $h(x)$ given by Theorem 6.4 taking $q = k + 1$. Since f and g are at least quadratic, it is clear that $\phi_1(x) = 0$ is an approximation to the Centre Manifold with errors $O(|x|^2)$. It suffices to plug it in eq.(6.3.2). Consider further the chain of approximations $\phi_{k+1}(x) = \phi_k(x) + \Delta_{k+1}(x)$, with $\Delta_{k+1}(x)$ a homogeneous polynomial of degree $k + 1$ in x. Our goal is to choose

$\Delta_{k+1}(x)$ in such a way that Theorem 6.4 is fulfilled to order $q = k+2$, so that the successive approximations $\phi_m(x)$ differ from the correct Centre Manifold $h(x)$ in $O(|x|^{m+1})$. We perform the construction inductively, starting with $\phi_1(x) = 0$. Assuming that a good choice of Δ_k and ϕ_k has been done up to order k, we seek for a $\phi_{k+1}(x)$ that satisfies eq.(6.3.2):

$$\begin{aligned}&\frac{d}{dx}(\phi_k(x) + \Delta_{k+1}(x))\,(Ax + f(x, \phi_k(x) + \Delta_{k+1}(x)) = \\ &B\,(\phi_k(x) + \Delta_{k+1}(x)) + g(x, \phi_k(x) + \Delta_{k+1}(x)) + O(|x|^{k+2}),\end{aligned}$$

which can be recast as

$$\begin{aligned}&\frac{d}{dx}(\Delta_{k+1}(x))Ax - B\Delta_{k+1}(x) = \\ &B\phi_k(x) + g(x, \phi_k(x)) - \frac{d\phi_k(x)}{dx}\,(Ax + f(x, \phi_k(x))) + O(|x|^{k+2}),\end{aligned}$$

Some terms involving $\Delta_{k+1}(x)$ (inside f and g) appear to be missing in this last equation when compared with the previous one. Actually, because of the form of f and g, those terms are of order $O(|x|^{k+2})$ or higher and need not be written explicitly. Both equations are indeed equivalent. Moreover, the RHS of the last equation has no contributions of degree k or lower, these terms have been taken care of by the inductive procedure, starting at $k = 1$. Hence the explicitly written part of the RHS is a known homogeneous polynomial of degree $k + 1$, while the LHS is also a homogeneous polynomial of the same degree and unknown coefficients. By equating both polynomials, we solve for the unknown coefficients. It remains to prove that the equation always has unique solution. This is equivalent to stating that the equation $\frac{d}{dx}(\Delta_{k+1}(x))Ax - B\Delta_{k+1}(x) = 0$ has the unique solution $\Delta_{k+1}(x) = 0$. When A and B are diagonalisable, the proof is rather straightforward, while it is more involved with technicalities for the general case.

Example 6.2. We compute approximations to the Centre Manifold for the system

$$\begin{aligned}\dot{x} &= y \\ \dot{y} &= -y + yx + x^2\end{aligned}$$

which has a non-hyperbolic fixed point at the origin. First we move to a coordinate system where the linearisation at the origin is diagonal, namely: $u = x + y$ and $v = y$ [1]. We rewrite:

$$\begin{aligned}\dot{u} &= u(u - v) \\ \dot{v} &= -v + u(u - v),\end{aligned}$$

[1] Eigenvectors are defined up to a constant. Hence there are different (although equivalent) choices for the subsequent calculations. For example, $u = x+y$, $v = -y$ work equally well.

looking for approximations $v = \phi_k(u)$. $\phi_1(u) = 0$ fulfills Theorem 6.4 with errors $O(|x|^2)$. Letting $\phi_{k+1}(u) = \phi_k(u) + a_{k+1}u^{k+1}$, we compute eq.(6.3.2) for this case, namely:

$$\begin{aligned}&\left(\frac{d\phi_k}{du} + a_{k+1}(k+1)u^k\right)(u(u - \phi_k(u) - a_{k+1}u^{k+1})) = \\ &- \left(\phi_k(u) + a_{k+1}u^{k+1}\right) + u(u - \phi_k(u) - a_{k+1}u^{k+1}).\end{aligned}$$

For $k = 1$ we have $2a_2u^2(u^2 - a_2u^3) = -a_2u^2 + u^2 - a_2u^3$. The equation is satisfied with errors $O(|x|^3)$ by letting $a_2 = 1$ and $\phi_2(u) = u^2$. For $k = 2$ we have $(2u+3a_3u^2)(u^2-u^3-a_3u^4) = -(u^2+a_3u^3)+u^2-u^3-a_3u^4$, giving $a_3 = -3$ and $\phi_3(u) = u^2 - 3u^3$. A subsequent step gives $a_4 = 11$ and $\phi_4(u) = u^2 - 3u^3 + 14u^4$. The dynamic splits approximately into an exponential decay to the manifold $v = \phi_4(u) + O(|x|^5)$ and a dynamics on the manifold given by $\dot{u} = u^2 - u^3 + 3u^4 - 14u^5 + O(|x|^6)$.

6.5.2. Centre Manifold Approximation for Maps. The equation to consider now is eq.(6.4). Again, we inductively construct successive approximations to h for $k \geq 1$ of the form $\phi_{k+1}(x) = \phi_k(x) + \Delta_{k+1}(x)$, with $\phi_1(x) = 0$ and $\Delta_{k+1}(x)$ a homogeneous polynomial of degree $k+1$ in x satisfying eq.(6.4) up to order $k+1$, after suitable rewriting:

$$\begin{aligned}\phi_{k+1}(x_{n+1}) &= \phi_{k+1}(Ax_n + f(x_n, \phi_{k+1}(x_n))) \\ &= (\phi_k + \Delta_{k+1})(Ax_n + f(x_n, \phi_{k+1}(x_n))) \\ &= B(\phi_k(x_n) + \Delta_{k+1}(x_n)) + g(x_n, \phi_{k+1}(x_n)) + O(|x|^{k+2}).\end{aligned}$$

Let us work now with the last equality. Using the same argument as before, $f(x_n, \phi_{k+1}(x_n)) = f(x_n, \phi_k(x_n)) + O(|x|^{k+2})$, and similarly for g. Also, $\Delta_{k+1}(Ax_n + f(x_n, \phi_{k+1}(x_n))) = \Delta_{k+1}(Ax_n) + O(|x|^{k+2})$. Hence,

$$\begin{aligned}\Delta_{k+1}(Ax_n) - &B\Delta_{k+1}(x_n) \\ &= B\phi_k(x_n) + g(x_n, \phi_k(x_n)) - \phi_k(Ax_n + f(x_n, \phi_k(x_n))) \\ &\quad + O(|x|^{k+2}).\end{aligned}$$

Induction and Theorem 6.7 assure that the RHS has no contributions of order $|x|^k$ or lower. Both $\Delta_{k+1}(x)$ and the LHS of the equation are unknown homogeneous polynomials of degree $k+1$ while the RHS is a known homogeneous polynomial of the same degree, fully computable after having solved step k. Equating both polynomials we obtain the coefficients of $\Delta_{k+1}(x)$. Again, it remains to prove that the equation always has unique solution. This is equivalent to stating that the equation $\Delta_{k+1}(Ax_n) - B\Delta_{k+1}(x_n) = 0$ has the unique solution $\Delta_{k+1}(x) = 0$. When A and B are diagonalisable, the proof is, again, rather straightforward, while it is more involved with technicalities for the general case.

Example 6.3. The system

$$\begin{aligned} x_{n+1} &= x_n + x_n y_n &&= x_n + f(x_n, y_n) \\ y_{n+1} &= y_n/2 - {x_n}^2 &&= y_n/2 + g(x_n, y_n) \end{aligned}$$

has a non-hyperbolic fixed point at the origin. The linearisation is already in diagonal form with eigenvalues 1 and 1/2. Letting $k = 1$, we have $\phi_1(x) = 0$, while eq.(6.4) reads $a_2 x^2/2 = -x^2 + O(|x|^3)$, obtaining $a_2 = -2$ and $\phi_2(x) = -2x^2$. We notice that $a_3 = 0$ since there are no third-order terms in the defining equation, while $a_4 = -16$ and the Centre Manifold reads $h(x) = -2x^2 - 16x^4 + O(|x|^5)$. The dynamics on the manifold is approximated by $x_{n+1} = x_n - 2x_n^3 - 16x_n^5 + O(|x_n|^6)$.

6.6. Relation of Centre Manifold Theory with Bifurcations

The moment has arrived to connect the Centre Manifold Theory with the bifurcation discussion at the beginning of the Chapter. The goal was to be able to study what happens on a parameter dependent family of dynamical systems when some linearisation eigenvalues become non-hyperbolic. Since this phenomenon affects a few eigenvalues at a time, it is desirable to study it on a reduced space, covering only the non-hyperbolic part of the system without being forced to carry the hyperbolic part behind.

Sositaisvili's Theorem as well as the reduction to the Centre Manifold hold for just one dynamical system, with fixed vector field. However, the bifurcation problem deals with a parametric family of vector fields, where the fixed point and its linearisation eigenvalues depend on the parameter. At some special parameter value, the system ceases to be hyperbolic and the goal is to understand the whole transition.

Centre Manifold theory is particularly adapted to study such a problem. The strategy is to extend the dynamical equations with an additional equation for the parameter, namely $\dot{\mu} = 0$ or $\mu_{n+1} = \mu_n$ for maps (of dimension p if μ consists of p real-valued parameters). In this way, we are treating the parameters as extra coordinates and the additional equation indicates that for each dynamical system the parameters are actually constant. The enhanced system treats then the whole family at once, at the cost of increasing the dimensionality.

At the special value μ_0, both the original system and the extended system have a non-hyperbolic fixed point, while the Centre Manifold for the extended system (of dimension $d + p$, where $x \in \mathbb{R}^d$) now includes the parameter dependency near the fixed point on the x equation for small deviations $\mu - \mu_0$. The hyperbolic part, collected in the vw (or y) coordinates, has been decoupled from the non-hyperbolic part. In the extended description, the non-hyperbolic part has a μ-dependency intertwined with the x-dependency. On this lower-dimensional system, with $d + p$ equations (where p of them are trivial), we will address in the next Chapter the study of the bifurcation transition. The extended description does not change the fact that the original system was assumed to become non-hyperbolic only for $\mu = \mu_0$, only that since now μ behaves essentially as a coordinate, we can study the transition hyperbolic → non-hyperbolic while varying μ. We address the details of this approach in the next Chapter.

6.7. Exercises

EXERCISE 6.1. Compute a reduction to the Centre Manifold up to order 5, i.e., error of size $O(|x|^6)$, for the system

$$\begin{aligned} \dot{x} &= xy - x^3 + xy^2 \\ \dot{y} &= -y + x^2 + x^2 y, \end{aligned}$$

near the origin. Compute a dynamical equation on the manifold. Is the origin stable?

EXERCISE 6.2. Compute a reduction to the Centre Manifold for the system

$$\begin{aligned} \dot{x} &= 10(y - x) \\ \dot{y} &= x - y - xz \\ \dot{z} &= xy - \frac{8}{3}z \end{aligned}$$

EXERCISE 6.3. Compute a reduction to the Centre Manifold for the system

$$\begin{aligned} x_{n+1} &= -x_n - {y_n}^2 \\ y_{n+1} &= -\frac{1}{2}y_n + x_n(y_n + x_n) \end{aligned}$$

EXERCISE 6.4. In the system

$$\begin{aligned} \dot{x} &= y \\ \dot{y} &= \mu x - y - x^2 \end{aligned}$$

consider μ as a third variable with dynamics $\dot{\mu} = 0$. Compute the reduction to the Centre Manifold (that now is 2-dimensional) and the approximated dynamics on the manifold.

EXERCISE 6.5. In the system

$$\begin{aligned} \dot{x} &= \mu x - y \\ \dot{y} &= x + \mu y + y^2 + x^2 \end{aligned}$$

consider μ as a third variable with dynamics $\dot{\mu} = 0$. Compute the reduction to the Centre Manifold (that now is 2-dimensional) and the approximated dynamics on the manifold.

CHAPTER 7

Local Bifurcations of Fixed Points

7.1. Local Bifurcations of Vector Fields

[**GH86, Wig90, SNM96**] We will now proceed with the bifurcation analysis starting from the simplest possible situation of a one-parameter family of dynamical systems. Some of the ideas presented here have been developed from [**SNM96**].

Having only one real parameter, the simplest (and typical) situation is to have a fixed point where all eigenvalues of its linearisation are different. Assume the fixed point is hyperbolic to begin with. Then the eigenvalues of the linearisation $D_x f(x, \lambda)$ are non–zero and by the Implicit Function Theorem there exists a curve $x(\lambda)$ of hyperbolic fixed points. Moreover, the eigenvalues vary independently of each other, and hence we may assume that only one real part of an eigenvalue is approaching zero at some value λ^*. Other situations (double eigenvalues, two "unrelated" eigenvalues crossing zero simultaneously, etc.) require additional constraints and break down in front of arbitrary small perturbations. We will start the analysis by assuming also that at a non–hyperbolic condition the dependency on λ is *transverse*, technically that $\det D_\lambda f(x_0, \lambda_0) \neq 0$. This situation is called a **codimension–one bifurcation**, i.e., there is a one–parameter family of fixed points *crossing* a non—hyperbolic condition transversally.

Remark 7.1. Codimension–one bifurcations are the typical phenomenon found in applications, since it is difficult to minutely control several parameters at the same time. However there are plenty of experimental studies where codimension–two (and even three) problems are encountered.

Since only one eigenvalue and consequently only one eigenvector or invariant manifold is failing hyperbolicity, after simplifying the problem as much as possible through the Theorems of Hartman–Grobman, Ladis and Sositaisvili, the simplest situation we encounter is where the linearisation block A is real and one–dimensional. The next simplest is where the eigenvalues of A are complex conjugate and cross transversally the imaginary axis. Let us study these two cases in detail. Having only one parameter to resort to, other situations than these two are exceptional and belong in higher–codimension studies.

7.1.1. The Saddle–node Family of Bifurcations. [GH86, SNM96]

The situation we encounter is that after a suitable coordinate transformation, on $n-1$ coordinates the dynamics is essentially linear, hyperbolic and comparatively simple, while there is a one–dimensional decoupled problem $\dot{x} = f(x, \lambda)$ (or, rather, two–dimensional adding the auxiliary equation $\dot{\lambda} = 0$) where changes are unavoidable (topological equivalence along the λ-axis cannot occur). We summarise the non—hyperbolic and transversality conditions (these will be our **basic assumptions**):

- Fixed point: $f(x_0, \lambda_0) = 0$
- Non–hyperbolicity: $\frac{\partial f}{\partial x}(x_0, \lambda_0) = 0$
- Transversality: $\frac{\partial f}{\partial \lambda}(x_0, \lambda_0) \neq 0$

The second condition signals the break down of the one–parameter curve of fixed points used in the fully hyperbolic case, since a basic condition of the Implicit Function Theorem is not fulfilled. However, the third condition allows the use of this Theorem in the "other" direction.

LEMMA 7.1. *Under the basic assumptions, the one–dimensional system $\dot{x} = f(x, \lambda)$ has a unique curve of fixed points $\lambda(x)$ for a sufficiently small interval around x_0, with $\lambda(x_0) = \lambda_0$. Moreover, $\frac{d\lambda}{dx}(x_0) = 0$.*

PROOF. The existence of the curve is assured by the Implicit Function Theorem. Along this curve we have that $f(x, \lambda(x)) = 0$. Taking the x–derivative on both sides, we have $\frac{\partial f}{\partial x} + (\frac{\partial f}{\partial \lambda})\frac{d\lambda}{dx} = 0$. Hence, using the basic assumptions again, we have $\frac{d\lambda}{dx}(x_0) = 0$. □

REMARK 7.2. Recall that the Implicit Function Theorem states also that the function $\lambda(x)$ is differentiable, with derivative

$$\frac{d\lambda}{dx}(x_0) = -\frac{\frac{\partial f}{\partial x}(x_0, \lambda_0)}{\frac{\partial f}{\partial \lambda}(x_0, \lambda_0)}.$$

The numerator on the right hand side is zero because of the basic assumptions.

Computing the second derivative of the fixed point equation at the bifurcation condition we arrive at:

THEOREM 7.1 (Saddle-node Bifurcation). [**GH86**, 3.4.1 p.148] *Under the basic assumptions and provided $\frac{\partial^2 f}{\partial x^2}(x_0, \lambda_0) \neq 0$, there exists a quadratic curve of fixed points $\lambda(x)$ for a sufficiently small interval around x_0. The fixed points are hyperbolic for $x \neq x_0$ and come in pairs with opposite stability for each fixed λ.*

PROOF. By Lemma 7.1 we know that the function $\lambda(x)$ exists and that its first derivative is zero at the bifurcation point. The second derivative of the fixed point equation reads

$$\frac{\partial^2 f}{\partial x^2} + 2\frac{\partial^2 f}{\partial \lambda\, \partial x}\frac{d\lambda}{dx} + \frac{\partial^2 f}{\partial \lambda^2}(\frac{d\lambda}{dx})^2 + \frac{\partial f}{\partial \lambda}\frac{d^2\lambda}{dx^2} = 0.$$

Hence, at the bifurcation condition we have

$$\frac{\partial^2 f}{\partial x^2}(x_0,\lambda_0) = -\frac{\partial f}{\partial \lambda}(x_0,\lambda_0)\,\frac{d^2\lambda}{dx^2}(x_0) \neq 0.$$

Therefore, the function $\lambda(x)$ reads:

$$\lambda(x) = \lambda_0 + c(x-x_0)^2 + O[3],$$

where

$$2c = -\frac{\frac{\partial^2 f}{\partial x^2}(x_0,\lambda_0)}{\frac{\partial f}{\partial \lambda}(x_0,\lambda_0)} \neq 0.$$

We use the notation $O[n]$ to indicate that a given expression is written accurately up to order $n-1$ in $(x-x_0)$, $(\lambda-\lambda_0)$, i.e., in all variables of the extended problem, while terms of power n and higher are concealed in the $O[\cdots]$. The goal of such a partition is that for sufficiently small deviations, the $O[\cdots]$ terms do not influence the qualitative behaviour of the system. In order to compute the stability of the fixed point(s) along the curve, from the x–derivative of the fixed point equation, we obtain:

$$\frac{\partial f}{\partial x} = -\frac{\partial f}{\partial \lambda}\,\frac{d\lambda}{dx}.$$

The first factor on the right hand side is non–zero at the bifurcation point because of the basic assumptions and by continuity it remains non–zero on a sufficiently small x–interval. The second factor reads $\frac{d\lambda}{dx} = 2c(x-x_0) + O[2]$ and changes sign at $x = x_0$. Hence, $\frac{\partial f}{\partial x}$ has different signs at each side of x_0. □

Another approach to arrive at the previous Theorem is the following. A perhaps simpler way to reproduce the saddle–node analysis is the following.

Lemma 7.2. *Let $f(x,\lambda)$ be expansible in Taylor series around (λ_0, x_0) as*

$$\begin{aligned} f(x,\lambda) &= a(\lambda-\lambda_0) + g(x-x_0) + b(\lambda-\lambda_0)(x-x_0) + c(x-x_0)^2 \\ &\quad + d(\lambda-\lambda_0)^2 + O[3]. \end{aligned}$$

Under the **basic assumptions***, and if $\frac{\partial^2 f}{\partial x^2}(x_0,\lambda_0) \neq 0$, then a suitable change of coordinates reduces the Saddle-node problem to the form*

$$f(y,\mu) = \mu + cy^2 + O[3].$$

where now the critical point lies at $(\mu_0, y_0) = (0,0)$.

Proof. Note that $a \neq 0$, $c \neq 0$ and $g = 0$ because of the assumptions of the Lemma. Let $y = (x-x_0) + \frac{b}{2c}(\lambda-\lambda_0)$ and rewrite:

$$f(y,\lambda) = a(\lambda-\lambda_0) + (d - \frac{b^2}{4c})(\lambda-\lambda_0)^2 + cy^2 + O[3]$$

The expression $\mu(\lambda) = a(\lambda-\lambda_0) + (d - \frac{b^2}{4c})(\lambda-\lambda_0)^2$ is monotone and crosses zero for $(\lambda-\lambda_0)$ sufficiently small. By shifting the origin,

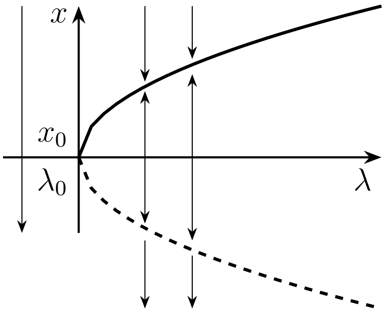


Figure 7.1. A bifurcation diagram for the saddle–node bifurcation. Full line corresponds to a stable fixed point. Vertical lines show typical trajectories for different values of λ.

stretching and rescaling the λ coordinate, we may recast it as μ. Hence, the function $f(y,\mu) = \mu + cy^2 + O[3]$ satisfies the assumptions of the Theorem at the critical point $(\mu_0, y_0) = (0,0)$. □

We will use this technique in the coming subsections.

7.1.1.1. *The Bifurcation Diagram.* [**GH86, SNM96**] A **bifurcation diagram** consists of a graph of the fixed point curve(s) along with the stability of the fixed points for each value of λ, depicted in (x,λ) space, at the vicinity of (x_0,λ_0). It is customary to place the phase space coordinate in the vertical axis and the parameter in the horizontal axis. For the case of a saddle–node bifurcation there are four possible diagrams, depending on the signs of $\frac{\partial f}{\partial \lambda}(x_0,\lambda_0)$ and $\frac{\partial^2 f}{\partial x^2}(x_0,\lambda_0)$. See Figure 7.1 for the case $\frac{\partial^2 f}{\partial x^2} < 0 < \frac{\partial f}{\partial \lambda}$ where the stable fixed point appears for $x > x_0$. Other combinations of signs produce mirror images of this picture along one or both axes.

The previous scenario may be altered when additional constraints are imposed on the system. We will address the two simplest cases since both yield novel results. Imposing further constraints does not alter the qualitative picture around the bifurcation.

7.1.1.2. *The Transcritical Bifurcation.* [**GH86, SNM96**]

Theorem 7.2. [**GH86**, p.149] *Let $f(x,\lambda)$ be expansible in Taylor series around (λ_0, x_0). Assume that the following saddle-node conditions hold, namely $f(x_0,\lambda_0) = 0$, $\frac{\partial f}{\partial x}(x_0,\lambda_0) = 0$ and $\frac{\partial^2 f}{\partial x^2}(x_0,\lambda_0) \neq 0$, along with the additional assumptions $\frac{\partial f}{\partial \lambda}(x_0,\lambda_0) = 0$, $\frac{\partial^2 f}{\partial \lambda\,\partial x}(x_0,\lambda_0) \neq 0$. Then f can be rewritten as*

$$f(y,\mu) = b\mu y + cy^2 + O[3],$$

after a suitable coordinate transformation, where now the critical point occurs at $(\mu_0, y_0) = (0,0)$. The bifurcation diagram displays two hyperbolic fixed points of opposite stability, collapsing into one at the non-hyperbolic condition $\mu = 0$.

PROOF. We leave to the reader to work out the coordinate transformation and parameter renaming-rescaling. We compute the bifurcation diagram noting that the system has a fixed point at $y_0 = 0$, for any μ-value, and another fixed point at $y_1 = -b\mu/c$. It is also straightforward to compute the linearisation eigenvalues $\pm b\mu$. □

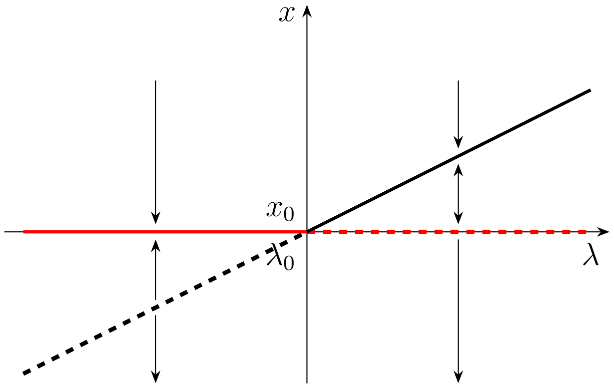


FIGURE 7.2. A bifurcation diagram for the Transcritical bifurcation. Full line corresponds to a stable fixed point. Vertical lines show typical trajectories for different values of λ.

The bifurcation diagram is shown in Figure 7.2 for the case $b > 0$, showing explicitly that the origin of coordinates lies at (x_0, λ_0) (corresponding diagrams occur for different choices of signs for b and c). This bifurcation is called the **Transcritical Bifurcation**. It has two fixed points that "collapse" at the bifurcation point, exchanging stability properties.

7.1.1.3. *The Pitchfork Bifurcation.* [**GH86, SNM96**]

THEOREM 7.3. [**GH86**, p.149] *Under the assumptions of the Transcritical Theorem, with the modified constraints* $\frac{\partial^2 f}{\partial x^2}(x_0, \lambda_0) = 0$ *and* $\frac{\partial^3 f}{\partial x^3}(x_0, \lambda_0) \neq 0$*, a suitable coordinate transformation recasts the problem as*

$$f(y, \mu) = b\mu y + dy^3 + O[4].$$

The bifurcation diagram displays one fixed point at one side of the bifurcation, changing its stability at the bifurcation point. A pair of hyperbolic fixed point of opposite stability (compared with that of the "original" point) branch out of the non-hyperbolic point.

PROOF. We leave again the computation of the new coordinates to the reader. The process gets slightly more cumbersome, since now we have to deal with cubic terms as well. **Hint:** Try with a transformation of the kind $x = y + m\mu + p\mu^2 + qy^2$ and $\lambda = \mu + n\mu^2$ and show for small (y, μ) the arbitrary parameters m, n, p and q can be adjusted to achieve the goal.

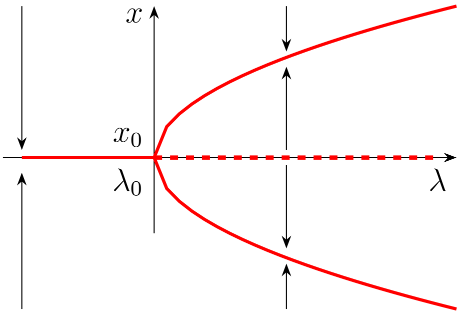


Figure 7.3. A bifurcation diagram for the Pitchfork bifurcation. Full lines correspond to stable fixed points. Vertical lines show typical trajectories for different values of λ.

Analysing the resulting equation we note that there is a fixed point at $y_0 = 0$, occurring for all values of the parameter μ, while a pair of fixed points appear at $y_{1,2} = \pm\sqrt{-b\mu/d}$, provided $-b\mu/d > 0$. We evaluate the y–derivative $g' = b\mu + 3dy^2$ at the fixed points to compute the stability, yielding $b\mu$ for y_0 and $-2b\mu$ for the pair $y_{1,2}$. The bifurcation diagram for $b > 0 > d$ is shown in Figure 7.3. □

The name **Pitchfork** arises from the shape of the fixed point curves. Again, there are four corresponding diagrams with changed orientation and/or stability depending on the signs of b and d.

Remark 7.3. Since both Transcritical and Pitchfork require one, respectively two additional constraints on top of the saddle–node conditions of Theorem 7.1, they are less likely to appear in experimental setups, except for the case of highly symmetric systems. A symmetric system presenting e.g., a Pitchfork, upon an arbitrarily small general perturbation breaking the symmetry may fall into a situation with a saddle–node bifurcation and an unaffected hyperbolic fixed point. See Figure 7.4.

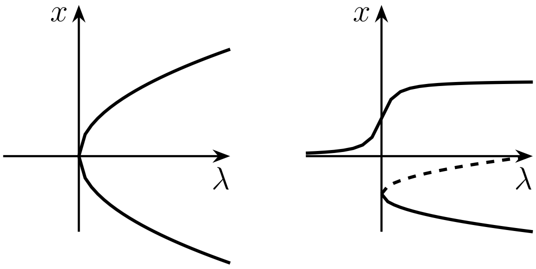


Figure 7.4. A non–symmetric perturbation may modify the Pitchfork into a saddle–node.

7.1.2. The Hopf Bifurcation. [**GH86, SNM96**] The next level of complication arises when we still have one parameter but now a single pair of complex conjugate linearisation eigenvalues become non–hyperbolic by crossing the imaginary axis. Sositaisvili's reduction becomes now two dimensional. The basic assumptions read (shifting the fixed point and the critical parameter value to the origin as it is now customary):

1. Fixed point: $f_1(0,0,\lambda)=0$, $f_2(0,0,\lambda)=0$.
2. Complex eigenvalues:

$$D\left(\begin{array}{c} f_1 \\ f_2 \end{array}\right)(0,0,\lambda)=\left(\begin{array}{cc} \alpha(\lambda) & -\beta(\lambda) \\ \beta(\lambda) & \alpha(\lambda) \end{array}\right)=A(\lambda)$$

where $\alpha(\lambda)=d\lambda$, $d\neq 0$ and $\beta(\lambda)=\omega+c\lambda$, $\omega\neq 0$.

A formal MacLaurin expansion of a general problem compatible with the basic assumptions, up to second order reads

$$\dot{x}=\alpha(\lambda)x-\beta(\lambda)y+a_1x^2+b_1xy+c_1y^2+O(3)$$
$$\dot{y}=\beta(\lambda)x+\alpha(\lambda)y+a_2x^2+b_2xy+c_2y^2+O(3)$$

Let us now work out a polynomial change of coordinates, simplifying the problem, a similar idea to what was sketched previously with the saddle–node family of bifurcations. We note on passing that the procedure repeated in this Chapter of simplifying the vector field as much as possible through coordinate transformations is part of a general method known as **Normal Form Theory**[**GH86, SNM96**].

We take a constructive approach by proposing a change of coordinates $x=u+h_1(u,v)$, $y=v+h_2(u,v)$ where the h's are sufficiently nice quadratic functions to be determined under the condition that the second order terms in the MacLaurin expansion above become as simple as possible (zero if possible). It is easier to work out formally the transformation in matrix form:

$$\left(\begin{array}{c} \dot{x} \\ \dot{y} \end{array}\right)=\left(\begin{array}{cc} 1+\frac{\partial h_1}{\partial u} & \frac{\partial h_1}{\partial v} \\ \frac{\partial h_1}{\partial u} & 1+\frac{\partial h_2}{\partial v} \end{array}\right)\left(\begin{array}{c} \dot{u} \\ \dot{v} \end{array}\right)=(1+Dh)\left(\begin{array}{c} \dot{u} \\ \dot{v} \end{array}\right)$$

Hence,

$$\begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix}=(1+Dh)^{-1}A(\lambda)\begin{pmatrix} u+h_1 \\ v+h_2 \end{pmatrix}+\begin{pmatrix} a_1u^2+b_1uv+c_1v^2+O(3) \\ a_2u^2+b_2uv+c_2v^2+O(3) \end{pmatrix}$$

We replace $(1+Dh)^{-1}$ by $1-Dh$ since this introduces errors only of $O(3)$ or higher. Rewriting this expression in order to make the modification explicit, we obtain

$$\begin{aligned}\begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix}=&A(\lambda)\begin{pmatrix} u \\ v \end{pmatrix}+ \\ &A(\lambda)\begin{pmatrix} h_1 \\ h_2 \end{pmatrix}-Dh\;A(\lambda)\begin{pmatrix} u \\ v \end{pmatrix}+\begin{pmatrix} a_1u^2+b_1uv+c_1v^2+O(3) \\ a_2u^2+b_2uv+c_2v^2+O(3) \end{pmatrix}\end{aligned}$$

If we demand the second order terms in this equation to be identically zero, we obtain a linear system of six equations in the six coefficients entering in h_1 and h_2. The right hand side of the system is given by the six corresponding problem-dependent coefficients a_i, b_i, c_i, $i = 1, 2$.

The good news is that this system has unique solution for a sufficiently small λ–interval around zero. This is because the determinant of the coefficient matrix reads $(\alpha^2 + \beta^2)^2(\alpha^2 + 9\beta^2)$ and is nonzero for $|\lambda|$ sufficiently small since $\beta(\lambda = 0) = \omega \neq 0$.

The final general form, up to third order reads,

$$\begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = A(\lambda) \begin{pmatrix} u \\ v \end{pmatrix} + (u^2 + v^2) \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} + o(3).$$

In the same spirit, we may extend the change of coordinates to $y = v + h_2(u, v) + h_3(u, v)$ where h_3 is an homogeneous cubic polynomial, to be determined under the condition that the *third* order terms in the MacLaurin expansion above become as simple as possible. The calculation is much lengthier. The general idea is exposed in [**GH86**]. The bottom line is that third-order terms cannot be completely eliminated, since the resulting equation system for the coefficients of h_3 does not have unique solution. Hence, some cubic terms will always be present in the final general form. Under the general condition $a \neq 0$ and sufficiently close to the bifurcation point $(0, 0, 0)$ in (x, y, λ)–space, we have

THEOREM 7.4 (Andronov–Hopf Bifurcation Theorem). [**GH86**, 3.4.2 p.151] *Let $f(x, y, \lambda)$ be a family of vector fields with singular solution $(x, y)(t) \equiv (0, 0)$ for all λ. Let $\alpha(\lambda) + i\beta(\lambda)$ be a pair of complex conjugate eigenvalues of the matrix $Df(0, 0, \lambda)$ that crosses the imaginary axis at $\lambda = 0$ with non–zero speed, i.e., $\alpha(0) = 0$, $\beta(0) \neq 0$ and $\alpha'(0) \neq 0$. Further assume that no other eigenvalue is an integer multiple of $i\beta$. Then a family (indexed by λ) of periodic trajectories bifurcates from $(x, y, \lambda) = (0, 0, 0)$. The periods of these orbits approach $\frac{2\pi}{\beta}$ as $\lambda \to 0$.*

PROOF. The proof can be constructed by working out the bifurcation diagram of the general form, sufficiently close to the bifurcation point, including the third order terms. It is more illustrative to recast the problem in polar coordinates, $x = r \cos \phi$, $y = r \sin \phi$, disregarding higher–order terms:

$$\begin{aligned} \dot{r} &= d\lambda r + ar^3 \\ \dot{\phi} &= \omega + br^2, \end{aligned}$$

where d, a and ω are nonzero. Note that the term in b may be disregarded as well for r small enough, since $\dot{\phi}$ will be nonzero for any choice of b if r is small enough.

The structure of a Pitchfork bifurcation can be read on the r–equation. This means that apart from a fixed point at $r = 0$, a solution with constant $r = \sqrt{-d\lambda/a}$ bifurcates whenever the argument of the root is positive (recall that r is non-negative, so there are no other solutions to $\dot{r} = 0$). This solution is a periodic orbit, since r is constant and positive while ϕ varies monotonically. The sign of ω determines the sense of circulation of the orbit, while the relative signs of d and a specifies which of four possible bifurcation diagrams occur. See Figure 7.5 for the case $d > 0 > a$, $w > 0$. □

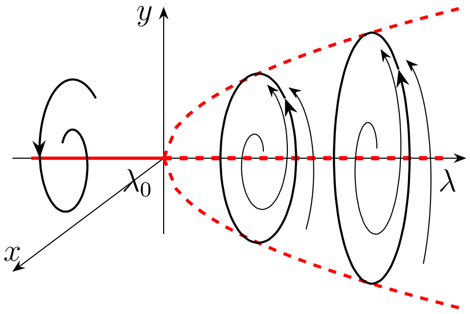


FIGURE 7.5. A bifurcation diagram for the Hopf bifurcation. Full line corresponds to a stable fixed point. When the fixed point changes stability, a stable periodic orbit spawns out.

REMARK 7.4. When the bifurcation generates a stable limit cycle, it is called **supercritical**, otherwise **subcritical**.

7.2. Local Bifurcations of Fixed Points of Maps

[**GH86, Wig90, SNM96**] Intuitively, since all flows generate time–one maps, one may expect that the effects of local bifurcations on flows will transport to maps one way or the other. Bifurcations of fixed points of maps must exhibit at least as much complication as that shown by a family of time–one maps when the underlying flow–family undergoes a bifurcation.

Since there is a map version of the Centre Manifold Theorems, we will start the analysis following the spirit of the previous sections, i.e., by considering a general map under non–hyperbolic conditions, simplifying its expression as much as possible via changes of coordinates and finally studying the resulting family of maps. Again, the simplest bifurcations occur when the hyperbolicity condition of Hartman–Grobman theorem breaks down as one eigenvalue of the linearisation of the map at the fixed point crosses 1 or -1, or when a pair of complex conjugate eigenvalues crosses the unit circle. Many features of the treatment resemble those of flows. We give here a summarised description, leaving the details to the reader, when they just mimic the corresponding computations for flows.

7.2.1. The Saddle–node Family of Bifurcations of Maps. [**GH86, SNM96**] The setup here is a map $x_{n+1} = F(x_n, \lambda)$ with the following **basic assumptions** (using coordinates such that the fixed point and the critical parameter value occur at the origin):

1. Fixed point at the origin: $F(0, 0) = 0$
2. Non–hyperbolicity: $\frac{\partial F}{\partial x}(0, 0) = 1$
3. Transversal crossing: $\frac{\partial F}{\partial \lambda}(0, 0) \neq 0$

Theorem 7.5. **(Saddle-node for maps)**[**GH86**, p.157] *Under the basic assumptions, and given $\frac{\partial^2 F}{\partial x^2}(0, 0) \neq 0$, a general map $F(x, \lambda)$ that admits a Taylor expansion in x and λ can be taken to the form*

$$x_{n+1} = \lambda + x_n + c{x_n}^2 + O([3])$$

by changes of coordinates. The associated dynamics is that of a pair of hyperbolic fixed points of opposite stability branching out of the non-hyperbolic situation.

Proof. The central part of the proof is the coordinate change, which can be performed paralleling the procedures of the previous section. □

As in the case of flows, the non–hyperbolicity condition leads in this case also to the breakdown of the Implicit Function Theorem conditions to find a unique curve $x(\lambda)$ of fixed points, while the transversal crossing condition permits to use the Theorem the other way around, to establish the existence of a curve $\lambda(x)$.

Theorem 7.6. **(Transcritical and Pitchfork for maps)**[**SNM96**, p.120] *Under the assumptions of the previous Theorem, modified as:*
(a) $\frac{\partial F}{\partial \lambda}(0, 0) = 0$ while $\frac{\partial^2 F}{\partial \lambda\, \partial x}(0, 0) \neq 0$ or
(b) $\frac{\partial F}{\partial \lambda}(0, 0) = 0$, $\frac{\partial^2 F}{\partial x^2}(0, 0) = 0$ while $\frac{\partial^2 F}{\partial \lambda\, \partial x}(0, 0) \neq 0$ and $\frac{\partial^3 F}{\partial x^3}(0, 0) \neq 0$
then a general map $F(x, \lambda)$ that admits a Taylor expansion in x and λ can be taken to the form (respectively)

$$\begin{aligned} (a):& \quad x_{n+1} = (1 + \lambda)x_n + c{x_n}^2 + O([3]) \\ (b):& \quad x_{n+1} = (1 + \lambda)x_n + d{x_n}^3 + O([4]) \end{aligned}$$

by changes of coordinates.

The bifurcation diagrams mimic completely those already displayed for flows.

7.2.2. The Flip or Period–Doubling Bifurcation. [**GH86, SNM96**] Probably the most famous difference between flows and maps is given by the Period–doubling, or Flip, bifurcation. Indeed, in maps there are two ways of obtaining a non–hyperbolic dynamical system where only one real eigenvalue is involved. Apart from the saddle–node type of non–hyperbolicity (in its three flavours) discussed above (paralleling in most respects the behaviour in flows), we have to consider the case of an eigenvalue $\lambda = -1$.

Assume then that $F(x, \lambda)$ has a non–hyperbolic fixed point at $(0, 0)$ with eigenvalue -1. By the Implicit Function Theorem, F will have a curve $x(\lambda)$ of fixed points and no other fixed points sufficiently close to this curve. Note that in this case the argument goes in a completely different way as compared with the saddle–node situation, since there is no "breakdown" of the conditions of the Implicit Function Theorem associated to the non–hyperbolicity condition.

Let us consider, however $G(x, \lambda) = F(F(x, \lambda), \lambda)$, the second iterated map. G has also a non–hyperbolic fixed point at $(0, 0)$ but now with eigenvalue $+1$. Suppose now that the conditions for a Pitchfork bifurcation are fulfilled by G. The fixed point of F is also a fixed point of G (the whole curve indeed), but the additional (unique) pair of fixed points (x_1, x_2) of G branching away from the bifurcation are *not* fixed points of F (F has a unique curve of fixed points $x(\lambda)$). Hence, we must have $F(x_1) = x_2$ and $F(x_2) = x_1$ (otherwise G would have more than three fixed points), i.e., a period–2 orbit. More formally:

THEOREM 7.7 (Period–Doubling Bifurcation of Maps). [**GH86**, 3.5.1 p.158] *Let $F(x, \lambda)$ be a C^3 map in both variables and let $G(x, \lambda) = F(F(x, \lambda), \lambda)$. Under the assumptions*

1. $F(0, 0) = 0$
2. $\frac{\partial F}{\partial x}(0, 0) = -1$
3. $\frac{\partial G}{\partial \lambda}(0, 0) = 0$
4. $\frac{\partial^2 G}{\partial x^2}(0, 0) = 0$
5. $\frac{\partial^2 G}{\partial x\, \partial \lambda}(0, 0) \neq 0$
6. $\frac{\partial^3 G}{\partial x^3}(0, 0) \neq 0$

there exists a curve of period–2 points of F for a sufficiently small interval around $(x, \lambda) = (0, 0)$. At both sides of the bifurcation in parameter space, the fixed point is hyperbolic and switches stability. The period–2 points have opposite stability compared to the fixed point.

PROOF. Most of the proof is already done. Only the final issue on stability is left. A general C^3 map satisfying the first two conditions reads

$$F(x, \lambda) = \lambda(\alpha + \beta\lambda + \gamma\lambda^2) + x(-1 + a\lambda + b\lambda^2) + x^2(c + g\lambda) + dx^3 + O[4].$$

Moreover, for $|\lambda|$ sufficiently small we may further simplify the expression to $F(x,\lambda) = \lambda + x(-1 + a\lambda) + cx^2 + dx^3 + O[4]$ by subsequent rescalings. Computing the iterated map G it is straightforward to verify that conditions 3 and 4 are fulfilled. Conditions 5 and 6 defining the Pitchfork scenario for G can be recast in terms of the derivatives of F at $(0,0)$:

5'. $2\frac{\partial^2 F}{\partial x\,\partial\lambda} + \frac{\partial F}{\partial\lambda}\frac{\partial^2 F}{\partial x^2} \neq 0$

6'. $\frac{1}{2}(\frac{\partial^2 F}{\partial x^2})^2 + \frac{1}{3}\frac{\partial^3 F}{\partial x^3} \neq 0$

By direct computation we can now verify that condition 5' guarantees that the common fixed point of F and G changes stability at the bifurcation (regarded as a fixed point of either F or G). □

7.2.3. The Hopf Bifurcation on Maps. [**GH86, SNM96**] Next in the series of non–hyperbolic situations is when a pair of complex conjugate eigenvalues crosses the unit circle in the complex plane. The corresponding bifurcation for maps has been popularised under the name "Hopf", despite the fact that Hopf Theorem deals with flows. The map theorem was proved by Neimark in 1959[**Nei59, Sac65**]. We deal here with a 2–dimensional problem with a fixed point at the origin. The simplest formulation of this bifurcation theorem reads,

Theorem 7.8. [**GH86**, 3.5.2 p.162] *Let $F(x,y,\lambda)$ be a map on $\mathbb{R}^2$ depending on one parameter with a smooth family of fixed points $(x(\lambda), y(\lambda))$ with a pair of complex conjugate eigenvalues $\sigma(\lambda)$. Without loss of generality let $x(0) = 0 = y(0)$. Assume*

H1. $|\sigma(0)| = 1$ *and* $\sigma(0)^j \neq 1$, *for* $j = 1,2,3,4$

H2. $\frac{\partial|\sigma|}{\partial\lambda}(0) = d \neq 0$

Then there is a smooth coordinate change mapping F to (in polar coordinates):

$$\bar{F}(r,\theta,\lambda) = (r(1 + d\lambda + ar^2), \theta + c + g\lambda + br^2) + \textit{higher order terms.}$$

If in addition

H3. $a \neq 0$

then an invariant circle for the map $(r_{n+1}, \theta_{n+1}) = \bar{F}(r_n, \theta_n\lambda)$ branches out from the bifurcation point.

Proof. We will only sketch parts of the proof here. Assumption H1 with $j = 1,2$ assures that the eigenvalue $\sigma(0)$ is not real, i.e., $c \neq 0$. Working out the coordinate transformation, one realises that H1 with $j = 3,4$ assures that all lower-order terms in r, θ, λ can be eliminated when going from F to $\bar{F}$, i.e., that $\bar{F}$ as above can be achieved. H2. assures also the change of stability of the fixed point at the bifurcation.

Consider next $\bar{F}$ without higher-order terms. H3 assures that $r = \sqrt{-d\lambda/a}$ is an invariant circle for the dynamics. Since $c \neq 0$ the dynamics on the invariant circle is that of a shift map.

The final and technical part of the proof is to realise that the truncated dynamics here sketched is equivalent to the dynamics of the full map $\bar{F}$ for sufficiently small λ and r. □

REMARK 7.5. Depending on the relative signs of d and a the change of stability of the fixed point and the occurrence of the invariant circle may be either supercritical or subcritical as in the flow case (four possible cases).

REMARK 7.6. The dynamics on the invariant circle is governed to lowest order by $\theta \mapsto \theta + c + (g - bd/a)\lambda$. Maps of this sort are not structurally stable. Both periodic and non–periodic behaviour will alternate when varying λ. The precise description of this map requires full knowledge of $\bar{F}$.

REMARK 7.7. If H1 with $j = 3, 4$ is not fulfilled, $\bar{F}$ will have additional quadratic and cubic terms in r, consequently altering the nature of the dynamics.

If H1 with $j = 1, 2$ is not fulfilled, we are actually dealing with a pair of eigenvalues $+1$ or -1, a problem which actually belongs in codimension–2 analysis.

7.3. Some General Comments on Local Bifurcations

7.3.1. Higher Dimensions. The presentation in this Chapter relies in the possibility of performing the reduction to the Centre Manifold described by Sositaisvili Theorem. This reduction is in general possible, but there exist some exceptional cases, called *resonances* where lower-order terms cannot be eliminated (as mentioned in the Hopf Theorem for maps) or a higher degree of smoothness cannot be achieved. Also, recall that the dynamics on the centre manifold does influence the details of the local behaviour on the hyperbolic subspace.

7.3.2. Higher-codimensional Bifurcations. [**GH86**] Codimension–one bifurcations are the starting cases of a "chain" of bifurcation problems subject to the influence of several independent and coexisting parameters. Regarding the situation from the point of view of applications, real–life problems are for practical reasons described by taking into account a small set of independent parameters at a time, so in any case, only low–codimension bifurcation problems are relevant.

Consider e.g., a laser device with injected signal. Such devices are useful in a variety of experimental situations, particularly in communication. The injected electromagnetic signal will have its own amplitude and phase, independent of the laser device, thus providing two parameters that can be controlled by the experimenter to some extent. In addition, the laser cavity itself may vary (by having a moving mirror on one end or some other sophisticated construction) and hence a mismatch between the natural frequency of the cavity and that of the lasing medium can be achieved. We have therefore three parameters that can be varied somehow freely, and the overall dynamics will occur in a subset of a three-dimensional family of dynamical systems. There exists experimental and theoretical work describing codimension–2 bifurcations as well as the effect of the "surrounding" codimension–3 environment.

7.4. Exercises

Exercise 7.1. Verify that the system

$$\dot{r} = \lambda r + 2r^3 - r^5$$
$$\dot{\theta} = 1$$

has a subcritical Hopf bifurcation.

Exercise 7.2. Study the occurrence of fixed point bifurcations at the origin for different values of a for the system

$$\begin{aligned} x_{n+1} &= ax_n + (6-a)y_n + {x_n}^3 \\ y_{n+1} &= -x_n - 2y_n + {y_n}^3. \end{aligned}$$

Exercise 7.3. Study the occurrence of fixed point bifurcations at the origin for different values of a for the system

$$\begin{aligned} \dot{x} &= (1+a^2)x + (2-6a)y + x^3 \\ \dot{y} &= -x - 2y - y^4. \end{aligned}$$

CHAPTER 8

Chaotic Systems

8.1. One-dimensional Chaotic Maps

We have seen in Chapter 4 that one-dimensional invertible maps have a rather simple dynamics, at least one that is understandable referring only to periodic and quasi-periodic trajectories. They have a rather predictable limiting behaviour. We may say that they do not produce *chaotic regimes.* We defer a proper definition until later, but we advance that one feature of chaotic dynamics is that the ω–limit set is more complicated than just a finite set of periodic trajectories and connecting trajectories.

We saw also with Poincaré–Bendixson Theorem 3.9 that 2–dimensional flows are also rather predictable and understandable with everyday tools.

To encounter "chaos" it will be necessary to search among problems with higher complexity. Fortunately there are plenty of such problems in nature and chaos does leave traces in experiments and natural phenomena. What we precisely mean by "chaos" will be specified after some acquaintance with its properties.

Let us start our search with one-dimensional non–invertible maps. Although invertibility is an important dynamical feature, this is not as serious as it may look. Many invertible dynamical systems display an "underlying" non–invertible map appearing after some sort of limiting procedure (e.g., "infinite dissipation", etc.).

We will start with some model classes.

8.1.1. Expanding Maps of the Circle. [**KH96**]

Definition 8.1. A differentiable map $f\colon \mathbb{S}^1 \to \mathbb{S}^1$ is called **expanding** if $|f'(x)| > 1$ for all $x \in \mathbb{S}^1$.

The simplest examples are linear maps, of the form

$$E_m\colon \mathbb{S}^1 \to \mathbb{S}^1 \quad E_m(x) = mx \mod 1,$$

for $m > 1$ a positive integer.

Theorem 8.1. [**KH96**, p.39, $m = 2$]*For a linear expanding map E_m of the circle,*

1. $\#\,\mathrm{Per}_n(E_m) = m^n - 1$, *where*
   $\mathrm{Per}_n(E_m) = \mathrm{Fix}(E_m^n) = \{x \in \mathbb{S}^1 \,:\, E_m^n(x) = x\}$

2. *The closure of the set of periodic points is the whole circle, i.e.,* $\overline{\mathrm{Per}(E_m)} = \mathbb{S}^1$.
3. *The rational numbers are* **pre-periodic***, i.e., there are numbers* $k, n \in \mathbb{N}$ *such that* $E_m^{k+n}(x) = E_m^k(x)$ *for rational* x*. Here* n *is the period and* k *is the pre-period.*
4. E_m *is transitive, i.e., there exists a dense orbit.*

Remark 8.1. Notice here a new phenomenon which might be considered as an indicator of *chaos*, namely the coexistence of periodic (i.e., non-dense) trajectories and dense trajectories.

Proof. As customary, we identify the unit circle $\mathbb{S}^1$ with the real unit interval mod 1 through the map $t \to e^{2\pi t}$. A periodic point of period n satisfies $E_m^n(x) = x$. This implies

$$m^n x = x \mod 1 \quad \text{or} \quad (m^n - 1)\, x = 0 \mod 1.$$

In other words, $(m^n - 1)\, x$ is an integer. The equation has exactly $m^n - 1$ different roots on $[0, 1)$ (all of them rational):

$$x = \frac{k}{m^n - 1} \quad k = 0, \cdots, m^n - 2.$$

This proves the first statement.

The second statement says that the set of periodic points is dense in the circle. Recalling the definition, this means that for any point y on the unit interval and any $\epsilon > 0$, there exists a periodic point p such that $|y - p| < \epsilon$. Note that the periodic points of period n are equally spaced on the circle with distance $d = \frac{1}{m^n - 1}$. This implies that when n grows they become more and more dense. For any given ϵ, taking n large enough we obtain $d < \epsilon$.

Regarding the third statement, for $x = p/q \in \mathbb{Q}$ there are only q different images $E_m^j \left(\frac{p}{q}\right) = \frac{m^j p}{q}$ mod 1, when varying the value of j. Hence, repetitions must occur, i.e., for some integers n, k $E_m^{n+k} \left(\frac{p}{q}\right) = E_m^n \left(\frac{p}{q}\right)$, what proves the statement. To depict the situation, recall that the map is expanding and thus not 1-to-1. A periodic point of period n might have more than one pre-image. The additional pre-images are not periodic of period n but will eventually become periodic after a few iterations of the map. In other words, we look for rational numbers $x = p/q$ such that $E_m^{k+n}(x) = E_m^k(x)$, (this is: $E_m^k(x)$ is n-periodic) while $E_m^k(x) \neq x$ (x is different from $E_m^k(x)$) and $E_m^n(x) \neq x$ (x itself is not n-periodic). The statement becomes that for $x = p/q \in \mathbb{Q}$, there exists $k, n \in \mathbb{N}$ such that $(m^n - 1)m^k p/q \in \mathbb{N}$ while $(m^k - 1)p/q \notin \mathbb{N}$ and $(m^n - 1)p/q \notin \mathbb{N}$. Example: Let $p = n = 1$, $m = 3$ and $q = 9$. The points $x_1 = 1/9$ and $x_2 = 1/3$ are *not* of period $n = 1$, since $E_3(1/9) = 1/3$ and $E_3(1/3) = 0$. x_1 is not of period $n = 2$ either, since $E_3^2(1/9) = 0$. However, $E_3^{2+j}(1/9) = E_3^{1+j}(1/3) = 0$ for all $j \geq 1$.

Hence x_1 has pre-period $k = 2$ and x_2 has pre-period $k = 1$. After k iterates of E_3 those points land on the period-1 point $x = 0$. There exists another point of pre-period $k = 1$, namely $y_2 = 2/3$. The points $p/9$, for $p = 1, 4, 7$ are pre-images of x_2 with pre-period $k = 2$, while the points for $p = 2, 5, 8$ are pre-images of y_2 also with pre-period $k = 2$.

Finally, let I be an arbitrary interval on the circle. Then there are $k, n \in \mathbb{N}$ such that

$$\left[\frac{k}{m^n}, \frac{k+1}{m^n}\right] \subset I.$$

But

$$E_m^n(I) \supset E_m^n\left(\left[\frac{k}{m^n}, \frac{k+1}{m^n}\right]\right) = \mathbb{S}^1.$$

Let J be any other interval in the circle. Then

$$E_m^n(I) \cap J = \mathbb{S}^1 \cap J = J \neq \varnothing.$$

This means that E_m is transitive. □

The next theorem shows that the limit behaviour of some trajectories might be even more complicated.

THEOREM 8.2. [**KH96**, 1.7.3 p.40] *For the expanding linear map E_3, there exists a point (actually many) $X \in \mathbb{S}^1$ with $\omega(X) = \mathcal{C}$, the* **Standard Cantor set**.

For the definition and properties of the Cantor set, refer to Appendix A.2.1. We discuss here two properties which will be of use in the proof of this Theorem. The proof uses a **coding method** which will be of great importance later on.

1. The Standard Cantor set $\mathcal{C}$ corresponds to all numbers in $[0, 1]$ whose base–3 expansion does not contain the digit 1, i.e.,

$$x \in \mathcal{C} \iff x = \sum_{i=1}^{\infty} \frac{a_i}{3^i} \quad a_i \in \{0, 2\}.$$

2. The action of E_3 on a point $x \in \mathcal{C}$ is as follows:

$$\begin{aligned} E_3(x) &= \left(3 \sum_{i=1}^{\infty} \frac{a_i}{3^i}\right) \bmod 1 \\ &= \left(a_1 + \sum_{i=1}^{\infty} \frac{a_{i+1}}{3^i}\right) \bmod 1 \quad = \sum_{i=1}^{\infty} \frac{a_{i+1}}{3^i} \in \mathcal{C} \end{aligned}$$

Hence, $E_3(\mathcal{C}) = \mathcal{C}$.

PROOF. Let us write a list of all finite "words" with letters $0, 2$. This can be done in length–lexicographical ordering:

$$0, 2, 00, 02, 20, 22, 000, 002, 020, 022, 200, 202, 220, 222, \cdots$$

i.e., first we order the words in groups by length, so that words of length $k+1$ are found after those of length k, for all $k \geq 1$. Among each length k the order is lexicographic, namely the natural ordering of the strings taken as integer numbers. The reader should check in Appendix A.2.1 that the words of length k label all intervals required in step k of the construction of the Cantor set. Recall also that these intervals have width $1/3^k$. The proof finishes by exhibiting the point X with the stated property. Let $X \in \mathbb{S}^1$ to be the point whose base–3 expansion has the form

$$X = .0200022022000002020022200202220222\cdots$$

built by placing the words of the previous list one after the other. By construction, $X \in \mathcal{C}$. The map E_3^n acting on X "erases" the first n symbols in the string of X, hence the whole trajectory $E_3^k(X)$, $k \geq 0$ lies in the Standard Cantor set. Moreover, we see that for any k and sufficiently large n, the trajectory of X, visits all intervals used in the construction of the Cantor set at step k. This suffices to verify that the orbit of X is dense in the Cantor set. Indeed, for any $\epsilon > 0$ we can find k such that $1/3^k < \epsilon$ and hence the orbit of X passes closer than ϵ from any point in $\mathcal{C}$. Hence, $\omega(X) = \mathcal{C}$. □

Remark 8.2. A general feature of any E_m map is that its action is very simple when $x \in [0,1)$ is written in base m. Indeed,

$$x = \sum_{i=1}^{\infty} \frac{a_i}{m^i} \Rightarrow E_m(x) = \sum_{i=1}^{\infty} \frac{a_{i+1}}{m^i}$$

where the a_i's take values in $0, \dots, m-1$. E_m eliminates the first symbol in the base–m expansion of x. Hence, the limit behaviour of a point x does not depend on the first initial symbols in base–m. All types of limit behaviour can be found in any small interval $(a,b) \subset \mathbb{S}^1$. An interval of the form $(\frac{k}{m^n}, \frac{k+1}{m^n})$ is defined by the first n digits in base m: Writing k in base m as $k = k_1 m^{n-1} + \cdots + k_{n-1} m + k_n$ (where the integers $k_i \in [0, m-1]$) then

$$\left(\frac{k}{m^n}, \frac{k+1}{m^n} \right) = \left\{ x \in \mathbb{S}^1 \,:\, a_1 = k_1, \cdots, a_n = k_n \right\}.$$

Remark 8.3. In general the ω–limit set of a point x can be very complicated. A deep theorem of *Jewitt and Krieger*[**Jew70, Kri72**] states that the maps E_m "contain" all **stationary stochastic processes** as the limit behaviour. This means that it is impossible to describe all trajectories. Therefore we should call these maps **chaotic**.

Now we want to analyse general expanding maps on the circle. Again, for an expanding map $f \colon \mathbb{S}^1 \to \mathbb{S}^1$ we can consider its lift $F \colon \mathbb{R} \to \mathbb{R}$.

THEOREM 8.3. [**KH96**, 2.4.6 p.73] *Any expanding map g of degree m is topologically conjugate to E_m.*

PROOF. Since g is expanding, any lift G has nonzero and continuous derivative, hence, this derivative never changes sign and G is a strictly monotone function. Let $F(x) = mx$ on $\mathbb{R}$. Then F is a lift of E_m. We also fix a lift G of g with the property that G has its (unique) zero a in $[-1/2, 1/2)$. We consider now G restricted to the interval $[a, a+1)$. Then

$$G(a) = 0 \quad G(a+1) = m.$$

We can partition the image by G of the interval $[a, a+1]$ in intervals of length unity, $[j, j+1]$, $0 \leq j < m$ and using the pre-images by G, $[a, a+1]$ can be partitioned as well in intervals $I_0^1, \cdots, I_{m-1}^1$ with

$$I_j^1 := G^{-1}\left([j, j+1]\right) \quad j = 0, \cdots, m-1.$$

Inductively, we define similar intervals for G^n:

$$I_j^n := \left(G^n\right)^{-1}\left([j, j+1]\right) \quad j = 0, \cdots, m^n - 1.$$

The same procedure for F yields equally spaced intervals Δ_j^n. This is illustrated in Figure 8.1. Letting $a'_{n,j} = j/m^n$ we have that $\Delta_j^n = [a'_{n,j}, a'_{n,j+1}]$. Let $a_{n,j}$ be the endpoints of I_j^n.

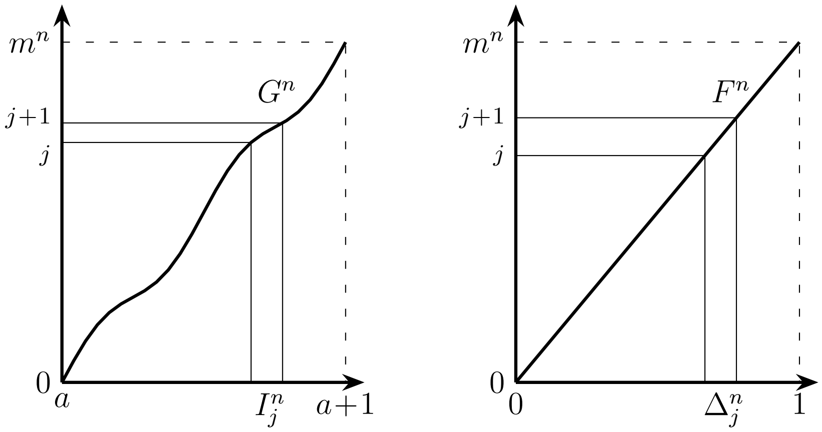


FIGURE 8.1. The pre-image I_j^n of a unit-length interval by G^n and the corresponding interval Δ_j^n for F^n.

Since G is expanding we have $|G'(x)| \geq \mu > 1$ for all $x \in \mathbb{R}$. This implies that $\left|I_j^n\right| \leq \mu^{-n}$. We define the conjugating map H on the endpoints of the intervals I_j^n and Δ_j^n:

$$I_j^n = [a_{n,j}, a_{n,j+1}] \quad \Delta_j^n = [a'_{n,j}, a'_{n,j+1}]$$

$$H(a_{n,j}) = a'_{n,j}.$$

Because of the properties of lifts, the intervals I and Δ can be replicated over the whole real line. We have that

$$\frac{mj}{m^n} = ma'_{n,j} = F(H(a_{n,j})) = a'_{n-1,j} = H(a_{n-1,j}) = H(G(a_{n,j})).$$

Note that the intermediate images $a'_{n-1,j}$ may lie outside $[0,1]$. Since the length of the intervals tends to 0 and H is monotone we can continuously extend H to a map $H\colon [a,a+1) \to [0,1)$ (a map on the whole real line, in fact). This map is monotone and bijective, hence a homeomorphism. Moreover,

$$H(a+1) - H(a) = F(1) - F(0) = \deg f = \deg g \in \mathbb{Z}.$$

H projects to a map $h\colon \mathbb{S}^1 \to \mathbb{S}^1$. Also by continuity,

$$f \circ h = h \circ g.$$

□

8.1.2. Coding and Symbolic Dynamics. [**GH86, Wig90, SNM96**] In this Subsection we will investigate a method to illustrate the dynamics in a combinatorial way. We have been using this method without explicit formalisation.

Is there some simple way to visualise the action of $f = E_m$? Express $x \in [0,1)$ in base $m = \deg g$:

$$x = .s_1(x)s_2(x)s_3(x)\cdots = \sum_{k=1}^{\infty} \frac{s_k(x)}{m^k}$$

The symbols s_k are integers in $\{0,1,\cdots,m-1\}$. We have:

$$F(x) = mx = \sum_{k=1}^{\infty} \frac{s_k(x)}{m^{k-1}} = s_1 + \sum_{k=1}^{\infty} \frac{s_{k+1}(x)}{m^k}$$

and consequently

$$f(x) = mx \mod 1 = \sum_{k=1}^{\infty} \frac{s_{k+1}(x)}{m^k} = .s_2(x)s_3(x)s_4(x)\cdots$$

With this notation, the action of f consists of dropping the first symbol of the expansion of x in base m. For this reason f is called a **shift map**: f shifts x one step along the symbol list. We will now rephrase the result of Theorem 8.3 in the shift-symbol language.

To any $x \in [a,a+1)$, we associate the infinite sequence of symbols

$$x \longrightarrow \underline{x} = s_1(x)s_2(x)s_3(x)\cdots$$

where $s_i(x)$ is the i–th symbol of the expansion of $h(x)$ in base–$m = \deg g$, i.e., using base–m expansion $\underline{x}$ is another way of regarding $h(x)$, only that the initial dot is missing. In this way, we are implicitly defining a space $\Sigma_m = \{0,1,\cdots,m-1\}^\infty$ whose elements are all infinite symbol sequences where each symbol is an integer in $\{0,1,\cdots,m-1\}$.

Then we have an almost 1–to–1 relation between $\mathbb{S}^1$, seen as $[a, a+1)$ and Σ_m:

$$\{\underline{x} \,:\, x \in \mathbb{S}^1\} = \Sigma_m = \{0, 1, \cdots, m-1\}^\infty,$$

The only exception are the m-adic points, i.e. the endpoints of the defining intervals where we have a 2-1 map from Σ_m to $\mathbb{S}^1$. This corresponds to $0.9999\cdots = 1$ in base 10.

There is a natural way to define distances on Σ_m (by recasting back a string as a point in $\mathbb{S}^1$ expressed in base–m and using the natural distance on the real numbers) by which all original continuity properties are also preserved on the symbol space. We may hence consider the map $\underline{h} : \mathbb{S}^1 \to \Sigma_m$ that in all practical respects acts as h. The new coding is useful to express the result of Theorem 8.3 in terms of $\underline{h}$:

$$\underline{g}(\underline{x}) \equiv \underline{h}(g(x)) = f(\underline{h}(x)) \equiv \sigma(\underline{x})$$

where $\sigma\colon \Sigma_m \to \Sigma_m$ is the left shift defined by

$$\sigma(x_1x_2x_3\cdots) = x_2x_3x_4\cdots$$

This provides a simple combinatorial description of the original system, i.e., g is conjugate to a shift map $\underline{g} \equiv \sigma$. The dynamical properties of g are thus transferred to a symbol space and to shifts on strings of symbols.

Using this coding, we can easily identify the ω–limit set of a point $\underline{x}$: A set

$$[i_1 \cdots i_n] := \{\underline{x} \in \Sigma_m \,:\, x_1 = i_1, \cdots, x_n = i_n\}$$

is called a **cylinder** of length n (Sometimes cylinders are called **words** or **blocks**). Now a cylinder has non-empty intersection with the ω–limit set if and only if there is a sequence $n_k \nearrow \infty$ such that

$$\sigma^{n_k}(\underline{x}) \in [i_1 \cdots i_n]$$

(a cylinder is a compact set and therefore it must contain limit point). But

$$\sigma^{n_k}(\underline{x}) = x_{n_k+1}x_{n_k+2}\cdots x_{n_k+i_n}\cdots$$

and we "see" the cylinder $[i_1 \cdots i_n]$ infinitely often in the infinite sequence $x_1x_2\cdots$.

The reader may recognise a cylinder in the labeling of an interval at step n of the construction of the Standard Cantor set.

8.1.3. An application to complex analysis (function theory).

We want to use dynamics to see that not every analytic function on the unit disc has an *analytic continuation*. Recall that for an analytic function f defined on an open set U, an analytic function F with domain $V \supset U$ is called an analytic continuation of f if F and f coincide on U. We consider the *lacunary Taylor series*

$$f(z) = \sum_{n=0}^{\infty} z^{2^n}$$

with radius of convergence $|z| < 1$. We are going to show that this series has no analytic continuation. For this we consider *radial limits* and show that they are infinite on a dense set of the unit circle. Hence there is no "way out" off the circle.

The radial limit (if it exists) is defined as follows. Let $z = re^{i\theta}$. Then

$$f(re^{i\theta}) = \sum_{n=0}^{\infty} r^{2^n} e^{i2^n\theta}.$$

The radial limit is

$$l(\theta) := \lim_{r\nearrow 1} \sum_{n=0}^{\infty} r^{2^n} e^{i2^n\theta}.$$

Let us consider a (rational) multiple of 2π, $\theta = (2\pi)a_0.a_1a_2\cdots a_N00000\cdots$; $a_i \in \{0,1\}$. Then

$$\begin{aligned} l(\theta) &:= \lim_{r\nearrow 1} \sum_{n=0}^{\infty} r^{2^n} e^{i2^n\theta} \\ &= \lim_{r\nearrow 1} \left(\sum_{n=0}^{N} r^{2^n} e^{i2^n\theta} + \sum_{n=N+1}^{\infty} r^{2^n} e^{0} \right) \\ &= \lim_{r\nearrow 1} l_N(\theta) + \lim_{r\nearrow 1} \sum_{n=N+1}^{\infty} r^{2^n} \\ &= a(N,\theta) + \lim_{r\nearrow 1} \sum_{n=N+1}^{\infty} r^{2^n} = \infty. \end{aligned}$$

Where $a(N,\theta)$ is a fixed number. Since those θ's are dense in $[0,1)$ we proved that there is no convergent Taylor series expansion on any sub-interval of the unit circle. Hence we cannot perform an analytic continuation.

There is a general principle for the behaviour of the radial limit: Let $S_n(\theta) = \sum_{j=0}^{n} e^{i2^j\theta}$.

Lemma 8.1. *Let $N = N(r) = \max\{n \geq 0 : r^{2^n} \geq e^{-1}\}$. We have*

$$\left| f(re^{i\theta}) - S_{N(r)}(\theta) \right| < C$$

where the constant C is uniform on r and θ.

Proof. We write

$$f(re^{i\theta}) \;=\; S_{N(r)}(\theta) + \sum_{n=0}^{N} (r^{2^n} - 1)e^{i2^n\theta} + \sum_{n=N+1}^{\infty} r^{2^n} e^{i2^n\theta}.$$

On one hand, since $r^{2^{N+k}} \leq e^{-2^k}$ for $k \geq 1$, we have

$$\left| \sum_{n=N+1}^{\infty} r^{2^n} e^{i2^n\theta} \right| \leq \sum_{n=N+1}^{\infty} r^{2^n} \leq \sum_{k=1}^{\infty} e^{-2^k} < \infty.$$

On the other hand, since $r^{2^N} \geq e^{-1}$, we have $r \geq e^{-\frac{1}{2^N}}$. So, using $e^{-x} \geq 1 - x$ we have

$$\left|\sum_{n=0}^{N}(1 - r^{2^n})e^{i2^n x}\right| \leq \sum_{n=0}^{N}\left(1 - \exp(-2^{n-N})\right) \leq \sum_{n=0}^{N} 2^{-(N-n)} \leq 2.$$

□

Notice that

$$N(r) \sim \log_2 \frac{1}{1-r}.$$

The dynamics of E_2 allows, with the help of the previous Lemma, a more detailed analysis of this lacunary Taylor series. In particular, a dense set can be found with bounded radial limiting behaviour. Moreover, a more sophisticated application of stronger dynamical and statistical arguments leads to Central Limit Theorems, the Law of Iterated Logarithms or more properties of the radial limit behaviour. This way, the limiting behaviour might be considered as chaotic. All these properties hold for general lacunary Taylor series and were proved without dynamics earlier.[**FS03**]

8.1.4. More General Maps – The Logistic Map. [GH86, SNM96]

If we drop the assumption of expansiveness we immediately turn into much more difficult situations. There is no final theory on general differentiable maps of the interval. Some classes of maps have attracted special attention. The most famous class is the **logistic family**,

$$g_a(x) = ax(1 - x).$$

The complete analysis of these maps requires involved theories (like **Teichmüller Theory** or perhaps **Renormalisation Theory**) which are beyond the scope of this book, but a large number of interesting and practically useful properties can be established using elementary mathematics.

Let us illustrate the relevance of the logistic family in applications by considering a metaphor for population problems. Given a population of size x, a basic simplifying assumption is that the offspring produced by the population as a whole at a given time will be proportional to x. The underlying assumption is that all individuals behave roughly the same, so roughly all tend to mate and generate offspring, and a fraction of them will actually succeed, if they are healthy enough and lucky enough. We ignore the individual deterministic conditions and consider an overall, population-wise rate of success. We say then that after one generation, the population changes as $x_{n+1} = ax_n$. We have $a = 1+b-d$, where b is the birth-rate and d the death-rate. Given an initial condition, this equation has the solution $x_n = x_0a^n$. If $a > 1$

x_n grows exponentially and for $a < 1$ it decreases exponentially towards extinction. Since exponential growth/decrease for an indefinite period of time is materially impossible, a possible improvement of the metaphor would be to impose a stop for the growth when the population gets too large. Let us assume that there is a maximum acceptable value M for the population and express x as a fraction of the maximal population. Hence $x \in [0, 1]$ and $x_{n+1} = g_a(x_n)$ is a more reasonable population metaphor since it does not allow for population growth beyond unity, but it still may present an "almost" exponential growth for $x \ll 1$. A fundamental flaw of this metaphor is that populations consist of individuals, i.e., non-negative integers, hence x should be a rational number with denominator M. However, many interesting properties of $g_a(x)$, and even conclusions about the fate of populations, are obtained assuming x to be an arbitrary real number in the unit interval. We will study some properties of the logistic maps for $a \in [0, 4]$ in a series of exercises.

8.1.5. Exercises.

Exercise 8.1. Location of fixed points for the logistic family.

(a) Compute the fixed points of the system $x_{n+1} = g_a(x_n)$ for $a \in [0, 4]$.
(b) For which values of a does the system have genuine period–2 trajectories, i.e., fixed points of $x_{n+2} = g_a(g_a(x_n))$ which are *not* a repetition of the previous fixed points?
(c) Which is the order of the polynomials involved in a similar question for period–4?

Exercise 8.2. Asymptotic stability of fixed points.

(a) Compute the eigenvalue of the linearisation at the fixed point(s) of the logistic family for $a \in [0, 4]$. Colour–code graphically the regions where the fixed points are stable, non–hyperbolic, or unstable.
(b) Repeat the previous computation for the genuine period–2 trajectories, i.e., for the fixed points of $x_{n+2} = g_a(g_a(x_n))$.
(c) Try to answer a similar question for period–$4, \cdots, 2^n$. Use a computer algebra programme to help with the calculations.
(d) Col-or–code and plot the position of the non-zero fixed point as a function of a on a 2–d plot x vs a. Show the period–2 and higher trajectories as well on the same plot.

Exercise 8.3. Another interesting set of plots.

(a) For the non-zero fixed point $x_0(a)$ of the logistic family, compute (numerically if necessary) the "trajectory" of the linearisation eigenvalue on the complex plane parameterised by a, i.e., use a as "time" and compute the trajectory of the curve

$a \to Dg_a(x_0(a))$, for the a values where the eigenvalue remains inside the unit circle.

(b) Repeat the previous computation for the genuine period–2 trajectories, i.e., for the fixed points of $x_{n+2} = g_a(g_a(x_n))$.

(c) Display the above computations on the same plot (using the union of the corresponding a–ranges).

8.1.6. The Tent Map. [**KH96**] For $a = 4$, the logistic map

$$g_4(x) = 4x(1-x)$$

is topologically conjugate to the **tent map**

$$T(x) = \begin{cases} 2x & \text{if } x \leq 1/2 \\ 2(1-x) & \text{if } 1/2 \leq x \ . \end{cases}$$

(see Figure 8.2).

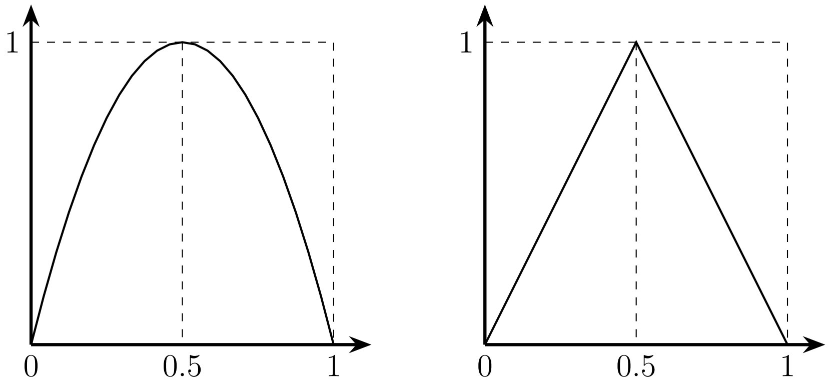


FIGURE 8.2. The logistic map g_4 and the tent map T.

The conjugating map h is given by

$$h(x) = \frac{1-\cos(\pi x)}{2}.$$

To see this we calculate for $x \leq 1/2$ (the other case is similar)

$$\begin{aligned} g_4(h(x)) &= 4h(x)(1-h(x)) \\ &= 4\left(\frac{1-\cos(\pi x)}{2}\right)\left(\frac{1+\cos(\pi x)}{2}\right) \\ &= 1-\cos^2(\pi x) = \sin^2(\pi x). \end{aligned}$$

On the other hand,

$$\begin{aligned} h(T(x)) &= \frac{1-\cos\left(\pi T(x)\right)}{2}) \\ &= \frac{1-\cos\left(2\pi x\right)}{2}) = \sin^2(\pi x). \end{aligned}$$

All dynamical properties of $T(x)$ have a correspondence in $g_4(x)$, except for those properties that depend on derivability at the maximum.

8.1.7. A First Approach to Chaotic Behaviour. A new property that we have seen in this section and was absent in 2–d flows (according to the Poincaré–Bendixson Theorem) and also absent in circle diffeomorphisms is the coexistence of dense orbits and periodic points in the ω–limit set.

Another property that may have revealed itself while performing the exercises on the logistic family is that, when a is allowed to vary, the stable nonzero periodic trajectory of period 2^n undergoes the following fate, for any n: For low values of a it does not exist. At some point it appears as a fixed point of g_a^n with linearisation eigenvalue unity. Then it is stable for some a–interval and finally it becomes unstable at the point where a stable trajectory of period 2^{n+1} appears. For a given value of a, the map displays a collection of unstable trajectories of periods $1, 2, 4, \cdots, 2^{n-1}$ and the stable trajectory of period 2^n. The ω–limit set becomes increasingly complicated as a grows and moreover, nearby points in $[0, 1]$ may have ω–limits lying far apart.

These properties are the "fingerprints" of chaos, in a loose sense they are indicators announcing that the expected dynamics will be very unpredictable and complicated. To understand when a system will be highly unpredictable is very important in applications, so that we do not create false expectations, nor we arrive to inadequate conclusions.

8.2. Higher-dimensional Maps

In dimension 1 we saw so far that interesting ("chaotic") dynamics is produced by local expansion and strong recurrence (the map cannot expand without "returning" since the phase space is bounded). This could only be achieved by non-invertible maps. In higher dimensions one can match the local expansion (at least in some direction) with invertibility. We will now see that already in 2 dimensions one can have "chaotic" invertible dynamics.

One of the main tools to analyse such systems will be the method of coding suggested, as we have seen, by the expansion of a number in base m for expanding maps of the circle of degree m. So the question arises what are the properties we want to have for a "good" coding.

1. We want to divide (partition) the phase space X into finitely many pieces (atoms) $(X_i)_{i=0}^{n-1}$.
2. The pieces X_i intersect only in small regions, essentially they have at most portions of their borders in common (this is needed for the uniqueness of coding on large relevant parts).
3. The intersection $\bigcap_{k\in\mathbb{Z}} f^k(X_{i_k})$ consists of at most one point for any bi-infinite sequence (i_k) (this is needed to assure that no two points have the same coding).

We will see that in some situations we can construct such a coding. In this way we obtain a continuous map $h\colon \Sigma_n = \{0, \cdots, n-1\}^{\mathbb{Z}} \to X$

with the property

$$h \circ \sigma = f \circ h,$$

where σ is an appropriate shift map.

8.2.1. Smale's Horseshoe. . [**GH86, Wig90**]

We consider the following map of the plane specified in the unit square Q (see Figures 8.3 and 8.4)

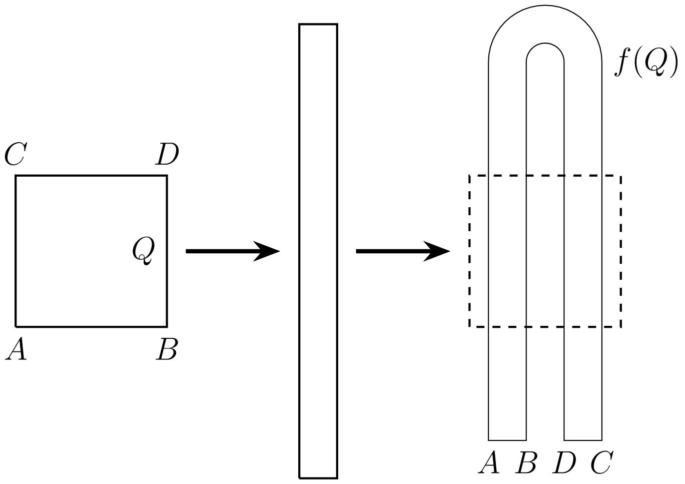


FIGURE 8.3. The horseshoe map.

Smale's original description was defined on a stadium-like region S of the plane, consisting of the square Q and two half disks defined by a half-circle connecting C and D above the square and another connecting A and B below. In this way the map f can take values inside S. Our presentation is nowadays standard practice [**GH86, Wig90**].

The action of f is better visualised in steps. First compress the horizontal direction by a factor of μ and expand the vertical direction by a factor of λ. Then bend the intermediate image of Q in the form of a horseshoe, laying it back on Q.

The inverse map bends back, compresses and expands in reverse order.

In order to succeed with this construction preserving continuity of the map (nothing on the square is teared apart), it is needed that $\lambda > 2$ and $\mu < \frac{1}{2}$. These numbers are called *expansion rate* and *contraction rate* respectively. These rates need not be uniform (non uniformity leads to curvy lines instead of straight lines and to transverse crossings at angles other than $\pi/2$) and their actual value is not important as long as they assure that Figures 8.3 and 8.4 can be drawn. However, we will see below that μ and λ are very useful to establish interesting properties of the invariant set of this map.

It is clear from the pictures that there are points of Q that are mapped back on Q by both f and f^{-1}. We may ask the question if

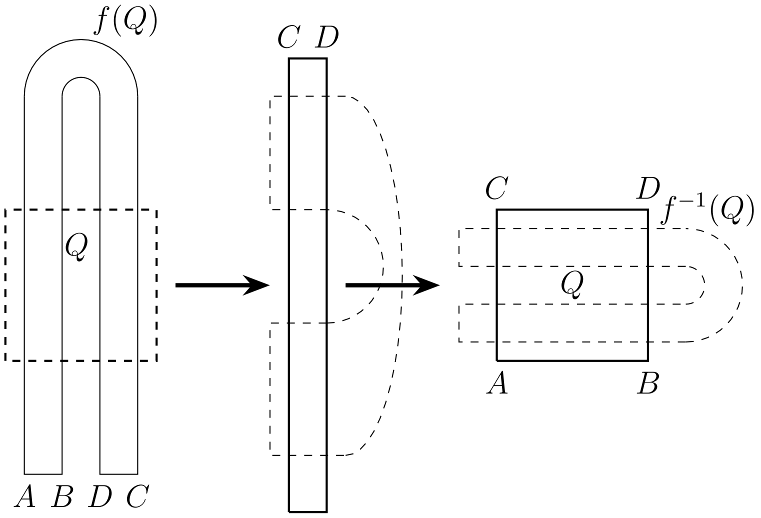


FIGURE 8.4. The inverse map.

there are any points that never leave Q upon repeated iteration of f and f^{-1}. The answer is positive. This will be the invariant set

$$\Lambda = \bigcap_{n\in\mathbb{Z}} f^n(Q)$$

called the **horseshoe**. Let us investigate its structure. The intersection

$$Q \cap f(Q) \cap \cdots \cap f^n(Q)$$

is the union of 2^n disjoint vertical rectangles (see Figure 8.5, the two gray-shaded rectangles correspond to $n = 1$ while the four black strips correspond to $n = 2$).

The intersection

$$Q \cap f^{-1}(Q) \cap \cdots \cap f^{-n}(Q)$$

is the union of 2^n horizontal rectangles, corresponding to a rotated picture similar to Figure 8.5. Therefore

$$f^{-n}(Q) \cap \cdots f^{-1}(Q) \cap Q \cap f(Q) \cap \cdots \cap f^n(Q)$$

is the union of 4^n small squares (if $\lambda^{-1} = \mu$ or rectangles of size $\lambda^{-n} \times \mu^n$ in general) (see Figure 8.6).

The reader may be reminded of the construction of a Cantor set. The vertical strips in the forward iteration n have width μ^n and are contained within the strips of the previous iteration. Hence,

$$\Lambda^+ \equiv \bigcap_{n\geq 0} f^n(Q) = \mathcal{C} \times [0, 1].$$

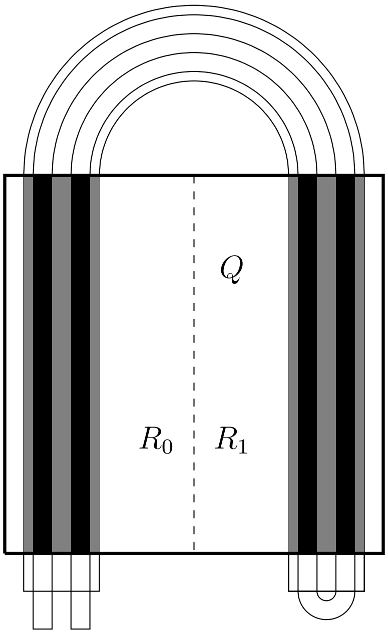


FIGURE 8.5. Two iterates of f. The gray shaded area is $Q \cap f(Q)$ while the black shaded area is $Q \cap f(Q) \cap f^2(Q)$.

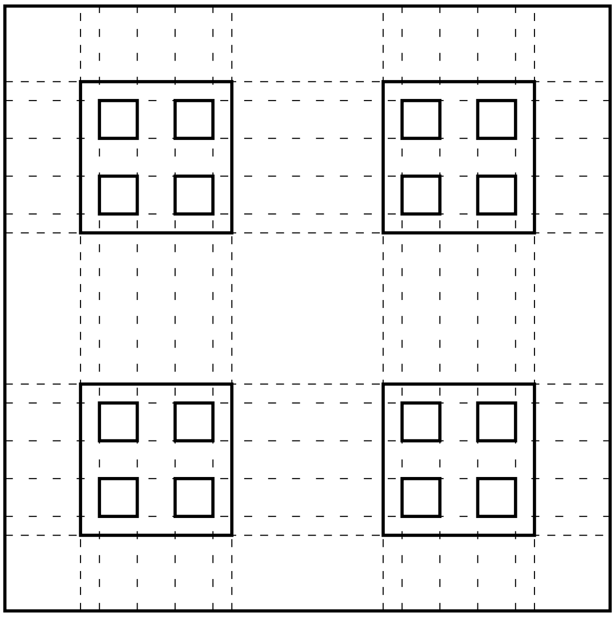


FIGURE 8.6. The first two iterations of f and f^{-1} put together.

In the same way, the horizontal strips of the inverse iterates have width λ^{-n} and

$$\Lambda^- \equiv \bigcap_{n \geq 0} f^{-n}(Q) = [0,1] \times \mathcal{C}.$$

Finally, the horseshoe can be seen as the intersection of the forward and backward iterates, $\Lambda = \Lambda^+ \cap \Lambda^-$, in other words

$$\Lambda = \bigcap_{n\in\mathbb{Z}} f^n(Q) = \mathcal{C} \times \mathcal{C},$$

where $\mathcal{C}$ is a **Cantor set**.

8.2.1.1. *Coding Space on Two Symbols.* We can introduce the following coding with the help of two vertical rectangles R_0 and R_1 (see Figure 8.5 again), which we take to be left and right halves of Q, separated by a common central vertical line. We note on passing that the left vertical strip of $f(Q)\cap Q$ lies inside R_0 while the right one lies inside R_1. By way of the horseshoe construction, these strips are horizontally separated by some positive distance $0 < \eta < 1 - 2\mu$. We could therefore consider an equivalent coding where the left strip of $f(Q) \cap Q$ "is" R_0 and the right one "is" R_1, since no point of Q outside these strips belongs to Λ. Either choice leads to the same dynamical properties. The two horizontal rectangles H_0 (lowermost) and H_1 associated to the inverse map will be useful as well. We define

$$x \longleftrightarrow \underline{x} = \cdots x_{-n} \cdots x_{-1}.x_0 x_1 \cdots x_n \cdots \in \Sigma_2 \qquad \Longleftrightarrow \qquad f^n(x) \in R_{x_n}$$

for all $n \in \mathbb{Z}$. We consider the bi-infinite past and future itinerary of each point $x \in \Lambda$, and we construct $\underline{x}$ by placing the label (0 or 1) of the rectangular strip where $f^n(x)$ lies in position n. Every central portion $x_{-M} \cdots x_0 \cdots x_M$ (M positive integer) of $\underline{x}$ identifies a small square inside Q. The bi-infinite sequence $\underline{x}$ identifies a point in Λ. The "dot" in the sequence identifies which element corresponds to $n = 0$. This coding is continuous (having an appropriate metric on Σ_2) and bijective, hence a homeomorphism:

$$h\colon \Lambda \to \Sigma_2.$$

Here we propose the following metric on Σ_2

$$d(\underline{x}, \underline{y}) = \frac{1}{2^N}, \quad N = \max\{n : x_i = y_i \quad |i| \le n\}.$$

N is the last position up to where both strings coincide, starting to compare the strings at $n = 0$ and moving away in both directions. The reader should verify that $d(\underline{x}, \underline{y}) = 0 \Leftrightarrow \underline{x} = \underline{y}$. An even more natural metric, when comparing with the base–2 coding introduced for the one–dimensional case, is:

$$d_{b2}(\underline{x}, \underline{y}) = \sum_{n\in\mathbb{Z}} \frac{|x_n - y_n|}{2^{|n|}}.$$

Note that $2d \ge d_{b2}$, i.e., $2d$ can be computed from d_{b2} by putting all subsequent numerators equal to one after the first nonzero numerator is reached, while moving away from $n = 0$. In any case, the metric turns Σ_2 into a compact metric space (actually a Cantor set). Two points in Σ_n are close if and only if their images under h are in some small square, i.e., close in Λ.

How does f look like if we represent Λ with Σ_2? Defining the *shift map* σ on sequences of Σ_2,

$$\sigma(\cdots x_{-n}\cdots x_{-1}.x_0x_1\cdots x_n\cdots) = \cdots x_{-n}\cdots x_{-1}x_0.x_1\cdots x_n\cdots$$

we have that

$$h\circ\sigma = f\circ h;$$

i.e., the action of f is just shifting the dot on a sequence one step to the right.

Corollary 8.1. *Horseshoe properties:*[**KH96**, 2.5.1 p.83]

1. $\overline{\bigcup_{n\in\mathbb{N}} Per_n(f)} = \Lambda$,
2. $\#Per_n(f) = 2^n$,
3. $f|_\Lambda$ *is mixing.*

Proof. The proof follows from the properties of the symbolic space Σ_2. In Σ_2 a point is periodic if and only if its coding is periodic, i.e.

$$\underline{x} = (x_0\cdots x_n)^\infty.$$

Therefore, any cylinder

$$C^n_{-n}(\underline{x}) := \{\underline{y}\in\Sigma_2 \,:\, y_k = x_k,\ -n\le k\le n\}$$

contains a periodic point (of period $2n+1$) and the first statement is proved. The second statement follows immediately since there are exactly 2^n words on two symbols having length n. These words serve as the building blocks for a period–n point.

To derive the mixing property, we consider two arbitrary cylinders $[x_{n_1}\cdots x_{m_1}]$ and $[y_{n_2}\cdots y_{m_2}]$. Let us verify that the image of one of the cylinders intersects the other cylinder for some n. Indeed, if $n > m_2 - n_1$ then

$$\varnothing \ne [y_{n_2}\cdots y_{m_2}]\cap\sigma^n([x_{n_1}\cdots x_{m_1}]) =$$
$$= \{\underline{z}\in\Sigma_2 \,:\, z_{n_2} = y_{n_2}\cdots z_{m_2} = y_{m_2}; z_{n+n_1} = x_{n_1}\cdots z_{n+m_1} = x_{m_1}\}.$$

□

Remark 8.4. Some of the periodic points in the second item, however, are not "genuinely" period–n, for example, the string 1111 counted at $n=4$ is just one of the (two) period–1 orbits. In any case, there are many genuinely period–n points for all n (for prime n, there are 2^{n-1} true period–n orbits).

Now we want to derive some useful analytic properties.

Definition 8.2. Let $f\colon X\to X$ be a homeomorphism. For $x\in X$ we define its **stable set (manifold)** as

$$W^s(x) := \{y\in X \,:\, d(f^n(y), f^n(x))\to 0 \text{ as } n\to+\infty\}.$$

The **unstable set (manifold)** $W^u(x)$ of x is defined as the stable set for f^{-1} at x.

Remark 8.5. These manifolds may be defined in a corresponding way for vector fields, by replacing discrete time n with continuous time t. We have already encountered invariant manifolds in Chapter 6.

Theorem 8.4. [**KH96**, 6.4.9 p.267] *The stable sets for the points in the horseshoe Λ are actually smooth manifolds.*

This result follows from a more general theorem which is proved in the same spirit as the Hartman–Grobman Theorem and is beyond the scope of this book. The Theorem holds for any diffeomorphism and any point having hyperbolic behaviour, i.e., having exponential expansion and exponential contraction along the orbit in some directions.

Remark 8.6. The stable and unstable manifolds are invariant:

$$\begin{aligned} f(W^s(x)) &= \{f(z) \in X \,:\, d(f^n(x), f^n(z)) \to 0 \text{ as } n \to +\infty\} \\ &= \{y = f(z) \in X \,:\, d(f^n(fx), f^n(y)) \to 0 \text{ as } n \to +\infty\} \\ &= W^s(f(x)). \end{aligned}$$

Moreover any conjugacy transports stable manifolds into stable manifolds, since

$$d_1(f_n(x), f^n(y)) = d_1(h^{-1} \circ g^n \circ h(x), h^{-1} \circ g^n \circ h(y) \to 0$$

if and only if

$$d_2(g^n(x), g^n(y)) \to 0$$

where $h\colon X_1 \to X_2$ is a continuous conjugacy.

Theorem 8.5. [**HP69**, Th. 1] *For all $x \in \Lambda$ we have*

$$\overline{W^s(x)} \cap \Lambda = \overline{W^u(x)} \cap \Lambda = \Lambda,$$

i.e., the stable and unstable manifolds are dense.

Proof. Since h is a homeomorphism we prove the theorem on Σ_2, since any homeomorphism transports stable (unstable) sets into stable (unstable) sets. By the definition of stable set $W^s(\underline{x})$, we have

$$\underline{y} \in W^s(\underline{x}) \iff (\exists\, n_0 \in \mathbb{N})\colon (\forall n > n_0) \quad x_n = y_n.$$

Otherwise, there would be a sequence $n_k \nearrow +\infty$ with $x_{n_k} \neq y_{n_k}$ and $d(\sigma^{n_k}(\underline{x}), \sigma^{n_k}(\underline{y})) = 1$ and the orbits would not converge. Now consider an arbitrary **cylinder set** of the form

$$C_n^m(\underline{y}) := \{\underline{z} \in \Sigma_2 \,:\, y_i = z_i \quad -m \le i \le m\} \quad n, m \in \mathbb{N}.$$

We need to find in each such cylinder a point of the stable set. In fact, such point will be

$$\underline{w} = w_0 w_1 \cdots \quad \begin{cases} y_i & \text{if } -m \le i \le n \\ x_i & \text{if } n+1 \le i \\ 0 & \text{otherwise.} \end{cases}$$

□

REMARK 8.7. Interesting facts about the horseshoe.

1. The horseshoe plays a similar role as a hyperbolic fixed point. Nearby any point of the horseshoe, there are trajectories that are "repelled" (away from that point) and those that are "attracted" (approaching the point).
2. The Lebesgue measure[1] of Λ is zero because for each $n \in \mathbb{N}$ we can cover the set Λ by 4^n squares of area a^{2n} with $a = \max(\mu, \lambda^{-1}) < \frac{1}{2}$, where μ (λ) is the contraction (expansion) rate. Hence, a complete covering of Λ has area $A_n = (4a^2)^n$ and we have $\lim_{n\to\infty} A_n = 0$
3. With the help of the coding we can conclude (similarly to the case of expanding maps on the circle) that the horseshoe is **structurally stable**: Any "nearby" map can be coded in the same way and hence it will have the same (conjugate) dynamics.

DEFINITION 8.3. A homeomorphism $f\colon X \to X$ is called **expansive** if there is a $\delta > 0$ such that

$$d(f^n(x), f^n(y)) < \delta \text{ for all } n \in \mathbb{Z} \iff x = y.$$

THEOREM 8.6. [**KH96**, 6.4.10 p.268] *The horseshoe is expansive.*

PROOF. If $x = y$, then $\underline{x} = \underline{y}$ and since d is a proper distance, it is immediate that $d(f^n(\underline{x}), f^n(\underline{y})) = 0$ for any n.

Let now $x \neq y$. To complete the proof we need to show that their images will eventually be moved to a distance at least $\delta > 0$. Clearly, $x \neq y \Leftrightarrow \underline{x} \neq \underline{y}$ and hence there will be a least $|n|$ such that $x_n \neq y_n$, this means that $f^n(x)$ and $f^n(y)$ lie on vertical strips belonging to different coding rectangles and therefore at a distance $\delta \geq \eta > 0$, where η is the separation between the first vertical strips defined previously. Hence, x and y lie further away than some positive distance. □

8.2.2. How do Horseshoes Appear? [**GH86, SNM96**] Let x be a hyperbolic fixed point of saddle type of a diffeomorphism f. A point $p \in W^s(x) \cap W^u(x)$ is called a **homoclinic point**. In the case of flows such points have rather simple properties, since stable and unstable manifolds cannot cross transversely. Either they are disjoint, or one of them is a subset of the other. In the latter case, we speak of a **homoclinic orbit**. In maps, these manifolds may cross transversely, since there is nothing corresponding to e.g. the Picard–Lindelöf Theorem to impose additional restrictions.

Let us make some acquaintance with these manifolds in the case of a horseshoe, as a preparation to answer the title question. Two different points $x \neq y$ of the horseshoe have manifolds that intersect. In fact,

[1] If the reader is unfamiliar with Lebesgue measure, just read "standard" measure of plane areas.

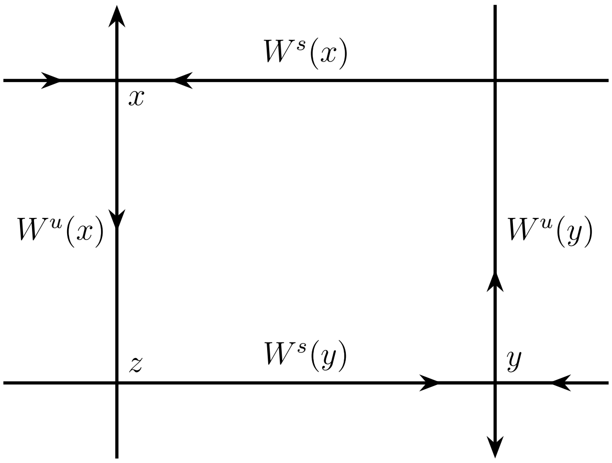


Figure 8.7. The point $z = W^u(x) \cap W^s(y)$.

$\underline{z} = \cdots x_{-2}x_{-1}.y_0y_1y_2\cdots$ approaches y for positive n and approaches x for negative n, see Figure 8.7. In the same way, we can produce points belonging to either $W^s(x)$ or $W^u(x)$ (but not to both) as well as points in $W^u(x) \cap W^s(y)$. Indeed, $\underline{z(u)} = \cdots\cdots x_{-2}x_{-1}.z_0z_1z_2\cdots$, with $z_i = |1 - x_i|$ for $i \geq 0$ lies on the unstable manifold of x while similarly $\underline{z(s)} = \cdots\cdots z_{-2}z_{-1}.x_0x_1x_2\cdots$ with $z_i = |1 - x_i|$ for $i < 0$ lies on the stable manifold of x. Any point $\underline{z(su)}$ such that for $|n| \geq N_0 > 0$ the coding symbols satisfy $z(su)_n = x_n$ while for $|k| < N_0$ we have $z(su)_k = |1 - x_k|$ will lie in the intersection of both manifolds, and in all cases $z(su) \neq x$.

Hence, in the general case we may expect that there will be diffeomorphisms f with homoclinic points p. We have $\alpha(p) = \omega(p) = x$. Note that this cannot happen in a linear system! If the intersection

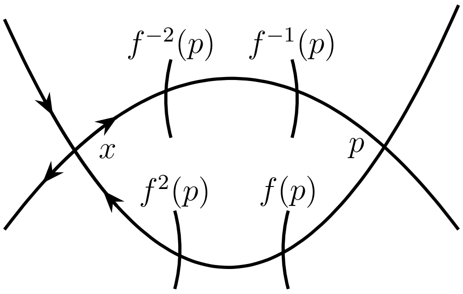


Figure 8.8. A sample of the infinitely many transversal homoclinic points existing along with p.

of $W^s(x)$ and $W^u(x)$ is transversal we call the point a **transversal homoclinic point**. We note that the entire orbit of p consists of transversal homoclinic points (see Figure 8.8) since $\omega(f^n(p)) = \omega(p)$ for any n (and the same holds for the α–limit). Note that for $n > 0$,

$f^n(p)$ is just a few steps ahead of p along $W^s(x)$, while for negative n it is a few steps behind. The angle between the stable and unstable manifold may vary when iterating with f but they can never be tangent.

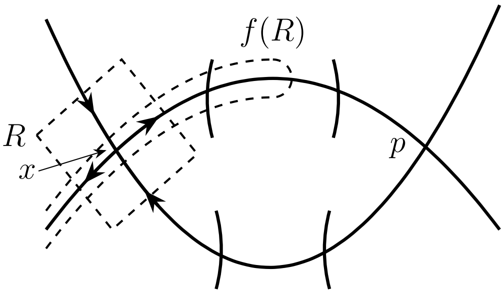


Figure 8.9. The action of f on a small rectangle R around the fixed point x.

Now we consider a small rectangle R around x and its image by a high positive iterate of f (see Figure 8.9). Since x is hyperbolic, the image of a sufficiently small rectangle R will be approximately linear, and hence expanded along the unstable direction and compressed along the stable direction.

Since a large portion of $W^u(x)$ lies within $f(R)$, higher iterates $f^k(R)$ will stretch the image along $W^u(x)$ until it eventually reaches p and for even higher iterates the image will return to R along with $f^n(p)$. The situation for large n will resemble that of Figure 8.10. The area

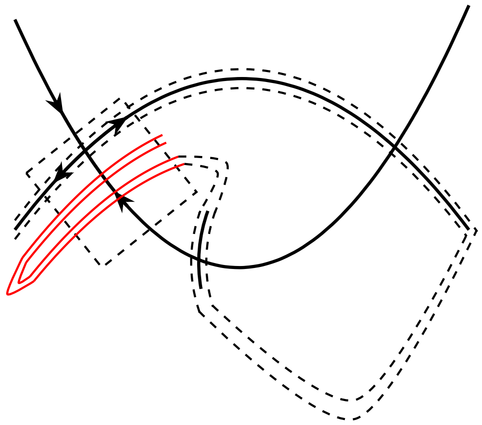


Figure 8.10. A higher iterate $f^n(R)$ returns to R bended in the form of a horseshoe (in red).

in red has all qualitative features of the horseshoe construction. This idea has been formalised by Smale in the 60's. The above discussion sketches the proof of the following (see also [**Sma67**] and [**KH96**, 6.5.5 p.276])

THEOREM 8.7. [**GH86**, 5.3.5 p.252] *If p is a transversal homoclinic point of a planar diffeomorphism f with hyperbolic fixed point x, then there exists a Cantor set Λ and a positive integer m such that $x \in \Lambda$, $f^m(\Lambda) = \Lambda$ and f^m restricted to Λ is conjugate to a shift map on bi–infinite sequences on 2 symbols.*

EXAMPLE 8.1. [**GH86**, 6.6 p.325] [**SNM96**, p.161-163] Let us perform a rigorous approximate calculation showing that, starting at some positive integer M, the Horseshoe has infinitely many periodic orbits, with all periods $k \geq M$.

Consider Figure 8.9, place the origin of coordinates at the fixed point, and the x- and y-axes along the stable respective unstable directions. Assume now that R in the figure is contained in the region where the Hartman-Grobman linearisation holds and pick a point $(0, Y) \in R$, pre-image of the homoclinic crossing p. Hence, after $n > 1$ iterates of the map f, the point $(0, Y)$ will map on some point $(X, 0) \in R$. By continuity, the map f^n on a small neighbourhood $U \subset R$ of $(0, Y)$ will take the approximate form

$$\begin{aligned} x' &= X + ax + b(y - Y) \\ y' &= cx + d(y - Y), \end{aligned}$$

where we take x and $y - Y$ small enough so that the nonlinear terms can be safely neglected. Here, $d \neq 0$ represents the condition of transversal intersection at $(X, 0)$ and invertibility is given by $ad - bc \neq 0$. This representation holds at least for the images (x', y') lying on R. Such points can be further mapped using the linearisation of the original map. After additional $m \geq 0$ iterates, the images will lie approximately on

$$\begin{aligned} x'' &= \lambda_-^m x' \\ y'' &= \lambda_+^m y', \end{aligned}$$

where λ_-, λ_+ are the stable and unstable eigenvalues of the original map f. Some images will return to the region U, the overall structure looking more or less as in Figure 8.10. Combining both maps, these images can be approximately computed as:

$$\begin{aligned} x'' &= \lambda_-^m (X + ax + b(y - Y)) \\ y'' &= \lambda_+^m (cx + d(y - Y)). \end{aligned}$$

Periodic points of period $k = n + m$ satisfy hence, the approximate equation

$$\begin{aligned} x &= \lambda_-^m (X + ax + b(y - Y)) \\ y &= \lambda_+^m (cx + d(y - Y)), \end{aligned}$$

where $(x, y) \in U \subset R$. Letting $M = n + m_0$ we recover the initial statement if we manage to prove that for all $m \geq m_0$ (i.e., for sufficiently large m) the periodic point equation above has a solution. To do this, setting $u = y - Y$ let us rewrite the equation as

$$\begin{pmatrix} -\lambda_-^m X \\ \lambda_+^{-m} Y \end{pmatrix} = \begin{pmatrix} \lambda_-^m a - 1 & \lambda_-^m b \\ c & d - \lambda_+^{-m} \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix}.$$

For every sufficiently large m this linear equation has a unique solution $(x, u)_m$ since the determinant of the matrix is nonzero (it is approximately equal to $-d$ since $|\lambda_-^m|$ and $|\lambda_+^{-m}|$ become arbitrarily small for sufficiently large m).

Remark 8.8. Summary of horseshoe properties. An invariant set of horseshoe type has

1. Periodic orbits of period p for all $p \in \mathbb{N}$.
2. The number of distinct genuine periodic orbits of period n grows (roughly) exponentially for n large.
3. Uncountably many dense, non–periodic orbits.

We have seen that a horseshoe has an extremely complicated structure Also, horseshoes appear as soon as we have a transversal homoclinic point. In some sense, we may say that chaotic systems always contain subsystems of horseshoe type. This means that in order to understand general chaotic dynamics it is useful to understand the horseshoe first. We will pursue this understanding with the help of probability theory in the next chapter.

8.2.3. Exercises.

Exercise 8.4. (a) Show that if p, q are different periodic points in Λ, then $W_s(p) \cap W_u(q) \neq \varnothing$ (**Hint**: Produce a symbolic sequence belonging to this intersection).
(b) Use the fact that there exist a dense orbit in Λ to prove that Λ cannot have *stable* periodic points. In fact, all periodic points are of *saddle* type, i.e., with both stable and unstable manifolds.

Exercise 8.5. **Homoclinic tangencies**[**SNM96**, p.164]: Assume that the origin is an hyperbolic fixed point of a μ-dependent family of planar maps such that $W_u(0)$ has a tangency point with $W_s(0)$ for $\mu = 0$. An approximation of this map with the same setup as in Example 8.1, mapping a small neighbourhood of $(0, Y)$ to a small neighbourhood of $(X, 0)$ reads

$$\begin{aligned} x' &= X + ax + b(y - Y) \\ y' &= \mu + cx + e(y - Y)^2. \end{aligned}$$

Note now that since there is a *tangency* we have no transversality (i.e., $d = 0$ in Example 8.1) and the lowest nonlinear contribution has to be included in the description. Compute a return map from a suitable neighbourhood of $(0, Y)$ to itself, find periodic orbits and their stability eigenvalues as a function of μ. Compute the values μ_k where saddle-node bifurcations occur. Show further that $\lim_{k\to\infty} \mu_k = 0$.

8.2.4. The Solenoid. [**GH86**] Another "Cantor set" object that is easy to describe with simple mathematics is the Solenoid. Consider a map f from the solid torus V into itself (see Figure 8.11)

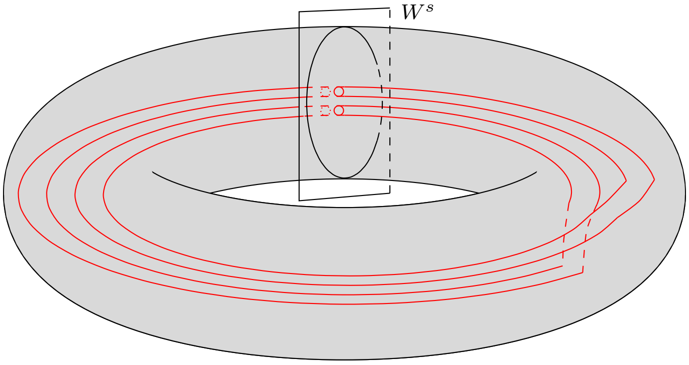


FIGURE 8.11. The Solenoid map.

$$f(t,x,y) = \left(2t \mod 1, \frac{1}{10}x + \frac{1}{2}\cos(2\pi t), \frac{1}{10}y + \frac{1}{2}\sin(2\pi t)\right),$$

where x, y are coordinates on the unit disc, $x^2 + y^2 \leq 1$ and $t \in \mathbb{S}^1$.

The set

$$\Gamma := \bigcap_{n\in\mathbb{Z}} f^n(V) = \bigcap_{n\in\mathbb{N}} f^n(V)$$

is called the **solenoid**, or Smale–Williams attractor. It is an **attractor** since it attracts all nearby points. Note first that $f(V) \subset V$, this is clear from the definition. Moreover, for fixed $t = t_0$, $f(t_0, x, y)$ is a contraction (see W^s depicted in Figure 8.11). Hence, all points in the torus V approach Γ. This set has locally the structure of a product of an interval and a Cantor set. It has a quite peculiar topological structure:

1. The intersection of Γ with the unit disc (for each fixed $t = t_0$) is a Cantor set.
2. Γ consists of uncountable many "threads" since the Cantor sets for each t continue locally into each other. However, each "thread" can intersect the unit disc at $t = t_0$ only countably many times.
3. Γ is itself connected, since it is the intersection of nested connected sets (linearly ordered by inclusion).

Sets of the kind of the solenoid are called **indecomposable continua**. They play an important role in topology.

8.2.4.1. *Coding.* Consider the "rectangles"

$$R_0 = [0, \frac{1}{2}] \times \mathbb{D}^2 \quad R_1 = [\frac{1}{2}, 1] \times \mathbb{D}^2,$$

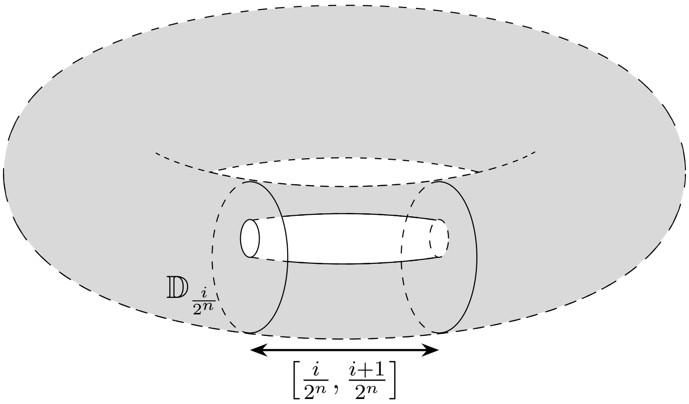

$\mathbb{D}_{\frac{i}{2^n}}$
$\left[\frac{i}{2^n}, \frac{i+1}{2^n}\right]$

FIGURE 8.12. coding cylinders on the solenoid.

where $\mathbb{D}^2$ is the unit disc. Then we get a coding map from the binary strings to the points of the solenoid:

$$\pi\colon \Sigma_2 \to \Gamma.$$

We list some properties of the coding map:

1. The coding is not one–to–one in the first coordinate. It is 2–to–1 at the dyadic points and otherwise one–to–one (compare with the expansion in base 2).
2. The cylinder sets have the form
$$[\frac{i}{2^n}, \frac{i+1}{2^n}] \times D_j^m$$
where D_j^m is a connected component of $f^m(V) \cap \mathbb{D}_{\frac{i}{2^n}}$ (see Figure 8.12).
3. The solenoid is for the same reasons as the expanding maps on the circle or the horseshoe, **structurally stable**.
4. The solenoid contains transversal homoclinic points. The simplest can be found as the intersection of the stable and unstable manifold of the fixed point. The fixed point is the image of 0^∞ under the coding:
$$\left(0, \frac{1}{2}\sum_{n=0}^{\infty} 10^{-n}, 0\right) = \left(0, \frac{5}{9}, 0\right).$$

8.3. Strange Attractors and Chaos

Let us now make some considerations about what is meant by expressions such as "chaos" or "chaotic dynamics". Given a dynamical system, the goal in applications is to understand its asymptotic behaviour for $t \to \infty$ (or such large time that the system behaviour cannot be distinguished in practice from that in the limit). In the end we are looking for ω–limit sets, attractors and their corresponding **basin of attraction**, i.e., which regions of phase space approach what portion of the ω–limit set. For simple systems it may suffice to compute fixed points and their stability. More complicated systems are less understood, while different techniques of analysis give different partial understanding. Intuitively, we would like to say that a system is chaotic when its attractor is more complicated than a finite set of fixed points, periodic orbits and connecting orbits, i.e., something "beyond" what is given by the Poincaré–Bendixson Theorem.

For vector fields in continuous time, one of the examples that became classical is that of the Lorenz equations. It is a nice exercise to show that the equations have a trapping region consisting of a fairly large ball encompassing the origin. Any initial condition on the surface of that ball will remain inside it for positive times. The question is what is the attractor lying inside the region? For certain parameter values, the attractor is just a fixed point. For other parameter values, the attractor gets modified, until we come to a situation where we still have a trapping region, but there exists neither a stable fixed point nor a stable periodic orbit inside the region. The attractor, whatever it is, must be more complicated than that.

8.3.1. Chaotic Behaviour. Another related question arising in applications is that of the **routes to chaos**. For families of dynamical systems depending on parameters, we seek to understand in which way the system "evolves" from simple dynamics to complicated dynamics when varying these parameters (it is a peculiar "evolution" since we are evolving in parameter space, time is always infinite when considering asymptotic behaviour). The classical example in maps has come to be the **period doubling cascade** of the logistic family: When a varies monotonically from 0 towards 4, the attractor of the logistic map g_a changes at fixed (discrete) values of a, going from a fixed point at zero, to a nonzero fixed point, to period–doubled attractors, and more.

Routes to chaos are somehow the opposite of structural stability. While in some parameter intervals we may have structural stability, along the route, the nature of the ω–limit set changes drastically at certain fixed parameter values. These changes may eventually "finish" in the appearance of a horseshoe. Let us be more precise.

Definition 8.4. A **strange attractor** is an attractor that contains a transversal homoclinic orbit (of a hyperbolic periodic point). A system is *chaotic* if it has a strange attractor.

As a consequence, a strange attractor contains within itself infinitely many periodic orbits and dense, non-periodic orbits. None of these periodic orbits is asymptotically stable, otherwise the set would be decomposable (not transitive) and hence not an attractor.

The intuitive image is that the orbits within the attractor are of saddle-type. When an initial condition passes close enough to the stable manifold of some orbit, the dynamics "mimics" this orbit until we get so close to it that the system "escapes" along the unstable manifold. Since we are never exactly "on" a stable manifold, this process goes on forever, shifting from one periodic orbit to another. In fact, lying on a dense orbit, the system eventually passes ϵ–close to any periodic orbit, for any given $\epsilon > 0$.

Alternatively, Devaney's definition of chaos gives a clear picture of chaotic behaviour [**Dev86**]:

Definition 8.5. The mapping $f : V \to V$ is said to be **chaotic** on the set V if:
1. f has sensitive dependence on initial conditions,
2. f is topologically transitive in V,
3. periodic points are dense in V.

8.3.2. Fingerprints of Chaos. For real systems, we can at best guess the presence of chaos. Even if the equations we use to describe our system are chaotic (they display a strange attractor), to what extent or precision the equations describe the system may always be a matter of debate.

When we can refer to a well-defined set of equations the quest for chaos goes through finding a trapping region, verifying that there is no stable fixed point or periodic orbit inside the region and finding a transverse homoclinic orbit within the attracting set. In some cases we can do part of the job, namely we find the trapping region and we find the transverse homoclinic orbit, but we cannot prove e.g., that this orbit belongs to the attractor. When nothing more precise can be done, we rely on "indicators" of crucial properties that we have encountered in other, better known, chaotic systems. However, we must bear in mind that necessary conditions should not be mixed up with sufficient conditions.

1. **Sensitivity to initial conditions**, meaning that $x \neq y \Rightarrow |f^n(x) - f^n(y)| > 1$ for n large enough (or a corresponding statement for flows). If this can be established for pairs x, y lying arbitrarily close to each other and for a large region (positive measure) of initial conditions in phase space, if not chaos we have at least one fundamental consequence of it, namely the impossibility of making long-term dynamical predictions (because our knowledge of initial conditions is never perfect).
2. **Sensitivity to parameter variation**, meaning that we have a family of dynamical systems where the ω–limit set changes drastically when changing parameters. These changes bear the collective name of **bifurcations**. Not all bifurcation cascades will

lead to the formation of e.g., a horseshoe, but again the impossibility of making long-term dynamical predictions is present, since similarly to initial conditions, the constant value of a given parameter cannot be specified beyond the precision of our experiment and/or numerical computation.

8.4. Exercises

Exercise 8.6. Let $A\colon \mathbb{T}^2 \to \mathbb{T}^2$ be given by

$$(x, y) \longrightarrow (2x + y, x + y) \quad \text{mod } 1.$$

What is the number of periodic points of period n? Are the periodic points dense?

Hint: Argue that the number of periodic points of period n is the number of pre-images of the point $(0, 0)$ under the (non-invertible) map

$$(A^n - Id) : \mathbb{T}^2 \to \mathbb{T}^2.$$

Then argue that the number of pre-images is for all points the same. From this conclude that the number of pre-images equals to the area of

$$\left[\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}^n - Id\right] ([0, 1] \times [0, 1]) .$$

For the density part, verify that all points with rational coordinates are periodic.

This map is area preserving, it expands in one direction and contracts in the other. It is called the *Arnold Cat Map* and it is a classical example in Dynamical Systems Theory. Search it in the web!

Exercise 8.7. Show that the homoclinic points, i.e., points $x \in \Lambda$ with $\alpha(x) = \omega(x) = p$ for some periodic point p, are dense in the horseshoe. Are the homoclinic points for a fixed point dense?

CHAPTER 9

Ergodic Theory

We have seen that some systems have a very chaotic, i.e., sensitive, dependence of the trajectories with respect to their initial data. In such cases, it is unreasonable to attempt predict the long-time-behaviour of all trajectories. One way out could be to consider only **typical** or **randomly chosen** trajectories and investigate their statistical properties. This idea goes back to Boltzmann. He considered a hard-ball-gas and conjectured that if one chooses the initial distribution arbitrarily then the system will "mix-up" during its time evolution and will approach the uniform disordered distribution (see Figure 9.1)

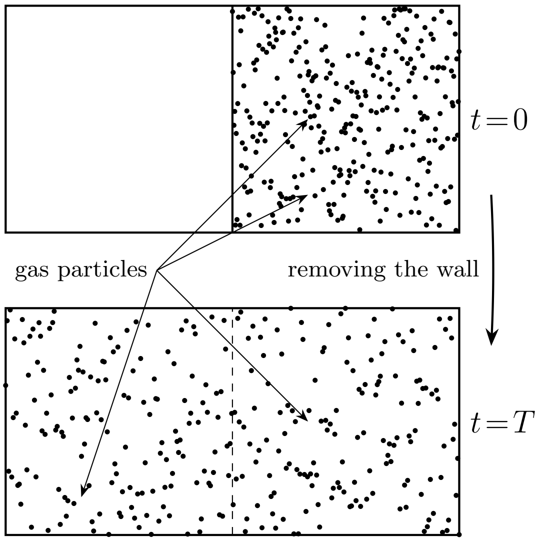


FIGURE 9.1. An expanding gas evolves towards equilibrium.

This could be expressed in the following way. Let ϕ be an observable (e.g., temperature, pressure, etc.) and x_0 the initial configuration (e.g., the configuration of the gas particles). The measured observable is then $\phi(x_0)$. The average over all configurations (or macro-observable) is given by $\int_X \phi(x)\,dx$ where X is the space of all configurations, i.e., the phase space, and dx is the volume measure. The **Boltzmann Ergodic Hypothesis** (BEH)[**Bol71**] claims that the time evolution approximates the state average.

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\phi(f^n x_0) = \int_X \phi\, dx.$$

Here $f\colon X\to X$ is the evolution law. Sometimes this formula is expressed as *"time average equals phase space average"*.

Before going into some details we will discuss the hypothesis on examples.

Example 9.1. Let $E_2\colon [0,1)\to[0,1)$ be the doubling map $x \longrightarrow 2x$ mod 1. We consider the observable

$$\phi(x) = \begin{cases} -1 & \text{if } \ 0\le x<\frac{1}{2} \\ +1 & \text{otherwise.}\end{cases}$$

Then

$$\int_{[0,1)}\phi(x)\, dx = \frac{1}{2}\cdot(-1)+\frac{1}{2}\cdot(+1) = 0.$$

If $x_0 = 0$ then

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\phi(E_2^n x_0) = \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\phi(x_0) = -1\neq 0.$$

Hence the BEH does not hold for the initial value $x_0 = 0$. However, we will see that $x_0 = 0$ is in some sense an exception.

Let $x = x_1x_2x_3\cdots\in[0,1)$ be a number with the "right" frequency of digits in base 2

$$\lim_{N\to\infty}\frac{1}{N}\#\left\{1\le n\le N \,:\, x_n=0\right\} = \lim_{N\to\infty}\frac{1}{N}\#\left\{1\le n\le N \,:\, x_n=1\right\} \\ = \frac{1}{2}.$$

By the **law of large numbers** the "probability of choosing such a number at random" is 1 (i.e., 100%). Now for such an x_0 we have

$$\begin{aligned}\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\phi(E_2^n x_0) &= \\ = \lim_{N\to\infty}\frac{1}{N}\big(\#\left\{1\le n\le N \,:\, x_n=0\right\}&\cdot(-1)+ \\ +\#\left\{1\le n\le N \,:\, x_n=1\right\}&\cdot(+1)\big) \\ = \frac{1}{2}\cdot(-1)+\frac{1}{2}\cdot(+1) = 0 &= \int_{[0,1)}\phi(x)\, dx.\end{aligned}$$

So for a "typical" point the BEH is valid.

Our main aim is to show that this situation holds generally. For this we need some preparation.

9.1. Naive Measure Theory

Measure theory is complicated in general and a fundamental part of mathematics. We will illustrate some of the basics being aware that this is no rigorous treatment.

A measure μ should be a function on subsets of X which assigns to it its "size" (like length, area, volume), i.e.,

$$\mu\colon \mathcal{B} \subset 2^X \to \mathbb{R}^+ \cup \{0\}. \tag{9.1}$$

Here 2^X denotes the set of all subsets of X. We ask this function to preserve some fundamental properties,. First, the measure should not change if we "cut a set into disjoint pieces":

$$\mu\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} \mu(A_i), \quad A_i \in 2^X, \quad A_i \cap A_j = \varnothing; \quad i \neq j. \tag{9.2}$$

Second, we want the total measure of the space to be equal to 1:

$$\mu(X) = 1. \tag{9.3}$$

Such a measure is called a **probability measure**, and reflects the fact that the probability of the whole sample space is equal to 1. Finally, $\mathcal{B}$ denotes the measurable subsets of X. Unfortunately, it is not possible to extend such a function to all subsets of X (Theorem of Ulam). One has to restrict the measure to the set of "constructible subsets". We first choose some elementary sets $E \in \mathcal{E}$ like intervals, rectangles, cubes, cylinder sets or similar. Then we apply a countable number of the operations *union* and *complement*:

$$E_1, E_2 \in \mathcal{E} \Longrightarrow E_1 \cup E_2 \in \mathcal{B}$$

and

$$E_1 \in \mathcal{E} \Longrightarrow X \setminus E_1 \in \mathcal{B}.$$

Moreover, at "each time we arrive at new sets" we allow countable unions and intersections of the new built sets. This way we consider the smallest collection of subsets of X that is closed under countable unions and intersections. In other words, we admit countable unions and countable intersections in $\mathcal{B}$ and for all new sets. In case our elementary sets are connected to the topology of X (like open sets), the set $\mathcal{B}$ is called the **Borel-σ-algebra** of X. It is the basic idea of measure theory to construct measures on such σ-algebras.

REMARK 9.1. In a compact metric space we always have a countable base of the topology consisting of balls. We can choose them as our elementary sets. We will from now on assume that **the cardinality of the elementary sets is countable.**

REMARK 9.2. The Lebesgue measure on $[0, 1]$ has the fundamental property that the measure of any finite interval $I = (a, b)$ with $0 \leq a < b \leq 1$ is $m(I) = b - a$. It is an example of a probability measure, where $\mathcal{E}$ is the collection of all open intervals. Lebesgue measure has the additional property that all subsets of a set of zero measure are measurable (and have also measure zero).

Example 9.2. The δ-measure on $[0,1]$ is another example:

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise.} \end{cases}$$

We want to remark that for this measure any set is measurable, i.e. we can choose $\mathcal{B} = 2^X$. This is not possible in general, in particular for Lebesgue measure!

9.1.1. Integration and Linear Functionals. From physics we can get the motivation that integration is the process of evaluating the average of an observable (a function). Already the Mean Value Theorem of integral calculus conveys this intuition. We will "borrow" this motivation to develop our ideas. An average of a function should be

1. 1 for the observable $\phi \equiv 1$,
2. positive for positive observables
3. linear in the space of all observables and
4. continuously dependent on the observables.

If we assume the observables are represented by elements from the space of continuous functions then we arrive to the idea of *integration as a positive linear continuous functional L of norm 1.* As in the naive measure theory we can extend the space of **potentially integrable functions $\mathcal{I}$**. For this we allow the following operations

$$C^0 \subset \mathcal{I}$$

and

$$\phi_n \in \mathcal{I} \Longrightarrow \phi(x) = \lim_{n\to\infty} \phi_n(x) \in \mathcal{I}. \tag{9.4}$$

We note that the functional does not need to be finite! Therefore, this is not the usual notion of integrable functions.

There is a 1–to–1 correspondence between measures and integrations (this is the message of **Riesz representation Theorem**[**KH96**, A.2.6 p.713]). For a Borel set A we have

$$\chi_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise.} \end{cases} \in \mathcal{I}$$

(this is called the **characteristic function** of the set A) and we can define

$$\mu(A) = L_\mu(\chi_A) := \int_X \chi_a(x)\, d\mu(x).$$

On the other hand, one can prove that the operations (9.4) applied to the linear combinations of the characteristic functions χ_A recover the whole set $\mathcal{I}$. Therefore it suffices to define the linear functional just on the set of characteristic functions.

$$L_\mu(\chi_A) := \mu(A).$$

This gives the notion of an integral with respect to an arbitrary measure:

$$\int_X \phi \, d\mu := L_\mu(\phi).$$

9.2. Invariant Measures

In the situation of a dynamical system we are interested in asymptotic statistics. So a measure which is concentrated on a set that is wandering (i.e., its iterates never return: $f^n(A) \cap A = \varnothing$) does not resemble the asymptotics. Therefore we restrict our attention to invariant measures (invariant under the action of f):

$$\mu(f^{-1}(A)) = \mu(A) \quad \forall A \in \mathcal{B}. \tag{9.5}$$

This is connected to the following integral equation

$$\begin{aligned}\int_{f^{-1}(A)} d\mu(x) &= \int_X \chi_{f^{-1}(A)}(x)\, d\mu(x) = \int_X \chi_A(f(x))\, d\mu(x) \\ &= \int_X \chi_A(y)\, d\mu(f^{-1}(y)) = \int_A d\mu(x).\end{aligned}$$

9.2.1. The Empirical Measure.

Assume we have fixed an initial value x_0. Then for a fixed elementary set E we can consider the **empirical frequencies**

$$\mathrm{Freq}_N(x_0, E) := \frac{1}{N} \# \{0 \le n < N \,:\, f^n(x_0) \in E\},$$

i.e., the visiting frequencies to the set E of x_0 and its iterates. In our situation we can assume that we have a countable set of elementary sets (e.g., rectangles with rational endpoints, cylinder sets, etc.) which generate $\mathcal{B}$. Then we can find by "diagonalisation" a sub-sequence N_k such that

$$\exists \lim_{k\to\infty} \frac{1}{N_k} \mathrm{Freq}_{N_k}(x_0, E) =: \mu_{emp}(E)$$

for all elementary sets E[1]. This function can be extended to a measure – called an **empirical measure** – on all $\mathcal{B}$. It has the property that

$$\begin{aligned}
\mu_{emp}(f^{-1}(E)) &= \lim_{k\to\infty} \frac{1}{N_k} \operatorname{Freq}_{N_k}(x_0, f^{-1}(E)) \\
&= \lim_{k\to\infty} \frac{1}{N_k} \operatorname{Freq}_{N_k}(f(x_0), E) \\
&= \lim_{k\to\infty} \frac{1}{N_k} \operatorname{Freq}_{N_k-1}(x_0, E) \pm \frac{1}{N_k} \\
&= \mu_{emp}(E)
\end{aligned}$$

for all elementary sets E. Hence an empirical measure is invariant.

Remark 9.3. The main difficulty is that an empirical measure is not unique. It may depend on x_0 and on the sequence N_k. For the doubling map E_2 we have for all N that

$$\operatorname{Freq}_N(0, E) = \begin{cases} 1 & \text{if } 0 \in E \\ 0 & \text{otherwise} \end{cases} = \delta_0(E).$$

For any "typical"point x_0 we have for the interval defined by the cylinder set $I = \{y \in [0,1) \ : y_1 = 0\}$ (an interval of length 2^{-1}, all points in the unit interval having zero as first element in its binary expansion)

$$\lim_{N\to\infty} \frac{1}{N} \operatorname{Freq}_N(x_0, I) = \lim_{N\to\infty} \frac{1}{N} \#\{0 \le n < N \ : \ x_n = 0\} = \frac{1}{2}.$$

Consider now a point x_1 with $x_{(2n)!} = 0, \cdots x_{(2n+1)!-1} = 0$ and $x_{(2n-1)!} = 1, \cdots x_{(2n)!-1} = 1$ for all $n \in \mathbb{N}$. Then

$$\lim_{k\to\infty} \frac{1}{(2k)!-1} \operatorname{Freq}_{(2k)!-1}(x_1, I) \ge 1 - \left(\lim_{k\to\infty} \frac{1}{(2k)!-1} (2k-2)! \right) = 1$$

and

$$\lim_{k\to\infty} \frac{1}{(2k+1)!-1} \operatorname{Freq}_{(2k+1)!-1}(x_1, I) \le \left(\lim_{k\to\infty} \frac{1}{(2k+1)!-1} (2k)! \right) = 0.$$

Hence, for this point the empirical measure is not unique (it depends on the sub-sequence N_k).

9.2.2. Invariant Measures for One–dimensional Maps. Consider a map on the interval with the following property: There are n disjoint intervals $(I_i)_{i=1}^n$; $\bigcup \overline{I_i} = [0,1]$ and f is linear on each of these intervals. Moreover, the image of each of the intervals is the union of some of those intervals

$$f(I_i) = I_{j_1} \cup \cdots \cup I_{j_{k(i)}}.$$

[1] Freq_N is a number between zero and one so the sequence has a convergent sub-sequence, in fact any set E in the countable family of elementary sets has a convergent sub-sequence. The idea is to pick a sub-sequence that is convergent for all sets in the family.

Such maps are called linear **Markov maps**. We can associate to this map a **transition matrix** $A = (a_{ij})_{1\le i,j\le n}$ with

$$a_{ij} = \begin{cases} 1 & \text{if } f(I_i) \supset I_j \\ 0 & \text{otherwise.} \end{cases}$$

The transition matrix for the doubling map E_2 (see Example 9.1) is

$$A_{E_2} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

The map g drawn in Figure 9.2 is also a linear Markov map. Its transition matrix is

$$A_g = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \end{pmatrix}.$$

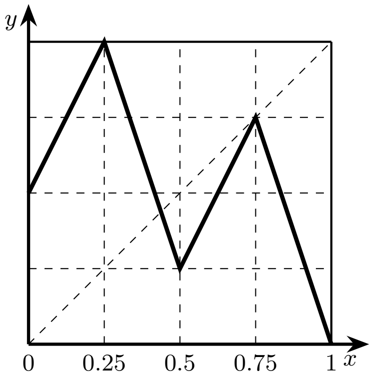


FIGURE 9.2. The linear Markov map g.

The transition matrix describes existing orbits. If the entry $a_{ij} = 1$ then at each time–step n there is a point with $f^n(x) \in I_i$ and $f^{n+1}(x) \in I_j$. More generally, the matrix $A^k = (a_{ij}^{(k)})$ gives all possible transitions within k time–steps. Not all transitions have to be equally likely. We can assign probabilities p_{ij} to each possible transition. Clearly

$$\sum_{j=1}^{n} p_{ij} = 1 \tag{9.6}$$

since the total probability of transitioning is 1. Also $p_{ij} = 0$ if $a_{ij} = 0$.

The new matrix $P = \{p_{ij}\}$ is called a **stochastic matrix**. If there is a positive integer M such that P^M has strictly positive entries this matrix is called irreducible. Then equation (9.6) and the Perron-Frobenius Theorem A.13 (see Appendix A.4.5) assure that 1 is a simple eigenvalue and the corresponding left and right eigenvectors are positive. The easy part of the result goes by noting that since all rows of P add up to one, the column vector with all entries equal to unity is a right eigenvector of P with eigenvalue one. Let $p = (p_1, \cdots, p_n)$ be the normalised left eigenvector, i.e., $\sum p_i = 1$ and $pP = p$. We can now construct an invariant measure on $[0, 1)$. Let

$$I_{i_1 \cdots i_k} = I_{i_1} \cap f^{-1} I_{i_2} \cap \cdots \cap f^{-(k-1)}(I_{i_k})$$

then we set

$$\mu_P \left(I_{i_1 \cdots i_k}\right) = p_{i_1} \cdot p_{i_1 i_2} \cdots p_{i_{k-1} i_k}.$$

This defines a function on the elementary intervals and can be extended to $\mathcal{B}$. For this we only have to show that the measure such defined has the invariance property on the elementary intervals.

Note first that the intervals can be recast in the cylinder notation of the previous chapter: $I_{i_1 \cdots i_k} \equiv [i_1 \cdots i_k]$ where the iterates are to be performed with f^{-1}. Verify first that since $[i_1 \cdots i_{n-1}] = \bigcup_{k=1}^{n} [i_1 \cdots i_{n-1} k]$ it holds that

$$\mu_P([i_1 \cdots i_{n-1}]) = \sum_{k=1}^{n} \mu_P([i_1 \cdots i_{n-1} k]).$$

Indeed,

$$\begin{aligned}
\mu_P([i_1 \cdots i_{n-1}]) &= p_{i_1} p_{i_1 i_2} \cdots p_{i_{n-2} i_{n-1}} \\
&= p_{i_1} p_{i_1 i_2} \cdots p_{i_{n-2} i_{n-1}} \left(\sum_{k=1}^{n} p_{i_{n-1} k} \right) \\
&= \sum_{k=1}^{n} p_{i_1} p_{i_1 i_2} \cdots p_{i_{n-1} k} \\
&= \sum_{k=1}^{n} \mu_P([i_1 \cdots i_{n-1} k]).
\end{aligned}$$

Now we will check that since p is an eigenvector this measure is invariant. Note that

$$(p_1, \cdots, p_n) \cdot P = \left(\sum_{k=1}^{n} p_k p_{k1}, \cdots, \sum_{k=1}^{n} p_k p_{kn} \right).$$

Hence,

$$\begin{aligned}
\mu_p\left(f^{-1}([i_1\cdots i_m])\right) &= \mu_p\left(\bigcup_{k=1}^{n}[ki_1\cdots i_m]\right)\\
&= \sum_{k=1}^{n}\mu_p\left([ki_1\cdots i_m]\right)\\
&= \sum_{k=1}^{n}p_k p_{ki_1}p_{i_1i_2}\cdots p_{i_{m-1}i_m}\\
&= \left(\sum_{k=1}^{n}p_k p_{ki_1}\right)p_{i_1i_2}\cdots p_{i_{m-1}i_m}\\
&= p_{i_1}p_{i_1i_2}\cdots p_{i_{m-1}i_m}\\
&= \mu_p\left([i_1\cdots i_m]\right).
\end{aligned}$$

Finally we can extend this result to all cylinders and therefore to all measurable sets in $\mathcal{B}$. Such a measure is called a **Markov measure**.

In case we can choose

$$p_{ij} = \frac{|I_j|}{|I_i|\ |f'|_{I_i}|}$$

then this probability is the amount of the measure that is spread from I_i to I_j. One can show that the resulting measure has a density $\rho(x)$ with respect to the Lebesgue measure

$$\int_{[0,1)}\phi(x)\,d\mu(x) = \int_{[0,1)}\phi(x)\rho(x)\,dx$$

and the density is connected to the right eigenvector p^+ normalised so that that $<p,p^+>= 1$. For this we conclude (note that $I_{i_1\cdots i_m} = [i_1\cdots i_m]$)

$$\begin{aligned}
\frac{\mu_p\left([i_1\cdots i_m]\right)}{|I_{i_1\cdots i_m}|} &= \frac{p_{i_1}p_{i_1i_2}\cdots p_{i_{m-1}i_m}}{|I_{i_1\cdots i_m}|}\\
&= \frac{1}{|I_{i_1\cdots i_m}|}p_{i_1}\frac{|I_{i_2}|}{|I_{i_1}||f'|_{I_{i_1}}|}\frac{|I_{i_3}|}{|I_{i_2}||f'|_{I_{i_2}}|}\cdots\frac{|I_{i_m}|}{|I_{i_{m-1}}||f'|_{I_{i_{m-1}}}|}\\
&= \frac{1}{|I_{i_1\cdots i_m}|}p_{i_1}\frac{|I_{i_m}|}{|I_{i_1}||f(m)|_{I_{i_1}}|}\\
&= \frac{1}{|I_{i_1\cdots i_m}|}p_{i_1}\frac{|I_{i_m}|}{|I_{i_1}|}|I_{i_1\cdots i_m}| = p_{i_1}\frac{|I_{i_m}|}{|I_{i_1}|}
\end{aligned}$$

where the left-hand side is bounded away from zero and finite independently of $i_1\cdots i_m$ and hence a density.

Example 9.3. For the map E_2 we can choose

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2}\\ \frac{1}{2} & \frac{1}{2}\end{pmatrix}.$$

The resulting density is $p^+ = (1, 1)$ and the measure is Lebesgue measure itself.

Remark 9.4. In the case that

$$p_{ij} = p_{lj} = p_j \quad \text{for all } i, j, l$$

the probability to be at "state" j does not depend on where you come from. Such a measure is called a **Bernoulli measure**. The general formula is

$$\mu_P\left(I_{i_1 \cdots i_k}\right) = p_{i_1} \cdot p_{i_2} \cdots p_{i_k}.$$

since

$$(p_1, \cdots, p_n) \begin{pmatrix} p_1 & \cdots & p_n \\ & \cdots & \\ p_1 & \cdots & p_n \end{pmatrix}$$

$$= (p_1(p_1 + \cdots + p_n), \cdots, p_n(p_1 + \cdots + p_n))$$

$$= (p_1, \cdots, p_n).$$

The Lebesgue measure for E_2 is actually a Bernoulli measure.

9.3. The Birkhoff Ergodic Theorem

We will present here (without proof) the most important theorem on orbit statistics. It can be seen as a mathematical justification of the Boltzmann Ergodic Hypothesis. The general form of this theorem and its proof can be found in the Appendix (Theorem A.19). Before proceeding we need some preparation.

We are only interested in the basic stones of an invariant measure. If e.g., we would have an invariant set with positive measure then also the complement is invariant and we can investigate the set and its complement separately. therefore we define

Definition 9.1. An invariant measure μ is called **ergodic** if it is indecomposable, i.e., if for any invariant Borel set A (this is, a set such that $A = f(A) = f^{-1}(A)$)

$$\mu(A)\mu(X \setminus A) = 0.$$

Remark 9.5. We can give also another equivalent formulation: μ is ergodic if and only if any invariant measurable function ϕ is constant μ-a.e.[2], i.e., there is a number a such that $\phi(x) = a$ for a set of points of measure 1. This follows since the sub- (sup-) level sets $L_c^{+(-)} := \{x \in$

[2] A property is said to hold μ-a.e. (almost everywhere with respect to the measure μ) when the set of points where the property holds has measure 1 or, equivalently, when the set of points where it does not hold has measure zero.

$X : \phi(x) \leq (\geq) c\}$ are invariant sets for an invariant function ϕ and hence, have measure 0 or 1. Then $\phi(x) = \sup_c(\inf_c)\{\mu(L_c^{+(-)}) = 0\}$.

Remark 9.6. It can be shown that for an irreducible A the Markov measures constructed in the previous section are ergodic.

It turns out that in general any invariant probability measure μ defined on the Borel-σ-algebra of a compact metric space can be uniquely decomposed into ergodic components:

Theorem 9.1 (Ergodic Decomposition). *The space $\mathcal{P}_{inv}(X)$ of invariant probability measures on the Borel-σ-algebra of a compact metric space X is a convex compact subset of the locally convex linear space of all finite signed Borel measures on X. The extreme points of $\mathcal{P}_{inv}(X)$ are precisely the ergodic measures $\mathcal{P}_{erg}(X)$. For any $\mu \in \mathcal{P}_{inv}(X)$ there exists a unique probability measure $\mathbb{P}_\mu$ on $\mathcal{P}_{inv}(X)$ that is concentrated on $\mathcal{P}_{erg}(X)$ and such that for any set A in the Borel-σ-algebra,*

$$\mu(A) = \int_{\mathcal{P}_{inv}(X)} \nu \, d\mathbb{P}_\mu(\nu) = \int_{\mathcal{P}_{erg}(X)} \nu \, d\mathbb{P}_\mu(\nu).$$

For a proof, see Section A.11 in the Appendix.

We are now ready to state [**KH96**, 4.2.1 p.136][**Sar20**, p.35]

Theorem 9.2 (Birkhoff's Ergodic Theorem). [**Bir31**] *Let μ be an ergodic probability measure for $f\colon X \to X$, X . Then there is a set A with $\mu(A) = 1$ such that for any continuous function ϕ and for any $x \in A$ we have*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \phi(f^n(x)) = \int_X \phi(x)\, d\mu.$$

Hence, when the hypotheses of the Theorem are fulfilled, that is, when we have an ergodic measure μ, the BEH holds for almost every point of X (with the possible exception of a set $X \setminus A$ of zero measure).

Corollary 9.1. *For $x \in A$ the empirical measure exists and is independent of x and the sub-sequence and equals μ.*

Proof. Let E be an elementary set and χ_E its characteristic function. Such a function is not continuous in general, but it can be approximated by continuous functions differing from χ_E on a set of measure smaller than ϵ, for any $\epsilon > 0$. So we can produce a sound argument showing that the Theorem holds also for χ_E. Actually, the full Ergodic Theorem stated in the Appendix is proved for a still wider class of functions. Hence, applying the Theorem we have

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \chi_E(f^n(x)) = \int_X \chi_E(x)\, d\mu = \mu(E)$$

for all $x \in E$. □

Remark 9.7. The Birkhoff Ergodic Theorem gives a method to evaluate integrals. This method is the essence that is behind the Monte–Carlo–Method. Suppose we have a map f with associated ergodic measure μ and we want to evaluate the integral $\int_X \phi(x)\,d\mu$ for some suitable function ϕ. We may choose a typical initial condition x_0 (i.e., randomly) and generate a sequence of **"random numbers"** using the map, namely $f^n(x_0)$. Then we evaluate the function at these numbers and average these values. The Birkhoff Ergodic Theorem claims that this value tends to the integral as $n \to \infty$. Of course, we must assure ourselves first that $x_0 \in A$.

The following theorem is actually simpler then the Ergodic Theorem and was proved before not using the ergodic theorem.

Corollary 9.2 (Poincaré Recurrence Theorem). [**KH96**, 4.1.19 p.142] *Let $A \in \mathcal{B}$ and μ be an ergodic measure. Then μ-a.e. point $x \in A$ (i.e., the set of points for which the result does not hold is a subset of A of zero measure) returns infinitely often to A.*

Proof. If $\mu(A) = 0$ then the assertion is obvious. Let us assume then that $\mu(A) > 0$. By the Birkhoff Ergodic Theorem we know that for a.e. point of A the empirical measure of A is equal to the measure $\mu(A)$. Hence a.e. point must return infinitely often to obtain the right "hitting frequency" (if $f^n(x)$ would return only finitely many times to A then the limit on the left hand side of the Ergodic Theorem would be zero, while the right hand side is positive). □

9.4. Exercise

Exercise 9.1. Show using the properties of measures and this definition that invariant sets with respect to an invariant ergodic measure have either full measure (measure 1) or measure zero.

9.5. Lyapunov Exponents

In this Section we consider a map of the interval $f\colon [0,1) \to [0,1)$ with an ergodic measure μ. We choose the function $\phi(x) = \log|f'(x)|$ (if $f \in C^1$ then ϕ is continuous). The Birkhoff Ergodic Theorem gives

$$\begin{aligned}\lim_{N\to\infty} \frac{1}{N} \log\left|\left(f^N(x)\right)'\right| &= \lim_{N\to\infty} \frac{1}{N} \log\left|\prod_{n=0}^{N-1} f'(f^n(x))\right| \\ &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \phi(f^n(x)) \\ &= \int_{[0,1)} \log|f'(x)|\; d\mu = \lambda_\mu\end{aligned}$$

for μ-a.e. x. The number λ_μ is called the **Lyapunov exponent** of f with respect to μ. It gives the average expansion along the orbit.

If we consider higher-dimensional maps the derivative $D_x f$ becomes a matrix and the norm is not multiplicative. We have to estimate $\lim_{n\to\infty} \frac{1}{n}\|D_x f^n v\|$. We note that in general

$$\|D_x f^n v\| \neq \prod_{i=1}^{n} \|D_{f^{i-1}(x)} f\| \|v\|.$$

Therefore we cannot apply the Birkhoff Ergodic Theorem in such a situation. However, the Oseledec Theorem[**Ose68**] can handle this more complicated situation, ensuring the existence of n numbers λ_i which give the average contraction/expansion rates in different directions. For more recent proofs of this Theorem see [**Rue79, KM99**].

9.6. Exercises

EXERCISE 9.2. Let $R_\alpha(x) = x + \alpha \mod 1$ be a rotation with irrational α. What are the invariant measures? What changes if $\alpha \in \mathbb{Q}$?

Hint: For irrational α show that the hitting frequency

$$\lim_{N\to\infty} \frac{1}{N} \#\{0 \leq n < N \,:\, R_\alpha^n(x) \in (a,b)\}$$

does not depend on x (but only on the length of the interval). Argue that there can be only one invariant functional.

EXERCISE 9.3. Two (Borel) measures μ, ν are said to be singular iff there is a (Borel) set A such that $\mu(A) = 1$ and $\nu(A) = 0$. Use the Birkhoff Ergodic Theorem to show that two different ergodic measures are singular.
Hint: Use the fact that the two linear functionals associated to the measures are different.

EXERCISE 9.4. Consider

$$\operatorname{supp} \mu := \{x \in X \,:\, \mu(B_\epsilon(x)) > 0 \text{ for all } \epsilon > 0\},$$

i.e., the smallest closed set of full measure. Show that it is an invariant set for a continuous function $f \colon X \to X$ on the compact metric space X, provided μ is invariant.
Hint: Show that the symmetric difference

$$(\operatorname{supp} \mu \setminus f(\operatorname{supp} \mu)) \cup (f(\operatorname{supp} \mu) \setminus \operatorname{supp} \mu)$$

has zero measure for any invariant measure. Conclude from continuity that the support must be invariant.

EXERCISE 9.5. Let μ be an invariant measure for $f \colon X \to X$, where f is continuous on the compact metric space X. Show that $\mu\left(X \setminus \bigcup_{x\in X} \omega(x)\right) = 0$.

APPENDIX A

Appendix

A.1. Basic definitions

A.1.1. Spaces.

Definition A.1. A **topological space** is a pair (X, τ), where X is a set and τ is a collection of subsets of X, satisfying.

1. The empty set and X itself belong to τ.
2. Any arbitrary (finite or infinite) union of members of τ belongs to τ.
3. The intersection of any finite number of members of τ belongs to τ.

The elements of τ are called *open sets* and τ is called a topology on X.

On a topological space the concepts of connectedness, continuity and convergence can be defined. The most general topological spaces are such that these properties are established *without* using the concept of *distance*. Topological spaces are characterised by describing how their open sets look like.

In science applications the concept of distance is often natural. Hence, metric spaces are used, where the topological properties are defined and refined using the additional concept of distance between points. The usual Euclidean space (ordinary space) is an example of a metric space.

The historical order, however, is the opposite: Metric spaces came first and gradually their topological underlying structure was unveiled.

Since we are dealing mostly with sets equipped with a distance we avoid the more general treatment of general topological spaces and concentrate on metric spaces.

Definition A.2. A set or a topological space is called **compact** when all open covers contain a finite sub-cover. For Euclidean space, a set is compact if and only if it is closed and bounded.

Definition A.3. A set X is called a **metric space** if there is a function called **metric**

$$d\colon X \times X \to \mathbb{R}^+ \cup \{0\}$$

with properties

- reflexivity:
$$d(x,y)=0 \qquad \Longleftrightarrow \qquad x=y$$
- symmetry:
$$d(x,y)=d(y,x)$$
- triangle inequality
$$d(x,z)\leq d(x,y)+d(y,z).$$

Definition A.4. A subset $Y \subset X$ of a metric space X is called **open** if and only if for any point $y \in Y$ there is an $\epsilon > 0$ such that the ball $B_\epsilon(y) := \{x \in X \,:\, d(x,y) < \epsilon\} \subset Y$.The complements of open sets are called **closed**. In particular $X \setminus Y$ is closed.

Example A.1. We consider some metric spaces:

- Let X be any non–empty set. The metric
$$d(x,y)=\begin{cases}0 & x=y\\ 1 & x\neq y\end{cases}$$
is called the telephone metric.
- Let $X=\mathbb{R}^n$. There are many metrics on this space. E.g.,
$$d((x_1,\cdots,x_n),(y_1,\cdots y_n))=\sqrt{\sum_{k=1}^{n}(x_k-y_k)^2}$$
or
$$d((x_1,\cdots,x_n),(y_1,\cdots y_n))=\max_{1\leq k\leq n}|x_k-y_k|$$
- Let $X=C[a,b]$ the space of continuous functions on the interval $[a,b]$. A metric can be defined as
$$d(f,g)=\max_{x\in[a,b]}|f(x)-g(x)|.$$
Since continuous functions on an interval are integrable we can also consider a different metric
$$d(f,g)=\int_{[a,b]}|f(x)-g(x)|\,dx.$$
We note that $d(f,g)=0 \iff f\equiv g$ since a non-negative continuous function has zero integral if and only if it is identically zero.

Definition A.5. A sequence $(a_n)_n$ is called a **Cauchy sequence** if for any $\epsilon > 0$ there is a number $N = N(\epsilon)$ such that for all $n, m > N$ we have $d(a_n, a_m) < \epsilon$. Equivalently, a sequence is Cauchy if
$$\lim_{n\to\infty}\sup_{m>n} d(a_n,a_m)=0.$$

Remark A.1. Any convergent sequence of a metric space is a Cauchy sequence. But if we consider the rational numbers there are many Cauchy sequences that do not converge in $\mathbb{Q}$, namely those having as limit (in a broader sense) irrational numbers, like for example $((1+1/n)^n)_{n\in\mathbb{N}}$. This will mean that the space of rational numbers has "holes".

Definition A.6. A metric space is called **complete** if any Cauchy sequence converges.

Proposition A.1. *The space $C[a,b]$ of continuous functions on the bounded interval $[a,b]$ with the* max*-distance is complete.*

Proof. Recall first that continuous functions in closed and bounded intervals are bounded. Consider now a Cauchy sequence $f_n(x)$ of functions of the space. This sequence induces a Cauchy sequence of real numbers, for each point $x_0 \in [a,b]$. Indeed, if $d(f_n, f_m) < \epsilon$ then for any $x_0 \in [a,b]$ $d(f_n(x_0), f_m(x_0)) < \epsilon$. Since a closed and bounded interval of the real numbers is a complete metric space, $f_n(x)$ converges point-wise to some function $f(x)$. It remains to be shown that such a function is continuous. There are many ways to do this.

We show first that $f_n \to f$ not only point-wise, but also in the max-distance defined above. Indeed, for $n, m > N_0$ sufficiently large,

$$d(f_n, f) = \max_{x\in[a,b]} |f_n(x) - f(x)| = \max_{x\in[a,b]} \lim_{m\to\infty} |f_n(x) - f_m(x)| \leq \epsilon.$$

The last step holds since f_n is a Cauchy sequence of functions, hence $|f_n(x) - f_m(x)| \leq \epsilon$ for *all* $x \in [a,b]$ and n, m sufficiently large.

Now it is straightforward to show that $f(x)$ is continuous. Consider

$$|f(x) - f(y)| \leq |f(x) - f_n(x)| + |f_n(x) - f_n(y)| + |f_n(y) - f(y)|.$$

For n large enough the first and third terms in the sum above can be made to be smaller than $\epsilon/3$, because of the previous argument. Also, since f_n is continuous, there exists $\delta > 0$ such that $|x - y| \leq \delta \Rightarrow |f_n(x) - f_n(y)| \leq \epsilon/3$. Putting everything together we find that $|x - y| \leq \delta \Rightarrow |f(x) - f(y)| \leq \epsilon$. This can be done for any positive ϵ, and hence f is continuous. □

Remark A.2. Cauchy sequences are not preserved by a continuous change of coordinates (homeomorphism, see Definition A.10) or even a differentiable change of coordinates (diffeomorphism, see Definition A.10). The map

$$\tan\colon \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \to \mathbb{R} \quad x \to \tan x$$

is a diffeomorphism of the open interval $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ onto the real line. The sequence $a_n = \frac{\pi}{2} - \frac{1}{n}$ is a (non–convergent) Cauchy sequence in the interval while $b_n = \tan\left(\frac{\pi}{2} - \frac{1}{n}\right)$ is not a Cauchy sequence in $\mathbb{R}$.

Although $\mathbb{R}$ is a complete metric space, its diffeomorphic image $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ is not.

The next theorem says that a compact metric space is not too large.

Theorem A.1. [**Tay06**, p.253] *Any compact metric space has a countable dense set $\{x_l\}$. Moreover the countable family $(U_{\frac{1}{m}}(x_n))_{n,m}$ forms a basis of the topology, i.e., for all $x \in X$ and all open sets $V \ni x$ there is $n, m \in \mathbb{N}$ such that*

$$x \in U_{\frac{1}{m}}(x_n) \subset V.$$

Recall that a set is *countable* when its elements can be put into 1–to–1 correspondence with the natural numbers. Also, a set is dense (in the metric sense) if for any $\epsilon > 0$ and for any $x \in X$ there is an element x_k of the set such that $d(x, x_k) < \epsilon$. Denseness can be characterised without distance, e.g., a set D is dense if for all $x \in X$ every open set containing x (this is called a *neighbourhood* of x) contains at least one element of D.

Definition A.7. A n-dimensional manifold is a topological space (or a set), where each point has a neighbourhood equivalent to an open set of $\mathbb{R}^n$. Equivalent means here that there exists a 1–to–1 (and hence invertible) map between the neighbourhood and the open set in $\mathbb{R}^n$. Such a map is called a *chart* and the set of all charts is called an *atlas*. Whenever two charts have a common domain, the corresponding maps should be compatible (conjugate).

The overall structure of the manifold may be complicated, but in the vicinity of each point its structure is just like Euclidean space. A manifold may be topological, where the conjugation between the overlaps of the charts is given by continuous maps, differentiable (differentiable maps), smooth (infinitely differentiable maps) or analytic (analytic maps).

A.1.2. Banach Spaces and Norms. Whenever we can regard the elements of a metric space X as vectors (i.e., X is a linear space, which means that we can define the sum and multiples by a scalar of elements in X as a new element in X) with a natural compatible ”length” (norm) $\|x\|$, there is a natural association connecting the ideas of “distance between points” and “length of vectors”. Indeed, the length $\|x\|$ is the distance between x and the zero vector.

These ideas are formalised in the concept of **norm** and **normed space**. Whenever a norm is defined on a vector space, we can define a corresponding distance as $d(x_1, x_2) = \|x_1 - x_2\|$. Subsequently we can define open sets and a topology using this distance. However, the inverse procedure is not always possible, there exist topologies (i.e., specifications of open sets) that are not normalisable. We refer to standard textbooks in topology and linear spaces for the proper definitions.

Also $n \times n$ matrices can be regarded as vector spaces, under the usual sum of matrices and the zero matrix in the role of "zero vector" of the space. We may want to estimate the size of matrices using some sort of norm.

From Linear Algebra we know that matrices describe linear maps between vector spaces. It is quite natural to define matrix norms as the largest modification in size they can act on a non zero vector.

Definition A.8 (Matrix Norm). We define the norm of a square matrix A as:

$$\|A\| := \sup_{\|x\|=1} \|A \cdot x\|.$$

Note that the norm of a matrix is defined via norms of vectors.

Definition A.9 (Banach Space). A Banach Space is a complete normed vector space.

Hence, it is a vector space where we can add its elements. In addition, it is equipped with a norm and out of it a distance between points and finally every Cauchy sequence in the space is convergent. The usual Euclidean $\mathbb{R}^n$ with the usual norm and distance is of course a Banach space.

A.1.3. Gronwall's Inequality.

Lemma A.1 (Gronwall's Inequality). [**Ver90**, p.5],[**CL55**, p.37] *If, for $t_0 \leq t \leq t_1$, $\phi(t) \geq 0$ and $\psi(t) \geq 0$ are continuous functions such that the inequality*

$$\phi(t) \leq K + L \int_{t_0}^{t} \psi(s)\phi(s)ds$$

holds on $t_0 \leq t \leq t_1$, with K and L positive constants, then

$$\phi(t) \leq K \exp\left(L \int_{t_0}^{t} \psi(s)ds\right)$$

on $t_0 \leq t \leq t_1$.

Proof. The hypothesis is equivalent to

$$\frac{\phi(t)}{K + L\int_{t_0}^{t} \psi(s)\phi(s)ds} \leq 1$$

Multiplying by $L\psi(t)$ and integrating,

$$\int_{t_0}^{t} \frac{L\psi(s)\phi(s)ds}{K + L\int_{t_0}^{s} \psi(\tau)\phi(\tau)d\tau} \leq L \int_{t_0}^{t} \psi(s)ds.$$

Thus

$$\ln\left(K + L\int_{t_0}^{t} \psi(s)\phi(s)ds\right) - \ln K \leq L \int_{t_0}^{t} \psi(s)ds$$

and finally

$$\phi(t) \leq K + L\int_{t_0}^{t} \psi(s)\phi(s)ds \leq K \exp\left(L\int_{t_0}^{t}\psi(s)ds\right).$$

□

A.1.4. The Induction Principle. The induction principle [**Pea89**] is a method of proof used to establish that a given statement is true for all natural numbers. It proves, so to speak, an infinite family of statements at once. Let us illustrate it with the following example: Let us prove that for all integers $N \geq 1$, $\displaystyle\sum_{k=1}^{N} k = \frac{1}{2}N(N+1)$.

The method starts by proving (checking) that the first statement, for $N = 1$, holds. In our example, for $N = 1$ the statement is $1 = \frac{1}{2}{\cdot}1{\cdot}2$, that is clearly valid.

The next and final step is to prove that if a statement in the infinite sequence of statements is true, then so is the next one. In our example, $\displaystyle\sum_{k=1}^{K} k = \frac{1}{2}K(K+1) \Rightarrow \sum_{k=1}^{K+1} k = \frac{1}{2}(K+1)(K+2)$. Indeed, we have that $\displaystyle\sum_{k=1}^{K+1} k = \sum_{k=1}^{K} k + (K+1) = \frac{1}{2}K(K+1) + (K+1) = \frac{1}{2}(K+1)(K+2)$. This completes the proof.

The induction principle is not supposed to be "proved" itself within the standard number framework that we learn at school. It is part of the foundational axioms defining the natural numbers, as formalised by Peano in the late XIX-th century. This is because this concept is as natural to human intuition as the positive integers are. However, the induction principle is a fruitful object of study for mathematical logic, and it can be proved in some axiomatic systems other than Peano's.

A.1.5. Maps.

Definition A.10. A continuous 1–to–1 map of a topological space whose inverse is also continuous is called a **homeomorphism**. If the map and its inverse are differentiable it is called a **diffeomorphism**.

Remark A.3. Such maps may be considered as (continuous respectively differentiable) changes of coordinates. If we apply a homeomorphism or a diffeomorphism to a small ball around the origin in $\mathbb{R}^n$ we map the (Cartesian) straight coordinate lines to (not necessarily straight) new curves. Since this done in a 1–to–1 way, we can give to each point in the image its coordinates defined by the pre-image.

A.2. More Topological Concepts

A.2.1. Cantor Sets. [**Wig90**, p.583] Let us consider the following construction. In the interval $[0,1]$ delete the open sub-interval $\left(\frac{1}{3},\frac{2}{3}\right)$. There remain two disjoint closed intervals I_0 and I_2. These intervals correspond to the real numbers in $[0,1]$ that have a 0 or a 2 in the first position when expressed in base–3:

$$I_0 = \{x \in [0,1] \,:\, x = .0x_2x_3\cdots \text{ in base–3}\}$$

and

$$I_2 = \{x \in [0,1] \,:\, x = .2x_2x_3\cdots \text{ in base–3}\}.$$

We can now repeat the procedure on each of the new intervals, eliminating their central open sub-interval of width $1/3^2$, namely $(1/9,2/9)$ and $(7/9,8/9)$. The four remaining closed intervals contain now real numbers having 0's or 2's in the first two positions. Continuing inductively by repeating the "rescaled" procedure at each remaining sub-interval, i.e., deleting the central third of each sub-interval, we have at step n exactly 2^n intervals, each one of length $\left(\frac{1}{3}\right)^n$. They correspond to real numbers having 0's or 2's in the first n positions and can be labeled by sequences $i_1i_2\cdots i_n$ denoting the first n positions in the base-3 expression, where $i_k \in \{0,2\}$ and

$$I_{i_1i_2\cdots i_n} \subset I_{i_1i_2\cdots i_{n-1}}.$$

In other words, we have

$$I_{i_1i_2\cdots i_n} = \{x \in [0,1] \,:\, x = .i_1i_2\cdots i_nx_{n+1}\cdots \text{ in base–3}\}.$$

The *standard Cantor set* is the set

$$\mathcal{C} = \bigcap_{n\in\mathbb{N}} \bigcup_{i_1i_2\cdots i_n} I_{i_1i_2\cdots i_n}.$$

Thus, it is the set of all real numbers in $[0,1]$ which do not have a 1 in their base–3 expansion. A general **Cantor set** is the homeomorphic image of the standard Cantor set. For future use, let us call x_n the string with the first n digits of the base–3 expression of a point $x \in \mathcal{C}$.

Definition A.11. A subset A of a space X is called totally disconnected if for any two points $x \neq y \in A$ we can find two open sets of X, $U \ni x$ and $V \ni y$ such that $A = (V\cap A)\cup(U\cap A)$ and $U \cap V = \varnothing$.

Definition A.12. A set A is called **perfect** if it is equal to the set of its limit points. In other words, every point of A is the limit of some sequence of distinct points in A, i.e., for any $x \in A$ and any open $U \ni x$ there is a point $y \in A \cap U$ distinct from x. In other words A does not contain isolated points.

Theorem A.2. [**Wig90**, p.583] *Any Cantor set is compact, perfect and is totally disconnected. Moreover, these three properties characterise a Cantor set.*

Proof. We prove the Theorem for the standard Cantor set.

The Cantor set is compact since it is a closed and bounded subset of the real line. You may note that it is closed since its complement in $[0,1]$ is open (it is a countable union of disjoint open intervals).

All endpoints of the closed intervals used to build $\mathcal{C}$ belong to $\mathcal{C}$ (they consist of a finite string x_n followed by an infinite string of 0's or of 2's). To see that $\mathcal{C}$ is perfect we will show that all points in $\mathcal{C}$ are accumulation points of elements of $\mathcal{C}$, namely that in any neighbourhood of $x \in \mathcal{C}$ there exists some other point $y \in \mathcal{C}$. Indeed, for any $1 > \epsilon > 0$ consider the open interval of radius ϵ around $x \in \mathcal{C}$ (any neighbourhood of x contains one such interval). Let $n > -\log_3 \epsilon$. Then the points $x_n 00\cdots$ and $x_n 22\cdots$ belong to the Cantor set, they are distinct and they lie closer to x than ϵ.

Since $\mathcal{C}$ is perfect it contains infinitely many points. We choose two: $x \neq y$. Since the points are distinct, they have different base–3 expansions, i.e., their expansions coincide up to some N but the digit $N+1$ is (without loss of generality) 0 for x and 2 for y. Consider now the closed interval $I_{x_N} \subset [0,1]$, where x_N is the common initial string of 0's and 2's in x and y. Both points belong to this interval. In the next step of the Cantor construction the open third of this interval is eliminated, namely all points of the form $0.x_N 1p$ (in base–3), where p is any string built from $\{0,1,2\}$. In particular, $x_* = 0, x_N 1100\cdots \notin \mathcal{C}$ and $x < x_* < y$. The sets $U = [0, x_*)$ and $V = (x_*, 1]$ satisfy the requirements of the definition (note that regarded as subsets of the unit interval, they are *open*). □

Let us consider some standard and less standard properties of the Cantor set.

Exercise A.1. Compute the Lebesgue measure (total length) of the eliminated intervals in the construction of the Standard Cantor set. Is the Cantor set measurable? What is its measure?

Exercise A.2. Try a Cantor-like construction by eliminating from the unit interval at each step a central open interval whose length is a fraction $p < 1$ of the closed interval in question. In other words, in the first step we eliminate from the unit interval of length $L_0 = 1$, the central portion of length pL_0. Two intervals of length $L_1 = (1-p)/2$ are left. In the second step we eliminate the central intervals of length pL_1, and so on. Is there a limit set? How does it look like? Compute the Lebesgue measure of the eliminated intervals and that of the limit set.

Exercise A.3. Let $p = 1/k$, where $k > 3$ is an integer. Try a Cantor-like construction by eliminating from the unit interval at each step $n \geq 0$ a central open interval whose length is p^{n+1}. Is there a limit set? How does it look like? Compute the Lebesgue measure of the eliminated intervals and that of the limit set.

A.2.2. On the Concept of Infinity. This subsection is intended as an informative addendum to previous discussions. In these days, however, books for a general audience can be successfully complemented by "googling" through the internet or inside Wikipedia (at your own risk, of course), using –for the present purpose– keywords such as *Axiom of Choice, Cantor, Well-ordering theorem, Zorn, Zermelo, Baire, Tarski, Hilbert, Constructivism*, etc.

Assumption: Throughout the book we assume the *Axiom of Choice (AC)* (stated below as Axiom A.1) to be valid. Some of the statements below rest on (AC) or some weaker form of this axiom.

DEFINITION A.13. We call two sets *equipotent* if there exists a bijective function mapping one set onto the other (element-wise). We say that a set is *infinite* if it is equipotent with a proper subset of itself.

EXERCISE A.4. Prove the following statements.

1. $\mathbb{Z}$ and $\mathbb{R}$ are infinite sets.
2. $\mathbb{Q}$ is equipotent to $\mathbb{Z}$.
3. There is no surjective function from $\mathbb{Q}$ (or $\mathbb{Z}$) to $\mathbb{R}$, i.e., $\mathbb{R}$ and $\mathbb{Q}$ are not equipotent.

EXERCISE A.5. Show that the standard Cantor set is equipotent to the real interval $[0, 1]$.

This definition of infinite set contributed to start one of the most amazing developments in mathematics around the turn of the XIX–XX-th centuries. In fact, the idea of infinity was present already in traditional analysis: the concept of *limit*, *accumulation point*, *continuous function*, or just *irrational number*, all involve some way or the other the "construction" of e.g., *infinite sequences* (existing at least within the universe of our imagination), as well as mathematical assertions about these sequences. Let us reason about the implications of the newer views on infinity. Before Cantor, the usual definition of infinity rested on counting:

DEFINITION A.14. A set is *finite*$_S$ (S for strongly) if it has n elements, for some non-negative integer n, i.e. there is a bijection to the set $\{1, \cdots, n\}$, whereas a set is *infinite*$_S$ if it is not *finite*$_S$.

The only set with zero elements is the empty set. A corollary of this definition is that a finite set with $n > 0$ elements can be put in 1–to–1 correspondence (i.e., bijectively) with the finite set $\{1, 2, \cdots, n\}$ of positive integers (this is just *counting*, the naive way). Now we restate the second half of Definition A.13 as:

DEFINITION A.15. A set A is *infinite*$_D$ (D for Dedekind) if there exists a proper subset B of A and a bijective function $f : A \mapsto B$, whereas a set is *finite*$_D$ if it is not *infinite*$_D$.

Again, the empty set is finite$_D$ since it has no proper subsets. However, it turns out that it is a non-trivial issue to decide wether Dedekind's concept of infinity is different from the naive one or not:

LEMMA A.2. *If a set X is infinite$_D$ then it is infinite$_S$.*

PROOF. The empty set and any set without proper nonempty subsets are ruled out since they are finite$_D$. Assume now that X is both infinite$_D$ and finite$_S$. Since it is finite$_S$, it has n elements for some positive n. Hence, any proper nonempty subset of X has $m < n$ elements for some positive m. Since X is infinite$_D$ then by composition of bijective functions there should exist a bijective function between the finite sets of integers $\{1, \cdots, n\}$ and $\{1, \cdots, m\}$. Hence, we arrive at a contradiction and infinite$_D$ ⇒ infinite$_S$. □

REMARK A.4. The question whether infinite$_S$ ⇒ infinite$_D$ or not (and hence whether equivalence between both concepts of infinity holds) is undecidable without an additional assumption beyond naive set theory. We will see in the next lines that such an assumption has unexpected consequences.

A.2.2.1. *Some Naive Comments on the Axiom of Choice.* [**Zer04**]

The next development spawned by Cantor's approach is exemplified by the third statement in Exercise A.4, namely that there exist infinite sets which are not pairwise equipotent. In particular, the *power set* of a set A (the set consisting of all subsets of A) is not equipotent with A. This statement is easy to prove for finite$_S$ sets without any additional assumption, since the power set of a finite set with n elements has exactly 2^n elements. A way to produce a proof that holds also for all sets includes Cantor's famous diagonal argument (this is the path to solve Exercise A.4.3), proving that the set of real numbers in $[0, 1]$ is not *denumerable* or not *countable*, i.e., not equipotent to the rationals (there are other proofs of the statement as well). Comparing with finite sets, we say that the power set of A is "larger" than A and that the set or real numbers is "larger" than the set of rationals. Having now infinite sets that are not equipotent with each other, the question arises whether these "infinities of different size" can be somehow organised or classified. To answer this question also requires the above mentioned additional assumption. Cantor made this assumption more or less automatically to start with. The idea was recognised by Peano and Bolzano among others and finally formulated clearly for the first time by Zermelo in 1904. In one of its forms it is called the *Axiom of Choice*:

AXIOM A.1. *For* **any** *collection of nonempty sets, there exists a (choice) function on the collection assigning to each set one of its elements.*

The "any" is important here since for certain collections, be them finite or infinite, it is more or less easy to produce choice functions without additional assumptions. It is enough to have an unambiguous rule that always selects an identifiable element. The Axiom of Choice is necessary to select a set from an infinite number of pairs of socks, but not from an infinite number of pairs of shoes (Bertrand Russell). No unambiguous rule can distinguish the socks in a pair, but shoes are different. The rule "choose always the left shoe" is unambiguous.

In fact, since the concept of choice seems to be natural for human beings and apparently also for other living entities, and since it is also provable in the mathematical sense for finite sets, it seems natural to assume that it holds for arbitrary collections. However, this requires a leap of faith. If we accept the axiom, this means that we believe that such a choice will always be possible, even in the cases where we cannot produce an explicit choice function or where it will never be produced because there is no available rule. To illustrate these ideas, ignore the Axiom for a moment and consider the following choices:

1. On the infinite collection of 2-element sets of the form $\{n, n+1\}$, for all non-negative integers n, define the choice function as that picking the odd number of each 2-element set (cf. Bertrand Russell's statement above).
2. Consider the equivalence relation defined on $[0, 1]$ as: $a \sim b$ if $a - b \in \mathbb{Q}$. All elements on the unit interval belong to just one equivalence class and the union of all equivalence classes coincides with the unit interval. We want a choice function on the collection of equivalence classes that picks one element out of each equivalence class.

The first choice can be done without difficulties. We can produce the result of applying this function, namely the set of odd positive integers (in every other set in the collection they appear as the first element, in the others as the second). We can even "list" this result in a complete and comprehensive way. The second choice is trickier. We may believe it can be done, but we have no rule to produce an example. We cannot prove that it is an impossible task either. However, accepting the Axiom of Choice we can prove that it can be done. Still, we will never be able to produce a complete explicit example of it. We know that e.g., 0, $\frac{\sqrt{2}}{2}$ and $\frac{\pi}{4}$ belong to different classes, so one choice function could start with these numbers, but we cannot produce the "full list" or a general rule to pinpoint any function value at will. We have difficulties already in explicitly organising the different equivalence classes, except for a few notorious ones.

There are a number of familiar and less familiar results whose proof invokes the Axiom of Choice (for some of the examples a weaker alternative is enough). Among them we have:

1. Given an infinite set E of (distinct) real numbers in $[0, 1]$ the following two statements are equivalent: (a) x is an *accumulation point* of E; (b) There exists a sequence $\{x_n\}_{n=1}^{\infty}$ of distinct elements of E such that $\lim_{n\to\infty} x_n = x$.
2. Every infinite set has a denumerable subset (also: infinite$_N \Rightarrow$ infinite$_C$).
3. Every vector space has a basis (Hamel Basis Theorem).
4. The generalised Cartesian product of a nonempty family of nonempty sets is nonempty (the Multiplicative Axiom).
5. A non-empty complete metric space is not the countable union of nowhere-dense sets (i.e., sets whose closure has dense complement). This is one form of Baire's Category Theorem.
6. There exists sets on the unit interval that are not measurable in the Lebesgue sense. The usual example is the *Vitali set*, defined using the second choice function exemplified above.
7. All sets can be well-ordered (also called Zermelo's Theorem).
8. The Induction principle can be extended to all sets (Transfinite Induction), in particular to Cantor's hierarchies of infinites by choosing a well-ordering.
9. There exist ways to partition the unit ball in $\mathbb{R}^3$ into a finite number of pieces and reassemble the pieces via rigid motions in such a way that a new sphere of any size, or even two unit spheres, are produced (Banach-Tarski Paradox, generalising an idea of Hausdorff).

This list is ordered in a non-rigorous way from "very natural statements" to "highly unnatural statements". The acceptance of the Axiom of Choice allows to prove that all these statements are correct.

The first one is particularly interesting. With elementary methods it can be proven that (b)⇒(a). However, to prove the converse (try it!) it is necessary at some point to pick an arbitrary element of an infinite and highly unspecified subset of E. The need of accepting the possibility of performing an arbitrary and unspecified choice is "central" to the proof. We cannot proceed without it. The interesting fact about this example is that both statements are encountered in *elementary* mathematics[1].

Less "drastic" versions of the Axiom of Choice such as the Axiom of Countable Choice (allowing Choice within countable collections) or the Axiom of Dependent Choice (which guarantees the existence of a countable choice set $\{a_n\}$ where a_{n+1} depends on a_n) are possible generalisations of finite choice that avoid some of the most puzzling consequences of Axiom A.1, such as e.g., points 6. or 9. in the list above. For a nice discussion, see ch 13 in [**Wag85**].

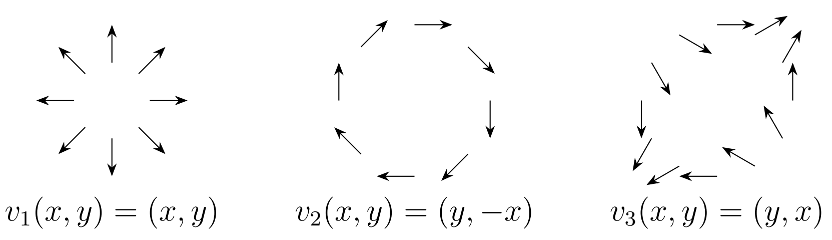
[1] However, some apparently harmless restrictions on the set E are enough to prove equivalence *within this restricted environment*, e.g., if we know that E has only one accumulation point.

The bottom line to take home is that the moment we start dealing with infinity (which in mathematics occurs quite frequently) we have to analyse carefully in which way the (familiar) results from finite mathematics can be extended to infinite sets. The question of accepting/rejecting the Axiom of Choice is usually not a central problem for the needs of applied mathematics, since many interesting and useful results can be produced without taking decisions governed by this Axiom. However, mathematicians dealing with foundational issues, set theory, etc., do have a lot of trouble with it. Lebesgue called it a *false statement* while simultaneously Fréchet claimed it was a *well-known* fact. Eventually, different schools of mathematics appeared, accepting or rejecting the Axiom (or some related statement) and its consequences.

A.2.3. The Index of a Vector–field. Let $v\colon \mathbb{R}^2 \to \mathbb{R}^2$ be a continuous vector field with a finite number of singular points. Let x be a singular point and C be a curve going exactly once (clockwise) around the the singularity x. We also assume that there is no other singularity than x inside and on the curve C. Let us consider the vectors $v(y)$; $y \in C$. We are interested in the number of complete revolutions this vector makes while we go once around the contour C. We count this number with its orientation: $+$ if the vector rotates clockwise and $-$ if it rotates counterclockwise. The reader may notice that this number does not depend on the starting point $y_0 \in C$. This number is called the **index** of the vector field at the singularity x.

If we "perturb" the curve a little bit (i.e., we make a small modification of the path of the curve without self intersections), the vectors $v(y)$ along the path will change continuously since we changed y continuously. The rotation changes therefore continuously. Hence, the index depends continuously on C as long as we do not cross another singularity. Since the index is an integer, it must remain constant as long as we do not meet another singularity.

Example A.2. Let $v_1(x,y) = (x,y)$, $v_2(x,y) = (y,-x)$ and $v_3(x,y) = (y,x)$ be three vector fields (see Figure A.1).



$v_1(x,y) = (x,y)$ $v_2(x,y) = (y,-x)$ $v_3(x,y) = (y,x)$

Figure A.1. Three planar vector fields with a singularity at the origin.

These vector fields have the following indices at the origin:

$$ind_{(0,0)}(v_1) = 1 \quad ind_{(0,0)}(v_2) = 1 \quad ind_{(0,0)}(v_3) = -1.$$

The vector fields in Figure A.2 have indices

$$ind_{(0,0)}(v_4) = 2 \quad ind_{(0,0)}(v_5) = 3 \quad ind_{(0,0)}(v_6) = -3.$$

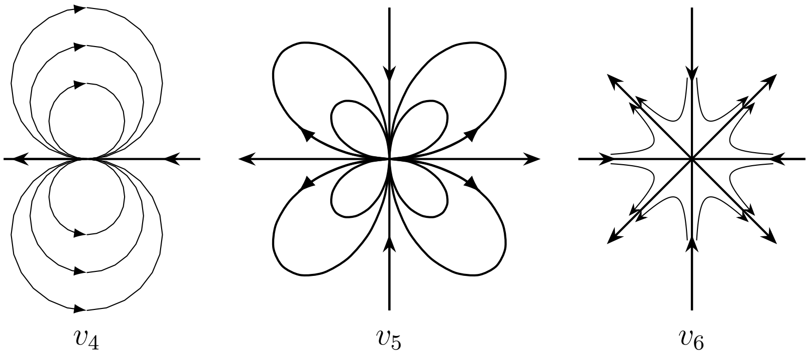


FIGURE A.2. More planar vector fields with a singularity at the origin.

Similarly, we can define the **index** ind_v of a non self–intersecting closed curve which does not pass through a singular point as the number of rotations (with orientation) that the vector $v(y)$ does while going around the curve. This index is also called the *winding number* of the vector field along the curve.

THEOREM A.3. [**GH86**, 1.8.4 p.51] *For a non self–intersecting closed curve C we have*

$$ind_v(C) = \sum_{x_i \text{ singular inside } C} ind_{x_i} v.$$

PROOF. Let us consider the curve C_1 which we obtain by continuous deformation from C as in Figure A.3.

C_1 has the same index as C since it does not pass through any singular point along the deformation. Then the straight segments of C_1 "cancel each other" since we move in opposite directions and they can be placed arbitrarily close to each other. By continuity of the vector field, whatever modification of v along a segment will be compensated by its companion segment. The curved arcs around each fixed point add up exactly to the sum of the indices. □

Calling the components of the vector field $v = (f_1, f_2)$, the index of a curve C can be computed as a line integral along the curve:

$$ind_v(C) = \frac{1}{2\pi} \int_C \frac{f_1 df_2 - f_2 df_1}{f_1^2 + f_2^2},$$

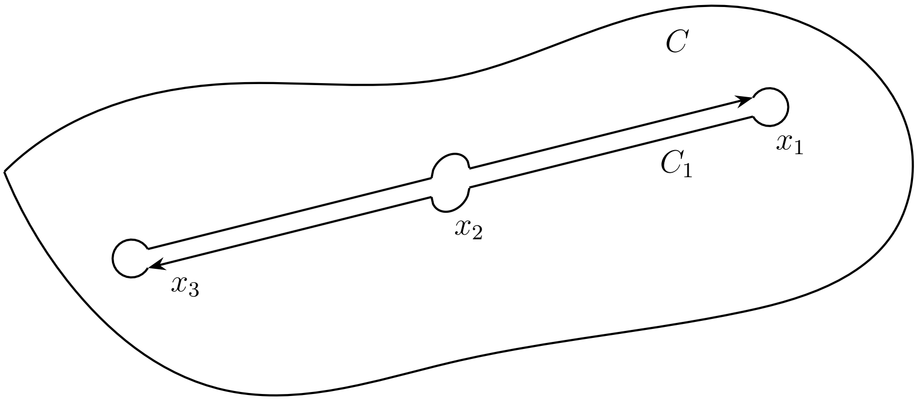


FIGURE A.3. Two curves with the same index.

since it is the total angular variation of the vector field along the curve. This relates to Green's Formula and contour integrals on the complex plane.

COROLLARY A.1. *If the index of a closed non self–intersecting curve in the plane is non–zero, the curve encloses at least one singular point.*

The "negative" reformulation of the previous Corollary reads:

COROLLARY A.2. *If a closed non self–intersecting curve in the plane does not encompass any singular point, its index is zero.*

This last result is intuitive for the case of sufficiently small curves. Then by continuity the vector field is almost constant along the curve (Straightening-out Theorem) and its total variation is zero.

THEOREM A.4. [**GH86**, 1.8.4 p.51] *The index of a periodic orbit of a planar vector field is* $+1$.

PROOF. The vector field is tangent to the orbit. When circulating along the orbit in the direction of the flow, it makes one complete revolution. □

COROLLARY A.3. *A periodic orbit of a planar vector field encompasses at least one singular point.*

A.2.3.1. *Flows on the Sphere.* We abandon for a moment the Euclidean plane and consider another 2-d surface, namely the sphere. The next Theorem might be known to the reader in terms of the Euler characteristics of the sphere $\mathbb{S}^2 \in \mathbb{R}^3$ (we will proceed along these lines below). We will interpret it in terms of vector fields on the sphere (i.e., at each point x of the sphere there is a vector $v(x) \in \mathbb{R}^3$ tangent to $\mathbb{S}^2$).

Theorem A.5. [**Arn73**, p.261] *Any continuous vector field on the sphere with finitely many singular points has sum of the indices equal to* 2*.*

Proof. We state without proof that the index of a curve or singular point is invariant in front of diffeomorphisms (this permits the computation of indices on other 2-d manifolds). We consider a nonsingular point x on the sphere. This point has a small neighbourhood $U(x)$ where all the vectors are almost the same as $v(x)$ (see Figure A.4, left). So we can choose the boundary of U as our contour $C = \partial U(x)$. The vector field has its singularities outside $U(x)$ and its boundary.

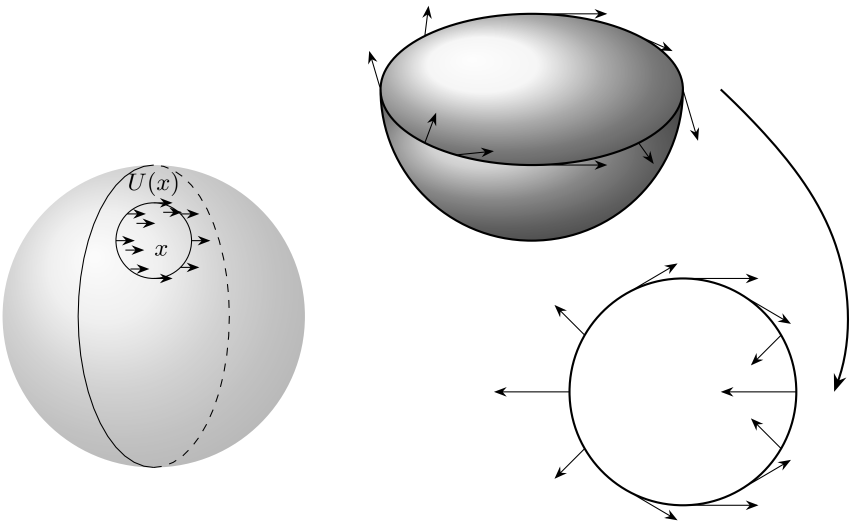


Figure A.4. A vector field on the sphere.

Consider now the stereographic projection mapping the complement of $U(x)$ to a plane, placed in such a way that x is the "north pole":

$$St(\theta, \phi) = \frac{2\sin\theta}{1-\cos\theta}(\cos\phi, \sin\phi).$$

In particular, the south pole $(\pi, 0)$ maps to the origin, $(0, 0)$. The map "flattens out" the complement of $U(x)$ to the interior region defined by the large contour $St(C)$. All (eventual) singularities of the original vector field on the sphere lie now inside this region. The image of the (essentially constant) original vector field along C is depicted on Figure A.4, right. The index of this curve 2, and it corresponds to the sum of indices of the singular points of the original field. □

The reader may wonder how the curve C encompassing no singular points on the sphere could be mapped to a new curve with index 2. The stereographic projection does not map the whole sphere diffeomorphically on the plane. One point, or a small neighbourhood, is missing (note that the map St above is not defined for $\theta = 0$). Moreover, the interior region defined by C on the sphere became the exterior on the plane. The sum of indices on the sphere is then $0 + 2$, the first term coming from the region inside C and the other from the rest of the sphere (computed after mapping it diffeomorphically to a planar region by St). Note that the theorem poses no specific condition on the vector field: The value 2 is a geometric property of the sphere, not of the dynamics.

In this way one may define the *Index at Infinity* for a planar vector field by mapping the plane to a sphere, letting infinity map to the north pole. We have then $ind_\infty = 2 - \sum_k ind_k$ where k runs over all (finitely many) fixed points of the vector field.

A.2.3.2. *The Hedgehog Theorem.* We provide here the proof of the Hedgehog Theorem stated in Chapter 3. The Theorem can now be seen as a consequence of Theorem A.5.

Theorem 3.7 (Hedgehog). *Any continuous vector field on the sphere has a singular point.*

Proof. By Theorem A.5 the sum of the indices of the vector field is 2. So it must have a singular point, by Corollary A.1. □

The name of the Theorem arises from the following metaphor. Consider a spherical hedgehog. The quills can be recast as vectors in 3-space, having one radial component, pointing outwards from the hedgehog and two tangent components. A smooth outer surface for a hedgehog quills corresponds then to a continuous planar vector field on the sphere. The corresponding statement is that a smooth hedgehog has at least one quill along the outer normal vector of its surface.

A.2.4. The Poincaré-Hopf Theorem. Theorem A.5 and 3.7 are of a very general character, they hold for *any* continuous vector field. Indeed, these dynamical results are linked to the underlying topological properties of the sphere, and hold, so to speak, *before* dynamics is even spoken of (or rather, independently of the particular nature of the dynamics in study). Poincaré realised that these Theorems could be formulated for more general surfaces and proved the Theorem that is now called *Poincaré-Hopf theorem* connecting topology and analysis. See [**USM95**] for a clear and simple approach. Here we will use ideas from dynamics to give a different approach and show the applicability of methods from dynamics.

First we consider orientable smooth closed surfaces S_g with g "holes" (g is called the *genus* of the surface). The sphere has zero holes while the torus has one and the pretzel pastry (in German: *brezel*) has three. See Figure A.7 for examples of surfaces with $g = 1, 2$.

We need to develop more topological background in order to state and prove The Poincaré-Hopf theorem.

A.2.4.1. *The Euler Characteristic of a Surface.* For a triangle in the plane, consider the integer computed by taking the number of faces (one in this case) minus edges (three) plus vertices (three). We have $\chi(\triangle) = f - e + v = 1$ and this holds for any planar triangle regardless of shape. Consider now a polygon in the plane, i.e., a closed, simply connected region bounded by a finite number of straight lines.

Definition A.16. A **triangulation** of a polygon is a finite partition of the polygon in non-overlapping triangles, glued to each other along the edges in such a way that each edge is shared by two adjacent triangles except the edges along the border of the polygon which belong to one triangle only (see Figure A.5).

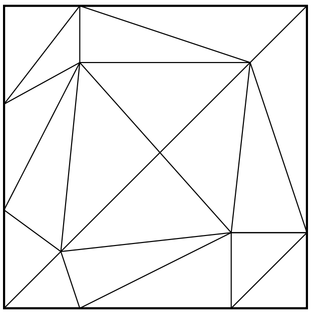


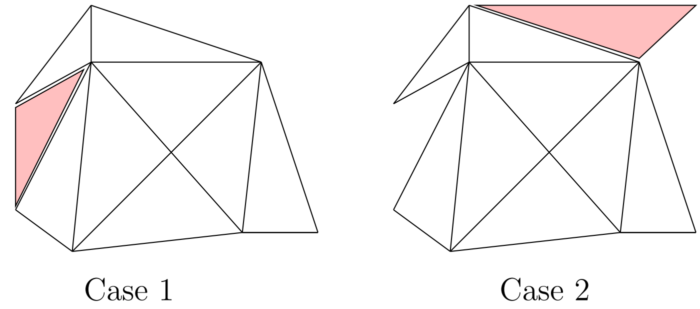
Figure A.5. A triangulation of a square.

Remark A.5. (a) All the original vertices of the polygon are vertices of the triangulation, but there may exist other vertices inside the polygon or along the original sides.
(b) For a closed compact surface (such as e.g., a sphere) each edge in a triangulation is shared by exactly two triangles (the surface has no boundary and hence there are no boundary triangles).

Lemma A.3. *For any triangulation of a polygon, the sum $f - e + v$ of faces, edges and vertices equals one.*

Proof. We proceed by induction. Given a triangulation of a polygon, let us pick one triangle arbitrarily and build the covering by adding adjacent triangles to it one at a time by gluing them to free edges, in such a way that at all intermediate steps we have a simply connected figure. The first triangle has $f - e + v = 1$. We will show that if at a step k in the construction we have $f_k - e_k + v_k = 1$ then the addition of another triangle does not change this fact.

There are exactly two ways to add a triangle to a collection of triangles, see Figure A.6. In the first case, the addition generates one new face, one new edge and no new vertices. Hence if $f_k - e_k + v_k = 1$ before the addition, now we have $f_k + 1 - (e_k + 1) + v_k = 1$. For the second case, we add one new face, two new edges and one new vertex. Thus, we obtain $f_k + 1 - (e_k + 2) + v_k + 1 = 1$. □



Case 1

Case 2

FIGURE A.6. Adding a triangle to a polygonal collection of triangles.

LEMMA A.4. *A (sub)triangulation of a triangle generated by (a) adding a new vertex along one side or (b) adding a new vertex at the interior of the triangle, does not alter the sum $f - e + v$.*

PROOF. In case (a) the edge where the vertex lies turns into two edges. To complete the triangulation, along with the new vertex we have to add an edge linking it to the opposite vertex of the triangle. The original face turns then into two as well. $f-e+v = 2-5+4 = 1$. In case (b) we add three edges from the new vertex to the original ones and get three faces instead of one. Hence, $f-e+v = 3-6+4 = 1$. □

LEMMA A.5. *Any triangulation of a surface S_g has the same sum $f - e + v$.*

PROOF. Given two triangulations T_1, T_2 we can construct a new triangulation $T = T_1 \cup T_2$ by taking the union of their vertices, keeping the edges and adding new edges to obtain a triangulation (the resulting triangulation need not to be unique!). Then T is a refinement of both T_1 and T_2 and has the same sum as any of the two. Hence the sum for T_1 and T_2 is the same. □

DEFINITION A.17. The **Euler characteristic** of a surface is the quantity $\chi(S) = f - e + v$ computed over any triangulation of the surface.

REMARK A.6. (a) We have shown that the Euler characteristic of a planar polygon is equal to one, regardless of the choice of triangulation. (b) The Euler characteristic is invariant under deformations of a triangulation that do not alter the numbers f, e and v.

LEMMA A.6. *For a sphere S, $\chi(S) = 2$.*

PROOF. Take an arbitrary triangulation for the sphere S. Pick one triangle and generate a hole on the sphere by eliminating the face of this triangle. Deform now the sphere to a polygon P on the plane having the edges of this triangle as boundaries, keeping the rest of the triangulation unchanged (except for the deformation). Since the triangulation of S has one more face than that of P we have $\chi(S) = 1 + \chi(P) = 2$. □

Lemma A.7. *For a torus T, $\chi(T) = 0$.*

Proof. Take a triangulation of the torus. Let us "cut" this surface in order to obtain a rectangle R. We need a longitudinal cut along the "equator" of the torus and a transversal cut as well. See Figure A.7.

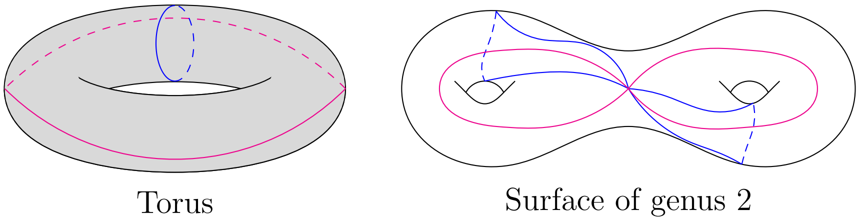


Figure A.7. Examples of "surface surgery" in order to compute $\chi(S_g)$. Magenta lines: equatorial cuts. Blue lines: transversal cuts.

Generate a sub-triangulation such that the two cut lines consist of edges and vertices of the triangulation. The triangulation of R coincides with that of T except that the vertices and edges along the cutting lines are duplicated since the two cutting lines are now the four edges of the rectangle. Let $e_t \geq 1$ be the edges along the top side of the rectangle and $e_l \geq 1$ those along the left side. Let $v_t = e_t + 1$, $v_l = e_l$ be the corresponding vertices (the vertex at the upper left corner has to be counted only once). Hence, $f_T - e_T + v_T = f_R - (e_R - e_t - e_l) + (v_R - (e_t + 1 + e_l)) = f_R - e_R + v_R - 1 = 0$. □

Definition A.18. **Barycentric Sub-triangulation**. For a given triangulation, consider the sub-triangulation produced by drawing the *medians* of each side, i.e., the lines going from the center of one edge to the opposite vertex. A standard result of planar geometry is that the medians cross at one point inside the triangle. The midpoint of each side and the crossing point become new vertices in the sub-triangulation. Hence, this triangulation modifies (f, e, v) of each triangle in the following way, see Figure A.8: $(1, 3, 3) \to (6, 12, 7)$.

With these elements we can state and subsequently prove the following Theorem:

Theorem A.6. [**USM95**] *(Poincaré-Hopf) Let S_g be a closed surface with genus g. For any vector field on S_g with finitely many critical points, the sum I_S of the indices of the critical points is equal to the Euler characteristic of S_g, namely $I_S = \chi(S_g) = 2 - 2g$.*

First we prove that there exists a vector field with the desired property.

Proposition A.2. *For any closed surface S_g of genus g there exists a vector field F with index I_F such that $I_F = \chi(S_g) = 2 - 2g$.*

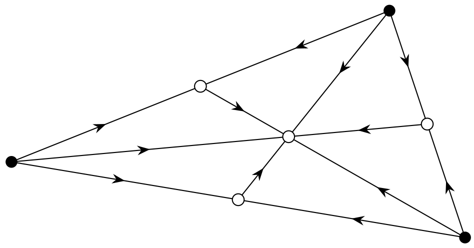


FIGURE A.8. The barycentric sub-triangulation and a flow on top of it. White dots: added vertices. For this flow, original vertices and the barycentric vertex have index 1 while the edge vertices have index -1. See Proposition A.2.

PROOF. Take an arbitrary triangulation of S_g and produce a barycentric sub-triangulation on each triangle. Generate a vector field F on it by letting the vertex inside each original face become a stable fixed point, while each original vertex becomes an unstable fixed point and the mid-side added vertices become saddle points. Let all edges become heteroclinic orbits connecting the fixed points such that an outgoing heteroclinic connects (a) each original vertex to the adjacent vertices (mid-side and in-face for each original triangle sharing the vertex) and (b) each mid-side vertex connects to the in-face vertex of each of the two triangles sharing the original edge along a (half) median, see Figure A.8.

The original triangulation has Euler characteristic $\chi(S_g) = f - e + v = 2 - 2g$. Let us now compute the index of the vector field F defined above. There are v unstable nodes, f stable nodes and e saddle fixed points. Hence, the index becomes $I_F = f + v - e = \chi(S_g)$. □

A.2.4.2. *Proof of Poincaré-Hopf Theorem.* It remains to be proven that two arbitrary continuous vector fields v, w with a finite number of singularities satisfying the conditions of the Theorem have the same index.

The argument used is a variation of the one presented for Theorem 3.7. Since we have a finite number of singular points there exists a point x_0 in the surface where both v and w are non singular. By smooth deformations of the original vector fields without altering the index we can even make both vector fields to coincide around x_0 (Straightening-out Theorem). Therefore the index along a small closed curve C encircling x_0 coincides for both vector fields. Note that since we are on a closed surface, this curve C encompasses all fixed points outside the region with x_0. A similar argument as in Theorem A.3

gives that

$$\begin{aligned} ind_v(C) = ind_w(C) &= \sum_{x_i \text{ singularities of } v \text{ "outside"} C} ind_{x_i} v \\ &= \sum_{y_i \text{ singularities of } w \text{ "outside"} C} ind_{y_i} w \\ &= I_v = I_w. \end{aligned}$$

This shows that the index of a vector field on S_g depends only on the genus and not on the vector field.

We are going to compute this index by using a special vector field v_g on S_g. First we decompose our surface by cutting it into pieces. The surface can be decomposed in $2g$ pieces, of which the $2(g-1)$ central ones look like a pair of pants (pair of pants decomposition) with two tubular pieces at each end. See Figure A.9.

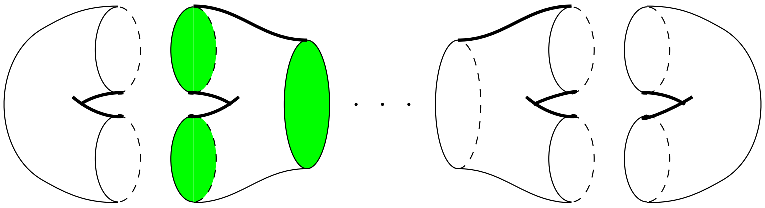


Figure A.9. "Pair of pants" decomposition. The inside volume of a pair of pants is coloured in green.

Thick lines indicate cuts on the surface in order to flatten it out, arriving to a "flat pair of pants decomposition", see Figure A.10.

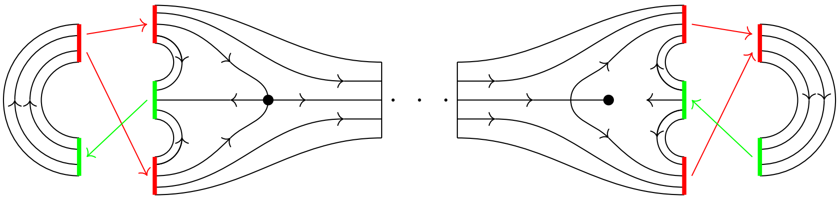


Figure A.10. Flow on the flattened "pair of pants" decomposition.

Equal colours indicate the correspondence of the cuts, i.e., in order to get back to the surface we have to glue the corresponding curves of the same color. The vector field indicated in the figure will be our desired vector field v_g. We see that we have $2(g-1)$ singularities, each of index -1. Therefore the index of this vector field (and consequently of any arbitrary continuous vector field with a finite number of singularities) equals $2(1-g)$. Since we already know that this is the same as the Euler characteristic $\chi(S_g)$ (consider the vector field F), the Theorem is proved.

A.3. Fixed Point Theorems

A.3.1. The Banach Fixed Point Theorem. The Banach Fixed Point Theorem is one of the basic theorems which helps to prove existence and uniqueness results in a broad generality.

It is of special interest for us since it is also a statement about a dynamical system. Moreover, the theorem is useful in a large number of branches of mathematics and science.

THEOREM A.7 (Banach's Fixed Point Theorem). [**Ban22**] *Let X be a complete metric space and $f\colon X \to X$ a* contraction*, i.e., a map such that there exists a number $0 \le \lambda < 1$ such that*

$$d(f(x), f(y)) \le \lambda d(x, y)$$

for all $x, y \in X$. Then the function f has a unique fixed point $x = f(x)$ on X.

PROOF. First we remark that because of its definition, a contraction is always a continuous Lipschitz function.

Let us fix a point $x_0 \in X$. We will show first that its trajectory $\{x_0, f(x_0), \cdots f^n(x_0), \cdots\}$ forms a Cauchy sequence. We have, for $n \in \mathbb{N}$

$$d(f^n(x_0), f^{n+1}(x_0)) \le \lambda d(f^{n-1}(x_0), f^n(x_0)) \le \lambda^n d(x_0, f(x_0)).$$

Hence, for $m > n$

$$\begin{aligned} d(f^n(x_0), f^m(x_0)) &\le d(f^n(x_0), f^{n+1}(x_0)) + d(f^{n+1}(x_0), f^{n+2}(x_0)) + \cdots \\ &\quad + d(f^{m-1}(x_0), f^m(x_0)) \\ &= \sum_{k=0}^{m-n-1} d(f^{n+k}(x_0), f^{n+k+1}(x_0)) \\ &\le \sum_{k=0}^{m-n-1} \lambda^{n+k} d(x_0, f(x_0)) \\ &= \lambda^n \frac{1}{1-\lambda} d(x_0, f(x_0)) \to 0, \quad \text{as} \quad \text{n} \to \infty. \end{aligned}$$

The last equality holds since $d(x_0, f(x_0))$ is finite and $\lambda < 1$, hence the sum is essentially a partial sum of the geometric series. This means that the trajectory of any point is a Cauchy sequence.

Since X is a complete metric space, this sequence is convergent. Let $x \in X$ be its limit. Then

$$f(x) = f(\lim_{n\to\infty} f^n(x_0)) = \lim_{n\to\infty} f(f^n(x_0)) = \lim_{n\to\infty} f^{n+1}(x_0) = x$$

thus proving that x is a fixed point.

It remains to prove that the fixed point is unique. Let us assume that $y \in X$ is also a fixed point. We will show that y coincides with x. Indeed, since both points are assumed to be fixed points, we have

$$d(x,y) = d(f(x), f(y)) \leq \lambda d(x,y).$$

Since $\lambda < 1$ this implies that $d(x,y) = 0$ or $x = y$. □

To gain some experience with Banach's Theorem, let us try it in some exercises.

A.3.2. Exercises.

Exercise A.6. Compute the fixed point of the sequence

(a) $x_{n+1} = \sqrt{1 + x_n}$, $x_0 = 0$.
(b) $x_{n+1} = \frac{1}{2}(x_n + \frac{3}{x_n})$, $x_0 = 1$.
(c) What happens if we change the initial condition?

Exercise A.7. Show that the sequence $a_{n+1} = \frac{1}{4} + a_n - {a_n}^2$, $n \geq 0$ converges for all initial conditions $a_0 \in [0,1]$. Compute its limit.

A.3.3. The 1–dimensional Fixed Point Theorem.

Theorem A.8. [**Rot88**, p.3] *Let $f \colon [0,1] \to [0,1]$ be continuous. Then f has a fixed point.*

Proof. If 0 or 1 are fixed points, we are done. Otherwise, consider the function $g(x) = f(x) - x$. We have $g(0) > 0 > g(1)$. Since g is also a continuous function, by the Mean Value Theorem and Rolle's Theorem, there exists a point x_0 such that $g(x_0) = f(x_0) - x_0 = 0$. □

A.3.4. Brouwer's Fixed Point Theorem.

Theorem A.9. [**Rot88**, p.5] *Let $f \colon \mathbb{D}^2 \to \mathbb{D}^2$ be a continuous map of the closed disc into itself. Then f has a fixed point.*

Proof. Assume there is no fixed point. Let $v_0(x) = \overline{xf(x)}$, i.e., the vector connecting x with its image $f(x)$. On the boundary of the disc we consider the vector fields $v_1(x) = \overline{xx^-}$ where x^- is the diametrically opposite point to x. For $0 \leq t \leq 1$ we can consider the vector field $v_t(x) = \overline{xx_t}$ where $x_t = t \cdot f(x) + (1-t)x^-$ (see Figure A.11)

None of the vectors v_t is zero ($x \neq f(x)$ for all x). hence v_t is a continuous deformation of v_0 into v_1. This implies that

$$ind_{v_0}\mathbb{S}^1 = ind_{v_1}\mathbb{S}^1 = 1.$$

By Corollary A.1 inside $\mathbb{S}^1$ the vector field v_0 must have a singular point, i.e., $x = f(x)$ contradicting our assumption. □

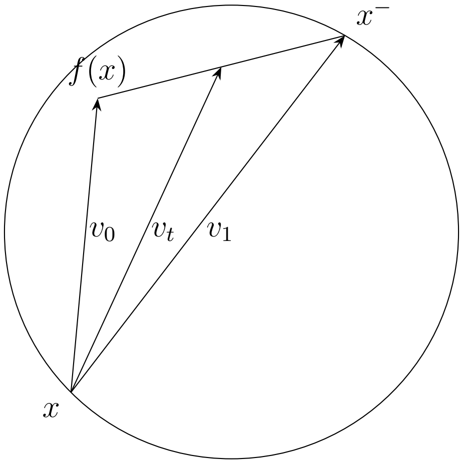


FIGURE A.11. Brouwer's fixed point theorem on the unit disc.

A.4. Linear Algebra and Matrix Algebra

Understanding square matrices is a fundamental ingredient of Dynamical Systems Theory. In many cases, simply looking at the linearisation of a system near a singular point is enough to understand its qualitative behaviour. We will deal with square real matrices throughout, although many results hold for complex matrices as well.

A.4.1. Basic Results on Square Matrices and Linear Equations. The following results are part of the Main Theorem in basic Linear Algebra courses. We state it here without proof.

THEOREM A.10. [**AR94**, 7.1.5 p.362] *Let A be a square real $n \times n$ matrix. The following statements are equivalent:*

- *A is invertible*
- $\det A \neq 0$
- *The equation $Ax = b$ has unique solution for any vector $b \in \mathbb{R}^n$*
- *The equation $Ax = 0$ has only the zero solution $x = 0$*
- *The columns of A are linearly independent*
- *The rows of A are linearly independent*

A.4.2. Eigenvalues and Eigenvectors. For a square $n \times n$ real matrix, the linear equation $(A - \lambda I)x = 0$ has (possibly complex-valued) nonzero solutions $x \in \mathbb{R}^n$ (or $\mathbb{C}^n$) if and only if $\det(A - \lambda I) = 0$ (see the previous Theorem). The latter is a real polynomial equation and by the Fundamental Theorem of Algebra, it has exactly n complex roots, when counted with their multiplicity. This polynomial is called the **characteristic polynomial** of the matrix A.

Definition A.19. Each distinct root of the polynomial equation $\det(A-\lambda I)=0$ is called an **eigenvalue**. The multiplicity of an eigenvalue regarded as a root of the characteristic polynomial is called **algebraic multiplicity**. Any member of the family of associated nonzero vectors x_λ for each eigenvalue is called an **eigenvector**.

A.4.3. Cayley-Hamilton Theorem and Other Results on Matrices. Other basic results from Linear Algebra include the following:

- A is invertible if and only if all its eigenvalues are nonzero.
- Each eigenvalue has at least one eigenvector. For a given eigenvalue λ_i, let $g_i \geq 1$, the **geometric multiplicity**, denote the number of associated linearly independent eigenvectors (note that $g_i = \dim\ \ker(A-\lambda_i I)$).
- Different eigenvalues have linearly independent eigenvectors.
- Dimension Theorem: For an $n \times n$ matrix A, $dim\ \ker(A) + dim\ ran(A) = n$, where $ran(A) = \{y \in \mathbb{R}^n : \exists x : y = Ax\}$ and $\ker(A) = \{x \in \mathbb{R}^n : Ax = 0\}$.
- $adj(A) \cdot A = A \cdot adj(A) = det(A) \cdot I$.

Theorem A.11 (Cayley-Hamilton Theorem). [**Gan59**, p.83] *Let the characteristic polynomial of the square matrix A be $p_A(\lambda) = det(A-\lambda I)$. Then, $p_A(A) = 0$ (the null matrix).*

Algebraic Proof. Consider $(A-\lambda I)\cdot adj(A-\lambda I) = p_A(\lambda)\, I$. The entries of the matrix $adj(A-\lambda I)$ are polynomials of degree $n-1$ in λ. Hence, we may rewrite this matrix as $adj(A-\lambda I) = \sum_{i=0}^{n-1} \lambda^i B_i$, where all matrices B_i commute with A. Hence,

$$(A-\lambda I)\cdot adj(A-\lambda I) = -\lambda^n B_{n-1} + \sum_{i=1}^{n-1} \lambda^i (AB_i - B_{i-1}) + AB_0 = p_A(\lambda)\, I.$$

This polynomial equality holds if and only if all (matrix) coefficients of the polynomials at each side coincide, i.e., $-B_{n-1} = I$, $AB_i - B_{i-1} = c_i I$ and $AB_0 = c_0 I$, where $p_A(\lambda) = c_0 + c_1\lambda + \cdots + \lambda^n$. Equality still holds after left-multiplying each equality by A^i and summing up. Hence, $A^n B_{n-1} + \sum_{i=1}^{n-1} A^i (B_{i-1} - AB_i) - AB_0 = 0 = p_A(A)$. □

Analytic Proof. It is immediate to verify that $p_A(A) = 0$ if A is a diagonal matrix and also if A is diagonalisable. For a general matrix, since diagonalisable matrices are dense in the set of all matrices, consider a sequence A_k of diagonal matrices having A as limit. Since the statement of the Theorem is closed, $p_A(A) = \lim_{k\to\infty} p_{A_k}(A_k) = 0$. □

A.4.4. The Jordan Canonical Form Theorem. Let us consider now further results related to eigenvalues and eigenvectors as a preparation for the Jordan Normal Form Theorem, which we discuss below.

LEMMA A.8. *There exist positive integers $\{k_i\}$ such that the set of eigenvalues of the matrix A decompose the vector space $\mathbb{V} = \mathbb{R}^n$ (or $\mathbb{C}^n$) as a direct sum of invariant subspaces:*

$$\mathbb{V} = \oplus_i \ker((A - \lambda_i I)^{k_i}).$$

PROOF. For each eigenvalue λ_i, consider the invariant subspaces $\ker(A - \lambda_i I)$ (spanned by a set of linearly independent eigenvectors associated to λ_i) and $\mathrm{ran}(A - \lambda_i I)$, of $\mathbb{V}$. If there exists a nonzero vector belonging to both subspaces, then some vector $x \in \mathbb{V}$ that is *not* in $\ker(A - \lambda_i I)$ is mapped by $A - \lambda_i I$ onto $\ker(A - \lambda_i I)$ (onto a linear combination of the eigenvectors associated to λ_i). Hence, we have that $\ker(A - \lambda_i I) \subset \ker((A - \lambda_i I)^2)$ (strict inclusion). This argument can be finitely repeated: there exists a maximal value $k_i \geq 1$ such that $\ker((A - \lambda_i I)^{k_i}) \cap \mathrm{ran}((A - \lambda_i I)^{k_i}) = \{0\}$, while for all integers $j > 0$, $\ker((A - \lambda_i I)^{k_i}) = \ker((A - \lambda_i I)^{k_i + j})$. Equality holds also for the range of these matrices, which shows that $\mathrm{ran}((A - \lambda_i I)^{k_i})$ does not contain eigenvectors associated to λ_i.

Hence, the invariant subspaces $\ker((A - \lambda_i I)^{k_i})$ and $\mathrm{ran}((A - \lambda_i I)^{k_i})$ decompose $\mathbb{V}$ as a direct sum, by the Dimension Theorem. Moreover, they are invariant under the action of A. We can hence further decompose $\mathrm{ran}((A - \lambda_i I)^{k_i})$ using another eigenvalue. It remains to be shown that when the set of eigenvalues is exhausted, the "residual" range contains only the zero vector. Assume that $\mathrm{ran}(\prod_i (A - \lambda_i I)^{k_i})$ has positive dimension. This subspace is A-invariant and hence it should contain an eigenvector to A which is not associated to any of its eigenvalues, but this would contradict the exhaustion. □

REMARK A.7. (i) The chain of subspace inclusions $\ker(A - \lambda_i I) \subset \cdots \subset \ker((A - \lambda_i I)^{k_i})$ is *strict*, i.e., each subspace has strictly larger dimension than the previous one. (ii) Restricted to the invariant subspace $\ker((A - \lambda_i I)^{k_i})$, the matrix $(A - \lambda_i I)^{k_i}$ is the null matrix. (iii) We will show below that $m_i = \dim \ \ker((A - \lambda_i I)^{k_i})$ equals the algebraic multiplicity of λ_i.

DEFINITION A.20. A **Jordan block** J_d of dimension d is a square $d \times d$ matrix of the following form:

$$J_d = \begin{pmatrix} \lambda & 1 & & 0 \\ 0 & \lambda & \ddots & \\ \vdots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & \lambda \end{pmatrix}$$

In other words, apart from λ along the diagonal and from 1 along the first upper sub-diagonal, all other matrix elements are zero. Jordan blocks of dimension 1 are trivially 1×1 (diagonal) matrices having λ as the only matrix element.

Theorem A.12 (Jordan Canonical Form). [**Gan59**, p.200-202] *Every square real matrix A can be reduced to a canonical form by a similarity transformation S:*

$$J = S^{-1}AS,$$

where J consists of Jordan blocks lying along the diagonal, each block with one of the eigenvalues of A along its diagonal. Jordan blocks associated to a given eigenvalue have dimensions that add up to the algebraic multiplicity of the eigenvalue.

Remark A.8. When the algebraic and geometric multiplicity of an eigenvalue λ_i coincide, there are g_i linearly independent eigenvectors (note also that from $m_i = g_i$ it follows that $k_i = 1$). The matrix S can be constructed by placing each eigenvector as a column of S. The corresponding Jordan blocks are of dimension 1, and hence diagonal matrices. If this occurs for all eigenvalues, we say that A is **diagonalisable**.

Before addressing the full problem let us consider the case of 2×2 matrices to gain insight.

Lemma A.9. *Every 2×2 real matrix A can be reduced to Jordan canonical form by a similarity transformation S:*

$$J = S^{-1}AS,$$

where J has the eigenvalue(s) of A in the diagonal, zero in the lower off–diagonal element and 0 *or* 1 *in the upper one, the latter case if and only if A has only one eigenvalue and one eigenvector.*

Proof. The characteristic polynomial is of degree 2 and hence it has either one or two distinct roots. In the latter case, we have two linearly independent eigenvectors and the matrix S can be built by placing these eigenvectors as columns. J is diagonal, as mentioned in the previous remark. If we have a double eigenvalue, then if its geometric multiplicity is two, we still have two linearly independent eigenvectors (constituting a base for the null-space of $A - \lambda I$) as before and J is diagonal. The remaining situation is when we have one eigenvalue and only one linearly independent eigenvector x_1.

Let $x_2 \neq cx_1$ be a vector linearly independent of x_1. There exist constants c_1, c_2 such that $(A - \lambda I)x_2 = c_1x_1 + c_2x_2$, since (x_1, x_2) is a basis. Let $b = (A - \lambda I)x_2$, which is a nonzero vector. Hence,

$$(A - \lambda I)b = (A - \lambda I)(c_1x_1 + c_2x_2) = c_2(A - \lambda I)x_2 = c_2b.$$

Therefore, b is an eigenvector of A, since $Ab = (\lambda + c_2)b$. But A has only one eigenvalue and only one linearly independent eigenvector, which means that $c_2 = 0$ and (by suitably scaling x_2 if necessary) $b = x_1$. We have then that $Ax_2 = x_1 + \lambda x_2$. Let $S = (x_1, x_2)$ be the invertible matrix having x_1 and x_2 as columns. Then, by direct computation we obtain,

$$AS = A(x_1, x_2) = (\lambda x_1, x_1 + \lambda x_2) = (x_1, x_2)\begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} = SJ.$$

□

Remark A.9. Note that $(A - \lambda I)^2 = 0$ since $(A - \lambda I)^2 x = 0$ for any vector x. A matrix B such that for some positive integer $n > 1$ satisfies $B^n = 0$ is called **nilpotent**. The reader may want to verify that in the definition of Jordan block, $(J_d - \lambda I_d)^d = 0$.

Proof of Jordan Theorem. Because of Lemma A.8, we may restrict the study to each invariant subspace $\ker((A - \lambda_i I)^{k_i})$. We consider the simplest situation first, i.e., when $g = 1$ (only one eigenvector). We also drop the index i for clarity.

By Remark A.8, we may reverse the ordering of the subspaces in the chain, i.e., we construct the set $K = \{y, (A - \lambda I)y, \cdots, (A - \lambda I)^{k-1}y\}$, where $y \in \ker((A - \lambda I)^k) \setminus \ker((A - \lambda I)^{k-1})$ (unless $k = 1$ and K consists of just one (eigen)vector). We also notice that the last vector of the chain K is an eigenvector. We claim that the set is linearly independent.

Consider $0 = c_0 y + \cdots + c_{k-1}(A - \lambda I)^{k-1}y$. Multiplying on the left by $(A - \lambda I)^{k-1}$ we find that $c_0 = 0$ (since all vectors other than y are mapped to zero). Multiplying by $(A - \lambda I)^{k-2}$ we find that also $c_1 = 0$. Continuing in the same way we find that all coefficients are zero.

Finally, we show that since there exists only one eigenvector, then $k = m$, i.e., the set K is a basis for $\ker((A - \lambda I)^k)$. Assume the contrary, i.e., that along with K there exists an additional linearly independent vector b in the basis. Still, $(A - \lambda I)^k b = 0$. Assume, without loss of generality, that $z = (A - \lambda I)^{k-1} b \neq 0$, then z has to be the (unique up to a nonzero constant) eigenvector in the set K. Hence, $(A - \lambda I)^{k-1}(b - a_0 y) = 0$. Again, since there is only one eigenvector, $(A - \lambda I)^{k-2}(b - a_0 y)$ has to be this eigenvector and hence $(A - \lambda I)^{k-2}(b - a_0 y - a_1(A - \lambda I)y) = 0$. Proceeding in the same way, we realise that $(A - \lambda I)(b - a_0 y - a_1(A - \lambda I)y - \cdots - a_{k-2}(A - \lambda I)^{k-2}y) = 0$, thus contradicting the assumption that b was linearly independent with the set K.

Summing up, building with the vectors of K the invertible matrix S, we can now verify by direct computation as in the case $n = 2$ that restricted to the invariant subspace $\ker((A - \lambda I)^k)$, $AS = SJ$ with J a Jordan block. By manually computing $\det(A - \lambda I)$ we realise that m equals the algebraic multiplicity of λ.

For the case $g > 1$, the proof along these lines gets cumbersome, but not impossible. We construct a basis for $\ker((A - \lambda I)^k)$ in the following way: (i) Pick a vector $y_1 \in \ker((A - \lambda I)^k) \setminus \ker((A - \lambda I)^{k-1})$ and construct a chain K_1 ending at an eigenvector. This chain has linearly independent vectors for the same reason as above. (ii) Consider now the invariant subspace $\ker((A-\lambda I)^k)\} \cap (span(K_1))^\perp$ and start all over again, i.e., compute k_2 (may be smaller than k), pick a vector $y_2 \in \{\ker((A-\lambda I)^{k_2}) \setminus \ker((A-\lambda I)^{k_2-1})\} \cap (span(K_1))^\perp$ (it is important that y_2 lies in the orthogonal complement of the subspace spanned by K_1). In this way we generate a new linearly independent chain K_2 ending in another eigenvector. (iii) Repeat this procedure inductively and generate new chains. Since $m = \dim\ \ker((A-\lambda I)^k)$ is finite, the process comes to an end when we ran out of eigenvectors.

Summing up now, we will have as many chains as the geometric multiplicity of λ since each chain ends in a linearly independent eigenvector. Some chains have length k, others may have smaller length (even length 1) and the sum of all lengths for all chains is m, since the set of vectors thus constructed is linearly independent and exhausts $\ker((A-\lambda I)^k)$. This argument is inspired in the following proposition:

Proposition A.3. *Every nilpotent operator B has a* chain basis *for ran(B) (i.e., a basis consisting of chains of vectors $\{v_j, Bv_j, \cdots, B^{s_j}v_j\}$, where the last vector in each chain is one of the linearly independent eigenvectors of B lying in ran(B)).*

A proof of this classical result can be found in e.g., [**Gan59**, p.200].

Using the chain basis constructed above, the matrix A is split up into diagonal blocks, one for each eigenvalue. Each of these blocks has dimension m_i and consists in a juxtaposition of g_i Jordan blocks of different dimensions, having λ in the diagonal and ones in the upper diagonal (for the chains of length larger than one). Again, computing by hand $\det(A-\lambda I)$ in this basis, we realise that m_i equals the algebraic multiplicity of λ_i. □

A.4.5. The Perron–Frobenius Theorem.

Theorem A.13 (Perron–Frobenius). [**KH96**, 1.9.11 p.52] *Let A be a $n \times n$ matrix with strictly positive entries. Then*

- *A has a positive eigenvalue r.*
- *There is a strictly positive eigenvector associated to the eigenvalue r.*
- *No other eigenvector has strictly positive coordinates.*
- *The eigenvalue r is* **simple** *(only one eigenvector and Jordan block of dimension one).*
- *All other eigenvalues ω satisfy $|\omega| < r$.*

Proof. Consider the set $\mathbb{T} \subset \mathbb{R}^n$ of non-negative rays (of vectors with non-negative coordinates). Since A is strictly positive, $A(\mathbb{T}) \subset \mathbb{T}$, moreover, $\partial\mathbb{T}$ (i.e., non-negative rays with some zero components) is mapped onto strictly positive rays. By the Brouwer fixed point theorem, there is a fixed point of A on $\mathbb{T}$, i.e., there exists a positive eigenvector for A with positive eigenvalue. This proves the first two statements. We call the eigenvalue r and the positive eigenvector v.

The matrix A^T has also the positive eigenvalue r with positive eigenvector u. This means that $u^TA = ru^T$. Hence, all positive eigenvectors w of A have the same eigenvalue r. Indeed, let $Aw = \mu w$, for some positive eigenvector w. Then $u^TAw = \mu u^Tw = ru^Tw$. Since $u^Tw > 0$, $\mu = r$.

Assume now that there exists an invariant plane π associated to r containing the eigenvector v. Consider the circle $\mathbb{S}^1$ on this plane and let L be the arc of non-negative rays on $\pi \cap \mathbb{S}^1$. The endpoints of L map onto positive rays, hence L and $\mathbb{S}^1$ are not point-wise invariant by A (after eventually rescaling along rays). Hence, r has only one eigenvector (i.e., its geometric multiplicity is one). This proves the third statement and part of the fourth one.

To complete the proof of the fourth statement, we will show that the equation $(A - rI)u = v$ has no solutions (and therefore, the algebraic multiplicity of r is also one). The matrix $A - rI$ has negative entries in the diagonal, since from the eigenvalue equation we obtain $r = a_{ii} + \sum_{j\neq i} a_{ij}v_j/v_i$. Let us now solve the equation $(A-rI)x = 0$ by Gauss elimination. We know that this equation has a unique strictly positive solution v up to a constant factor. Pivoting by the first row amounts to adding positive numbers to rows and columns 2 to n in $A-rI$. The resulting sub-matrix must still have negative diagonal entries, otherwise the eigenvalue equation cannot have a positive solution. Repeating this argument we realise that the *echelon* form of $A-rI$ has negative diagonal elements, except the last one that is zero (only one row of zeroes because the geometric multiplicity is one) and positive upper-diagonal elements . Attempting now to solve $(A - rI)u = v$ by Gauss elimination, the elimination procedure adds positive terms to each entry in v. Hence the equation corresponding to the last row has no solutions.

Let us split the proof of the final statement of the theorem in two parts. Let μ be a real eigenvalue of A with eigenvector w and u^T be the left eigenvector of A associated to r. It is clear from earlier arguments that eigenvector w must have negative entries. Let $|w|$ be the (non-negative) vector obtained from w by taking the absolute value of each entry. We have $|\mu|\ |w| = |\mu w| = |Aw| < A|w|$. By multiplication with u^T we obtain: $|\mu|\ u^T|w| < u^TA|w| = ru^T|w|$, hence $r > |\mu|$ since $u^T|w| > 0$.

Concerning complex eigenvalues η, the previous argument shows that $|\eta| \leq r$. Assume there exists a true complex eigenvalue where equality holds. Let w be the associated eigenvector and $|w|$ the non-negative vector obtained by taking the absolute value of each entry in w. First, we note that $A|w| = r|w|$. Indeed, $r|w| = |\eta w| = |Aw| \leq A|w|$. Hence, $z = A|w| - r|w|$ is non-negative. However, $u^Tz = u^TA|w| -$

$ru^T|w| = 0$, what proves that z is the null vector. This proves in addition that $|w| = v$. Moreover, since $A|w| = |Aw|$, equality holds in the triangular inequality. Therefore, $v = |w| = e^{i\theta}w$, which is a contradiction since v and w should be linearly independent. Then strict inequality holds, namely $|\eta| < r$. □

A.4.6. On the Eigenvalues of 2×2 matrices. Real 2×2 matrices are particularly simple and many stability calculations for maps from Chapter 5 can be done with paper and pencil. Let the trace (resp determinant) of such a matrix be T (resp D). The characteristic equation becomes: $\lambda^2 - \lambda T + D = 0$. We will consider two examples: (a) the condition to obtain one eigenvalue with modulus larger than one and the other eigenvalue with modulus smaller than one, and also (b) the condition giving both eigenvalues with modulus smaller than one.

A.4.6.1. *(a) Unstable (saddle) singular points.* Whenever $T^2/4 = 0$ or $T^2/4 \leq D$ then the modulus of both eigenvalues is $|\lambda_i| = \sqrt{|D|}$, so they cannot have different size. Hence, to obtain eigenvalues of different modulus it is necessary that $T^2/4 > \max(D, 0)$. Moreover, the eigenvalues are then real. For positive T the largest-modulus eigenvalue is positive and for negative T it is negative. Hence the proposed condition reads:

$$\frac{|T|}{2} + \sqrt{\frac{T^2}{4} - D} > 1 > \left|\frac{|T|}{2} - \sqrt{\frac{T^2}{4} - D}\right|.$$

For positive D the smaller eigenvalue is positive and the condition above reads:

$$\sqrt{\frac{T^2}{4} - D} > \left|1 - \frac{|T|}{2}\right|,$$

which leads to $|T| > 1 + D = |1 + D|$. For negative D the condition can be recast as:

$$\frac{|T|}{2} + \sqrt{\frac{T^2}{4} - D} > 1 > \sqrt{\frac{T^2}{4} - D} - \frac{|T|}{2} \Rightarrow \frac{|T|}{2} > \left|1 - \sqrt{\frac{T^2}{4} - D}\right|,$$

which after squaring a couple of times can be written as $T^2 > (1 + D)^2$, and hence $|T| > |1 + D|$. Hence, in either case the condition for having one eigenvalue with modulus larger than one and the other with modulus smaller than one reads: $|T| > |1 + D|$ **and** $T^2/4 > \max(D, 0)$.

A.4.6.2. *(b) Stable singular points.* Following the same reasoning as above, whenever $T^2/4 < D$ both eigenvalues have modulus $\sqrt{D}$. Hence, the stability condition reads in this case $1 > D > T^2/4 > 0$. Otherwise, when $T^2/4 \geq D$ the eigenvalues are real and the largest-modulus eigenvalue satisfies:

$$|\lambda_1| = \frac{|T|}{2} + \sqrt{\frac{T^2}{4} - D}.$$

In this case, the stability condition is $1 > T^2/4 \geq D > |T| - 1 \geq -1$.

A.5. Linear operators

A.5.1. Operator norm. When matrices are regarded as linear mappings between finite-dimensional vector spaces there is a natural generalisation to arbitrary vector spaces through the concept of linear operators. They are just continuous mappings between vector spaces. Spaces of linear operators are also vector spaces and a norm can be defined on them following the intuition from matrix norm.

Definition A.21 (Operator Norm). Let X be a Banach space. A linear operator $A : X \to X$ has norm

$$\|A\| := \sup_{\|x\|=1} \|A(x)\|.$$

Again, the norm of the operator is defined through vector norms.

A.5.2. Norm Estimates.

Theorem A.14. [**Lax02**, p.176] *Let L, G be linear operators in a Banach space E with $\|L\| \leq a < 1$ and $\|G^{-1}\| \leq a < 1$ and* Id *be the identity operator* $\operatorname{Id}(x) = x$. *Then the operators* $\operatorname{Id} + L$ *and* $\operatorname{Id} + G$ *are isomorphisms with*

$$\|(\operatorname{Id} + L)^{-1}\| \leq \frac{1}{1-a}, \qquad \|(\operatorname{Id} + G)^{-1}\| \leq \frac{a}{1-a}.$$

Proof. For a fixed element $y \in E$ define the map

$$F(x) := y - L(x).$$

Then

$$\|F(x_1) - F(x_2)\| = \|L(x_1) - L(x_2)\| \leq a\|x_1 - x_2\|.$$

By the Banach Fixed Point Theorem there is a unique solution to the fixed point problem $x = F(x)$. Hence for each y there is a unique x such that $x = F(x) = y - L(x)$, or equivalently, the equation $(\operatorname{Id} + L)x = y$ has a unique solution x. Therefore $\operatorname{Id} + L$ is a bijection.

We now want to estimate the norm of $(\operatorname{Id} + L)^{-1}$. We have

$$\|(\operatorname{Id} + L)^{-1}\| = \sup_{\|x+L(x)\|=1} \|x\|.$$

But we have

$$1 = \|x + L(x)\| \geq \|x\| - a\|x\|.$$

This implies that $\|(\operatorname{Id} + L)^{-1}\| \leq \frac{1}{1-a}$.

For the second statement we remark that

$$\operatorname{Id} + G = G\left(\operatorname{Id} + G^{-1}\right).$$

By the above considerations for $L = G^{-1}$ we have that $\operatorname{Id} + G^{-1}$ is a bijection and also G is invertible since G^{-1} is assumed to exist. Then

$$\|(\operatorname{Id} + G)^{-1}\| \leq \| \left(\operatorname{Id} + G^{-1}\right)^{-1} \| \cdot \|G^{-1}\| \leq \frac{a}{1-a}.$$

□

A.6. Algebra

A.6.1. Closed subgroups of the real numbers. We will use the following fact about closed subgroups of the real numbers.

Definition A.22. A set G is called an additive **group** if there exists an operation $(\cdot,\cdot) : G \times G \to G; \quad (a,b) = a + b \in G$ called **addition** defined with the following properties

1. Associativity: $a + (b + c) = (a + b) + c$ for all $a, b, c \in G$,
2. Neutral element: There is an element $0 \in G$ such that
$$a + 0 = 0 + a = a \quad \text{for all } a \in G.$$
3. Inverse element: For any $a \in G$ there is an element $-a \in G$ with
$$a + (-a) = -a + a = 0.$$

Remark A.10. The real numbers form an additive group with the usual addition. Moreover, the real numbers are a topological space. Therefore we can talk about open and closed sets.

Theorem A.15. [**GL87**, p.18] *Any proper closed subgroup of* $\mathbb{R}$ *as an additive group has the form* $\tau \cdot \mathbb{Z}$*, where* $\tau \in \mathbb{R}$*. Such groups are called* **lattices***.*

Proof. We will proceed in several steps.

Step 1. Any subgroup G of $\mathbb{R}$ without a smallest positive element is dense in $\mathbb{R}$.

Let us fix a number $r \in \mathbb{R}$ and a number $\epsilon > 0$. Since there is no smallest positive element we can find a sequence $G \ni g_n \to 0$ with $g_n > 0$. Let $m \in \mathbb{N}$. Obviously we have $mg_n \in G$. We want to show that for some numbers $m, n \in \mathbb{N}$ we have $mg_n \in (r - \epsilon, r + \epsilon)$. Since r and ϵ are arbitrary, this implies that G is dense in $\mathbb{R}$. The last inclusion can be rewritten as

$$g_n \in \left(\frac{r-\epsilon}{m}, \frac{r+\epsilon}{m}\right).$$

If we show that

$$\{g_n\} \cap \bigcup_{m \in \mathbb{N}} \left(\frac{r-\epsilon}{m}, \frac{r+\epsilon}{m}\right) \neq \varnothing, \tag{A.1}$$

then we are sure that some element $mg_n \in G$ lies closer to r than ϵ. Now we have that for $m > \frac{r-\epsilon}{2\epsilon}$

$$\frac{r-\epsilon}{m} < \frac{r+\epsilon}{m+1}.$$

This means that two consecutive intervals overlap and hence

$$\bigcup_{m\in\mathbb{N}} \left(\frac{r-\epsilon}{m}, \frac{r+\epsilon}{m}\right) \supset (0, 2\epsilon).$$

Since $0 < g_n \to 0$ for n large enough we have $g_n < 2\epsilon$ and eq.(A.1) is proved.

Step 2. Since G is closed and proper it is not dense. Hence, we know by step 1 that G has a smallest element $\tau > 0$. Clearly $\tau \cdot \mathbb{Z} \subset G$. Assume that there is a number $g \in G \setminus \tau \cdot \mathbb{Z}$. If we set

$$n = \max\{k \in \mathbb{Z} \,:\, k\tau < g\}$$

then

$$g = n\tau + r, \qquad 0 < r < \tau.$$

But $r = g - n\tau \in G$ contradicting that τ is the smallest element. □

A.7. The Liouville Theorem of Hamiltonian Dynamics

Towards the second half of the 19th. century there were a lot of developments arising from the original work of Newton in Mechanics, with contributions by Euler, D'Alembert, Lagrange and many others, culminating with the work by Hamilton on the equations of motion. We develop here the tools for proving the Liouville Theorem on fully integrable hamiltonian systems.

A.7.1. Example: The Harmonic Oscillator. The most thoroughly described example in Classical Mechanics regards the motion of a point particle of constant mass m along a straight line, subject to a force $F = -kx$, where x is the displacement of the particle along the line from an equilibrium position and k is a constant. Newton's equation reads

$$m\frac{d^2x}{dt^2} = -kx.$$

With the introduction of the auxiliary variable $p = m\dfrac{dx}{dt}$ the equation turns into the dynamical system

$$\begin{aligned}\frac{dx}{dt} &= \frac{p}{m}\\ \frac{dp}{dt} &= -kx\end{aligned}$$

The auxiliary function $H(x,p) = \frac{1}{2}\left(\frac{p^2}{m} + kx^2\right)$, called the *Hamiltonian* has two interesting properties,

- H is constant along trajectories, i.e.,
$$\begin{aligned}\frac{dH}{dt} &= \frac{\partial H}{\partial x}\frac{dx}{dt} + \frac{\partial H}{\partial p}\frac{dp}{dt} \\ &= kx\frac{p}{m} + \frac{p}{m}(-kx) = 0.\end{aligned}$$
- The equations of motion can be rephrased as
$$\begin{aligned}\frac{dx}{dt} &= \frac{\partial H}{\partial p} \\ \frac{dp}{dt} &= -\frac{\partial H}{\partial x}\end{aligned}$$

Another interesting fact is that there exist coordinate transformations (called *canonical transformations*) $(x,p) \to (r,s)$ such that they preserve the above structure. In other words, $H(r,s) = H(x(r,s),p(r,s))$ is still a constant along trajectories and the dynamical equations in the new coordinates read

$$\begin{aligned}\frac{dr}{dt} &= \frac{\partial H}{\partial s} \\ \frac{ds}{dt} &= -\frac{\partial H}{\partial r}\end{aligned}$$

Consider the new coordinate $J = \oint p dx$ where the integral is taken along a trajectory. With the substitution $x = \sqrt{\frac{2E}{k}}\sin\theta$ and integrating $\theta \in [0, 2\pi]$ we obtain $J = 2\pi H\sqrt{\frac{m}{k}}$. Where $H = E$ is the value of H along the given trajectory. Let W be the companion coordinate of J. Whatever W is, we will have that $\frac{\partial H}{\partial W} = 0$ since we have already managed to express $H = \sqrt{\frac{k}{m}}\frac{J}{2\pi}$ as a function of J only, which also turns to be a constant along trajectories. The new equations of motion read

$$\begin{aligned}\frac{dW}{dt} &= \frac{\partial H}{\partial J} = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \\ \frac{dJ}{dt} &= -\frac{\partial H}{\partial W} = 0\end{aligned}$$

with immediate solution.

The idea behind Liouville Theorem is to state the conditions in which a $2n$-dimensional dynamical system of this kind will have immediate solution.

A.7.2. Liouville Theorem.

DEFINITION A.23. An autonomous $2n$-dimensional dynamical system is called *Hamiltonian* if there exists a function $H(q_1 \cdots q_n, p_1 \cdots p_n)$ such that the dynamical equations can be rephrased as

$$\begin{aligned} \frac{dq_i}{dt} &= \frac{\partial H}{\partial p_i} \\ \frac{dp_i}{dt} &= -\frac{\partial H}{\partial q_i}, \quad i = 1, \cdots, n. \end{aligned}$$

DEFINITION A.24. Given a Hamiltonian system, a function $f(q_1 \cdots q_n, p_1 \cdots p_n)$ that is constant along trajectories, i.e, such that

$$\begin{aligned} \frac{df}{dt} &= \sum_{i=1}^{n} \frac{\partial f}{\partial q_i}\frac{dq_i}{dt} + \frac{\partial f}{\partial p_i}\frac{dp_i}{dt} \\ &= \sum_{i=1}^{n} \frac{\partial f}{\partial q_i}\frac{\partial H}{\partial p_i} - \frac{\partial f}{\partial p_i}\frac{\partial H}{\partial q_i} = 0 \end{aligned}$$

is called a *constant of the motion*. In particular, in view of Definition A.23 the Hamiltonian function H is a constant of motion.

THEOREM A.16 (Liouville, Arnold). [**Arn89**] *A Hamiltonian system having n independent constants of the motion $f_1, \cdots, f_n$ in involution (of which we say that the first one, $f_1 = H$) satisfies that*

- *If the n-dimensional level surface of first integrals $\mathcal{M}_f = \{(q_1 \cdots q_n, p_1 \cdots p_n) : f_k = c_k\}$ is compact and connected then it is diffeomorphic to a torus $T^n = S^1 \times \cdots \times S^1$.*
- *A coordinate system $I_1, \cdots, I_n, \phi_1, \cdots \phi_n$ can be introduced in phase-space such that the angles $\phi_1, \cdots \phi_n$, $0 \leq \phi_k \leq 2\pi$ are coordinates on $\mathcal{M}_f$ and $I_1, \cdots, I_n$ are adequate functions of $(f_1, \cdots, f_n)$ and also independent constants of the motion.*
- *The dynamical equations in this coordinate system become*

$$\frac{dI_k}{dt} = 0, \quad \frac{d\phi_k}{dt} = w_k(I_1, \cdots, I_n), \qquad k = 1, \cdots, n$$

Hence, the original system is solvable by quadratures (integration of known functions).

PROOF. We start by making the following facts explicit:

- For each possible choice of constants $c_1, \cdots, c_n$ the surface $\mathcal{M}_f$ (which is compact by assumption) where the motion takes place is defined.

- By *independent* constants of the motion it is meant that the normal vectors (given by ∇f_k) to the hyper-surfaces are linearly independent. By the constants of motion being *in involution* it is meant that the flow on $\mathcal{M}_f$ associated to the vector fields

  $$V^{f_k}(p\cdots,q) = \left(\frac{\partial f_k}{\partial p_1},\cdots,\frac{\partial f_k}{\partial p_n},-\frac{\partial f_k}{\partial q_1},\cdots,-\frac{\partial f_k}{\partial q_n}\right),$$

  $(k = 1,\cdots,n)$, commute, i.e., that

  $$\varphi^{V^{f_k}}(t,\varphi^{V^{f_j}}(s,x_0)) = \varphi^{V^{f_j}}(s,\varphi^{V^{f_k}}(t,x_0)).$$

Since on the manifold $\mathcal{M}_f$ the flow given by the vector fields are defined for all times, this flow is described by a mapping $\Phi^{t_1,\cdots,t_n} : \mathcal{M}_f \to \mathcal{M}_f$ given by $\Phi^{t_1,\cdots,t_n}(x) = \varphi^{Vf_n}(t_n,\cdots\varphi^{Vf_1}(t_1,x))$. In other words, the flow is described by the action of the additive group $\mathbb{R}^n = \{t_1,\cdots,t_n\}$ on $\mathcal{M}_f$. Since $\mathcal{M}_f$ is compact while $\mathbb{R}^n$ is not, there are elements $\tilde{t}_1,\cdots,\tilde{t}_n$ such that $\Phi^{\tilde{t}_1,\cdots,\tilde{t}_n}(x_0) = x_0$. This property is defines an additive, discrete, proper subgroup Γ_{x_0} of $\mathbb{R}^n$ of the form $\tau\cdot\mathbb{Z}^n$ in the same way as it was done for $n = 1$ in Theorem A.15. We note that the groups Γ_{x_0} and Γ_{y_0} are conjugate via the element $\Phi^{\tilde{t_1},\cdots,\tilde{t_n}}(x_0) = y_0$ since the vector field acts transitively on the energy surfaces. Also, in the same way as the quotient group $\mathbb{R}/\mathbb{Z}$ is the unit circle $\mathbb{S}^1$, we have that $\mathbb{R}^n/\mathbb{Z}^n = \mathbb{T}^n = \mathbb{S}^1\times\cdots\times\mathbb{S}^1$. Hence, the compact manifold $\mathcal{M}_f$ is shown to be diffeomorphic to $\mathbb{T}^n$ by the action of Φ. This proves the first statement.

By a suitable change of coordinates, the dynamics can be separated in n independent dynamics on each coordinate ϕ_i, $i = 1,\cdots,n$ of $\mathbb{T}^n$ with constant RHS. With this coordinate change, the constants $f_1,\cdots,f_n$ transform to new constants $I_1,\cdots,I_n$ and the dynamics on the Torus is separated as

$$\frac{d\phi_i}{dt} = V^{I_i}$$

with constant RHS depending solely on the I_i's.

By the implicit function theorem, equations $f_k = c_k$ can be simultaneously solved for $p_i(q_1,\cdots,q_n,c_1,\cdots,c_n)$ and the relations $f_i(q_i,p_i(q_1,\cdots,q_n,c_1,\cdots,c_n)) = c_i$ hold identically on $\mathcal{M}_f$. Moreover, since

$$\frac{\partial f_i}{\partial q_j} + \sum_{k=1}^{n}\frac{\partial f_i}{\partial p_k}\frac{\partial p_k}{\partial q_j} = 0$$

we have that

$$\sum_{j}^{n}\frac{\partial f_m}{\partial p_j}\frac{\partial f_i}{\partial q_j} + \sum_{j,k}^{n}\frac{\partial f_m}{\partial p_j}\frac{\partial f_i}{\partial p_k}\frac{\partial p_k}{\partial q_j} = 0.$$

By interchanging i and m we obtain

$$\sum_{j,k}^{n}\left(\frac{\partial f_m}{\partial p_j}\frac{\partial f_i}{\partial q_j}-\frac{\partial f_i}{\partial p_j}\frac{\partial f_m}{\partial q_j}\right)+\sum_{j,k}^{n}\frac{\partial f_m}{\partial p_j}\frac{\partial f_i}{\partial p_k}\left(\frac{\partial p_k}{\partial q_j}-\frac{\partial p_j}{\partial q_k}\right)=0.$$

The first sum vanishes because of the involution condition. Hence, the second sum implies that $\left(\frac{\partial p_k}{\partial q_j}-\frac{\partial p_j}{\partial q_k}\right)=0$ and hence,

$$\oint_{\Gamma}\sum_{j}p_j dq_j$$

is an exact differential integrating to zero on any closed loop Γ on $\mathcal{M}_f$ contractible to a point. However, since $\mathcal{M}_f$ is diffeomorphic to $\mathbb{T}^n$ the previously invoked coordinate change corresponds to n distinct closed loops γ_k that are not contractible to a point (one for each unit circle building $\mathbb{T}^n$) and the new constants of motion to the quantities

$$I_k=\oint_{\gamma_k}\sum_{j}p_j(q)dq_j.$$

By construction, the I_k's are constants depending only on $c_1,\cdots,c_n$ (and hence constants of the motion). Moreover, this relation is invertible to express $c_k(I_1,\cdots,I_n)$, what in particular gives that the Hamiltonian $c_1=f_1=H$ depends solely on the I_k's. This proves the second statement.

Propagating the coordinate change $(\phi_1,\cdots,\phi_n,I_1,\cdots,I_n)$ and using that

$$\sum_{i=1}^{n}\frac{\partial \phi_j}{\partial q_i}\frac{\partial I_k}{\partial p_i}-\frac{\partial I_k}{\partial q_i}\frac{\partial \phi_j}{\partial p_i}=\nabla\phi_j\cdot V^{I_k}=\delta_{jk},$$

the dynamical equations in these coordinates read

$$\frac{d\phi_i}{dt}=\frac{\partial H}{\partial I_i};\quad \frac{dI_i}{dt}=0\qquad i=1,\cdots,n.$$

□

Remark A.11. The technical formulation of the property that two constants of motion f_1 and f_2 are *in involution* is that

$$\sum_{i=1}^{n}\frac{\partial f_1}{\partial q_i}\frac{\partial f_2}{\partial p_i}-\frac{\partial f_2}{\partial q_i}\frac{\partial f_1}{\partial p_i}=0.$$

Remark A.12. The modern, differential geometric, coordinate-free version of this Theorem using Lie algebras and their commutators can be found in [**Arn89**].

A.8. Uniform Distribution

The notion of uniform distribution comes from Monte–Carlo–Methods. It deals with pseudo–random sequences.

DEFINITION A.25. A sequence $(x_n) \in [0,1]$ is **uniformly distributed** if for any continuous function $f\colon [0,1] \to \mathbb{C}$ we have

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(x_n) = \int_{[0,1]} f(x)\,dx.$$

THEOREM A.17 (H. Weyl). [**Wey16**] *A sequence (x_n) is uniformly distributed if and only if*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} e^{2\pi i h x_n} = 0$$

for all $0 \neq h \in \mathbb{Z}$.

PROOF. By Weierstraß' theorem, trigonometric polynomials are dense in the set of continuous functions of the interval, i.e., for any continuous f and any $\epsilon > 0$ there are numbers $k \in \mathbb{N}$, $z_j \in \mathbb{C}$, $n_j \in \mathbb{N}$; $1 \leq j \leq N$ such that

$$\max_{x\in[0,1]} \left| \sum_{j=1}^{k} z_j e^{2\pi i n_j x} - f(x) \right| < \epsilon.$$

Hence,

$$\begin{aligned}
&\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(x_n) - \int_{[0,1]} f(x)\,dx = \\
&\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(x_n) - \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \left(\sum_{j=1}^{k} z_j e^{2\pi i n_j x_n} \right) + \\
&\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \left(\sum_{j=1}^{k} z_j e^{2\pi i n_j x_n} \right) - \int_{[0,1]} \left(\sum_{j=1}^{k} z_j e^{2\pi i n_j x_n} \right) dx + \\
&\int_{[0,1]} \left(\sum_{j=1}^{k} z_j e^{2\pi i n_j x_n} \right) dx - \int_{[0,1]} f(x)\,dx.
\end{aligned}$$

Next note that for $0 \neq h \in \mathbb{Z}$

$$\int_{[0,1]} e^{2\pi i h x}\,dx = 0$$

and also

$$\left|\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}f(x_n)-\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\left(\sum_{j=1}^{k}z_je^{2\pi in_jx_n}\right)\right|<\epsilon,\quad\text{and}$$

$$\left|\int_{[0,1]}\left(\sum_{j=1}^{k}z_je^{2\pi in_jx_n}\right)dx-\int_{[0,1]}f(x)\,dx\right|<\epsilon.$$

Hence, we may write

$$\left|\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}f(x_n)-\int_{[0,1]}f(x)\,dx\right|\leq 2\epsilon+\left|\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\left(\sum_{j=1}^{k}z_je^{2\pi in_jx_n}\right)\right|$$

and

$$\left|\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\left(\sum_{j=1}^{k}z_je^{2\pi in_jx_n}\right)\right|\leq 2\epsilon+\left|\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}f(x_n)-\int_{[0,1]}f(x)\right|,$$

thus proving the statement. □

A.9. Number Theory

A standard theorem in Number Theory is the following

THEOREM A.18 (**Kronecker**). [**Kro84**] *Let $\alpha\in\mathbb{R}\setminus\mathbb{Q}$, then the sequence $n\alpha \mod 1$ is dense in the interval $[0,1]$.*

PROOF. We are not going to prove this theorem directly. We will prove a stronger fact. Namely, the sequence $n\alpha \mod 1$ is uniformly distributed. Since

$$\begin{aligned}\left|\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}e^{2\pi ihn\alpha}\right| &= \left|\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}e^{(2\pi ih\alpha)^n}\right|\\ &= \left|\lim_{N\to\infty}\frac{1}{N}\frac{1-e^{2\pi ihN\alpha}}{1-e^{2\pi ih\alpha}}\right|\\ &\leq \lim_{N\to\infty}\frac{1}{N}\frac{2}{|1-e^{2\pi ih\alpha}|}=0\end{aligned}$$

because $1\neq e^{2\pi ih\alpha}$ for $0\neq h\in\mathbb{Z}$ ($\alpha\notin\mathbb{Q}$). Now Weyl's theorem A.17 implies the uniform distribution.

Since any uniformly distributed sequence is dense, the Theorem is proved. □

REMARK A.13. If a sequence is not dense, it misses an interval I. We can find a continuous function which is identically zero outside the interval but positive inside. Then the sum in Definition A.25 is identically zero while the integral is positive. Therefore, the sequence is not uniformly distributed.

A.10. The general Birkhoff Ergodic Theorem

In this section we are going to prove Birkhoff's Ergodic Theorem in full generality. We present a combinatorial–functional theoretic proof. We believe that this kind of proof gives a recipe for other related theorems.

Theorem A.19 (General Birkhoff Ergodic Theorem). [**KH96**, 4.2.1 p.136][**Sar20**, p.37] *Let $T\colon X \to X$ be a measure preserving map of a measurable space X with probability measure μ. Let $\phi \in L^1(\mu)$. Then the following limit exists μ-a.e.:*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \phi(T^n x) \equiv \hat{\phi}(x)$$

Before proving this theorem we will show how Theorem 9.2 follows from this theorem. First we observe that

$$\begin{aligned}\int_X \left(\frac{1}{N} \sum_{n=0}^{N-1} \phi(T^n x) \right) d\mu(x) &= \frac{1}{N} \sum_{n=0}^{N-1} \int_X \phi(T^n x)\, d\mu(x) \\ &= \int_X \phi(x)\, d\mu(x)\end{aligned}$$

since μ is invariant. Therefore

$$\int_X \hat{\phi}(x)\, d\mu(x) = \int_X \phi(x)\, d\mu(x). \tag{A.2}$$

The space of continuous functions is separable, i.e., there is a dense sequence $\{\phi_n\}$ of continuous functions. Theorem A.19 states that for each $n \in \mathbb{N}$ we can choose a set A_n of measure $\mu(A_n) = 1$ such that $\hat{\phi}_n(x)$ exists for all $x \in A_n$. Let $A = \cap_{n\in\mathbb{N}} A_n$. Then $\mu(A) = 1$ and for all $x \in A$ the limit $\hat{\phi}_n(x)$ exists for all $n \in \mathbb{N}$. Since the the averaging operator is continuous, i.e.,

$$\|\phi - \psi\| < \epsilon \implies \left\| \frac{1}{N} \sum_{n=0}^{N-1} \phi(T^n x) - \frac{1}{N} \sum_{n=0}^{N-1} \psi(T^n x) \right\| < \epsilon$$

we have that $\hat{\phi}(x)$ exists for all $x \in A$ and all continuous functions ϕ.

Moreover, for $x \in A$

$$\begin{aligned}\hat{\phi}(x) &= \lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} \phi(T^n x)\\ &= \lim_{N\to\infty} \frac{1}{N}\left(\phi(x) - \phi(T^N x) + \sum_{n=0}^{N-1} \phi(T^{n+1}x)\right)\\ &= \lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} \phi(T^n(Tx))\\ &= \hat{\phi}(Tx)\end{aligned}$$

showing that $\hat{\phi}$ is an invariant function. By ergodicity, it must be constant μ–a.e. By equation (A.2) this constant must be equal to the integral of ϕ. Hence, Theorem 9.2 is proved. Now we return to the proof of the present Theorem.

Proof of the general Birkhoff Ergodic Theorem. We are going to prove the general Birkhoff Ergodic Theorem in the case that $\phi \in L^2(\mu)$. The L^1–case can be obtained by approximating an L^1 function by bounded functions (truncations). The advantage of L^2 is that we can use Hilbert space techniques.

The idea of the proof is the following. First we provide a class of functions were the theorem holds (more or less) trivially. Then we prove that this set of functions is dense. Finally we use Theorem A.20 (to be stated and proven next) to prove that the averaging operator is continuous and hence the theorem holds for all functions in L^2.

Let $\mathcal{I} := \{\phi \in L^2 : \phi(x) = \phi(Tx)\ \mu\text{–a.e.}\}$ be the space of invariant L^2–functions. Then for $\phi_1 \in \mathcal{I}$ and μ–a.e. $x \in X$ we have

$$\left\| \frac{1}{N}\sum_{n=0}^{N-1} \phi_1(T^n x) - \phi_1(x)\right\| = \left\| \frac{1}{N} N \cdot \phi_1(x) - \phi_1(x)\right\| = 0.$$

Let $\mathcal{C}$ be the space of functions θ which have a representation

$$\theta(x) = h(Tx) - h(x) + c$$

where $h \in L^2 \cap L^\infty$ is a bounded square–integrable function and $c \in \mathbb{R}$. Such functions are called **Co-boundaries**. Note that $\mathcal{C} \subset L^2$ since $\mu(X) = 1$. For $\phi_2 \in \mathcal{C}$ we have

$$\begin{aligned}\left\| \frac{1}{N}\sum_{n=0}^{N-1} \phi_2(T^n x) - c\right\| &= \left\| \frac{1}{N}\sum_{n=0}^{N-1} \big(h(T^{n+1}x) - h(T^n x) + c\big) - c\right\|\\ &= \left\| \frac{1}{N}\big(h(T^N x) - h(x) + Nc\big) - c\right\|\\ &\le \frac{2\sup_{x\in X}|h(x)|}{N}\end{aligned}$$

Moreover, using the invariance of μ we verify that

$$\begin{aligned}\int_X \phi_2 \, d\mu &= \int_X (h(Tx) - h(x) + c) \ d\mu \\ &= \int_X h(Tx) \, d\mu - \int_X h(x) \, d\mu + \int_X c \, d\mu \\ &= \int_X h(x) \, d\mu - \int_X h(x) \, d\mu + c = c.\end{aligned}$$

Hence, the following result holds for $\phi_2 \in \mathcal{C}$:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \phi_2(T^n x) = c = \int_X \phi_2 \, d\mu.$$

Therefore, if $\phi \in \mathcal{D} := \mathcal{I} + \mathcal{C}$ (i.e., $\phi = \phi_1 + \phi_2$ is the sum of an $\mathcal{I}$–function and a $\mathcal{C}$–function) then the Birkhoff average exists μ-a.e., since the averaging operator is linear and we can "split" its action over both terms in the sum.

We are now going to show that the set of functions $\mathcal{D}$ is dense in L^2. For this we prove first that the orthogonal complement of $\mathcal{C}$ consists of invariant functions, i.e., $\mathcal{C}^\perp \subset \mathcal{I}$. Let $\psi \in \mathcal{C}^\perp$, i.e.,

$$< \psi, \phi >:= \int_X \psi(x) \cdot \phi(x) \, d\mu = 0$$

for **all** $\phi \in \mathcal{C}$. In particular

$$< \phi, \phi - \phi \circ T >= 0.$$

Then

$$\begin{aligned}\|\phi - \phi \circ T\|_{L^2}^2 &=< \phi - \phi \circ T, \phi - \circ T >= \|\phi\|_{L^2}^2 - 2 < \phi, \phi \circ T > \\ &= 2\|\phi\|_{L^2}^2 - 2 < \phi, \phi - (\phi - \phi \circ T) > \\ &= 2\|\phi\|_{L^2}^2 - 2\|\phi\|_{L^2}^2 \\ &= 0.\end{aligned}$$

This implies that $\phi = \phi \circ T$ for μ-a.e. $x \in X$.

It is now left to prove that the averaging operator is continuous. Since this operator is linear, we have to check this only at $\{0\}$.

Let $A_c(\phi) := \{x \in X \, : \, \sup_{N\in\mathbb{N}} \sup_{1\le n\le N} \frac{1}{n} \sum_{k=0}^{n-1} |\phi(T^k x)| > c\}$, for $\phi \in L^1$. By the Maximal Ergodic Theorem A.20 we have that

$$\|\phi\|_{L^1} = \int_X |\phi| \, d\mu \ge \int_{A_c} |\phi| \, d\mu \ge c\mu(A_c(\phi))$$

or

$$\mu\{x \in X \, : \, \sup_{N\in\mathbb{N}} \sup_{1\le n\le N} \frac{1}{n} \sum_{k=0}^{n-1} |\phi(T^k x)| > c\} \le \frac{1}{c} \|\phi\|_{L^1}.$$

Now Fix $c \in \mathbb{R}^+$, $\epsilon > 0$ and let $\phi_m \in \mathcal{D}$ and $\|\phi - \phi_m\|_{L^1} \to 0$. We denote $\hat{\phi}_m(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \phi_m(T^k x)$ for μ–a.e. $x \in X$. Let $m_c \in \mathbb{N}$ be such that $\|\phi - \phi_m\|_{L^1} < \frac{c\epsilon}{2}$ for all $m > m_c$. Then by the previous inequality we get for $m > m_c$

$$\begin{aligned}
&\mu\left\{x \in X \,:\, \limsup_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1}\phi(T^k x) - \liminf_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\phi(T^k x) > c\right\}\\
&\le \left\{x \in X \,:\, \left|\limsup_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\phi(T^k x) - \hat{\phi}_m(x)\right|\right.\\
&\qquad\qquad \left. + \left|\liminf_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\phi(T^k x) - \hat{\phi}_m(x)\right| > c\right\}\\
&\le \left\{x \in X \,:\, 2\left|\limsup_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\phi(T^k x) - \widehat{\phi_m}(x)\right| > c\right\}\\
&\le \mu\left\{x \in X \,:\, \sup_{N\in\mathbb{N}}\sup_{1\le n\le N}\frac{1}{n}\sum_{k=0}^{n-1}|\phi(T^k x) - \phi_m(T^k x)| > \frac{c}{2}\right\}\\
&\le \frac{2}{c}\|\phi - \phi_m\|_{L^1} < \epsilon
\end{aligned}$$

Since $c, \epsilon > 0$ were arbitrary we conclude that

$$\mu\left\{x \in X \,:\, \limsup_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\phi(T^k x) - \liminf_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\phi(T^k x) > 0\right\} = 0$$

and the theorem is proven. □

A.10.1. The Maximal Ergodic Theorem.

THEOREM A.20. [**KP06**] *Let $c \in \mathbb{R}$ and μ be an invariant measure (invariant for the dynamics given by T).*
Let $A_c(\phi) := \{x \in X \,:\, \sup_{N\in\mathbb{N}}\sup_{1\le n\le N}\frac{1}{n}\sum_{k=0}^{n-1}\phi(T^k x) > c\}$. Then for any $\phi \in L^1(\mu)$ we have

$$\int_{A_c(\phi)} (\phi(x) - c)\, d\mu \ge 0.$$

PROOF OF THEOREM A.20. (After K. Petersen[**KP06**]) Let us denote

$$\phi^N(x) := \sup_{1\le n\le N}\frac{1}{n}\sum_{k=0}^{n-1}\phi(T^k x)$$

and assume that $\phi \in L^1 \cap L^\infty$. The general case is a question of a standard approximation procedure. For $c \in \mathbb{R}$ fixed we set, for $N \in \mathbb{N}$

$$E_N := \left\{x \in X \,:\, \phi^N(x) > c\right\}.$$

Then $(\phi(x) - c)\chi_{E_N}(x) \ge \phi(x) - c$, since both sides are equal if $x \in E_N$, while for $x \notin E_N$ we have that $\phi(x) - c \le 0 = (\phi(x) - c)\chi_{E_N}(x)$.

Let us split the sum $\sum_{k=0}^{m-1}\left[(\phi(T^k x) - c)\chi_{E_N}(T^k x)\right]$ into several parts for $m - 1 > N$ (actually, m will be much larger than N):

- The first part consists of zeroes as long as $T^i x \notin E_N$.
- Then there is a first k_1 where $T^{k_1}x \in E_N$. By the definition of E_N we have to wait at most $l_1 \leq N$ steps until
$$\sum_{k=0}^{l_1} \left[(\phi(T^{k_1+k}x) - c)\chi_{E_N}(T^{k_1+k}x)\right] = a_1 > 0.$$
This gives a second string of length l_1. Set $r_1 = k_1 + l_1$.
- There might be again some $n > r_1$ where $T^n x \notin E_N$. Let us denote with $k_2 > r_1$ the first step where $T^{k_2}x \in E_N$ and then repeat the previous splitting.
- Assume we have done this procedure n times, i.e., we have defined the strings from k_i till r_i; $i \leq n$ with $r_i - k_i = l_i \leq N$, such that
$$\sum_{k=0}^{l_i} \left[(\phi(T^{k_i+k}x) - c)\chi_{E_N}(T^{k_i+k}x)\right] = a_i > 0 \text{ and}$$
$$\sum_{k=r_{i-1}+1}^{k_i-1} \left[(\phi(T^k x) - c)\chi_{E_N}(T^k x)\right] = 0.$$
Then we can continue by induction until we reach $m-1$. Say we stopped at r_{last}. The rest of the sum, i.e., the sum from $r_{last}+1$ till $m-1$ can have at most $q < N$ terms in χ_{E_N} (otherwise r_{last} was not the last one) and the worst scenario is that those q terms are negative.

We can estimate:

$$\begin{aligned}
&\sum_{k=0}^{m-1} \left[(\phi(T^k x) - c)\chi_{E_N}(T^k x)\right] = \\
&\sum_{k=0}^{k_1-1} \left[(\phi(T^k x) - c)\chi_{E_N}(T^k x)\right] + \sum_{k=k_1}^{r_1} \left[(\phi(T^k x) - c)\chi_{E_N}(T^k x)\right] + \\
&\sum_{k=r_1+1}^{k_2-1} \left[(\phi(T^k x) - c)\chi_{E_N}(T^k x)\right] + \sum_{k=k_2}^{r_2} \left[(\phi(T^k x) - c)\chi_{E_N}(T^k x)\right] + \cdots \\
&+ \sum_{k=1+r_{last}}^{m-q-1} \left[(\phi(T^k x) - c)\chi_{E_N}(T^k x)\right] + \sum_{k=m-q}^{m-1} \left[(\phi(T^k x) - c)\chi_{E_N}(T^k x)\right] \\
&= 0 + a_1 + 0 + a_2 + \cdots + 0 + \sum_{k=m-q}^{m-1} \left[(\phi(T^k x) - c)\chi_{E_N}(T^k x)\right] \\
&\geq N\left(-\|\phi\|_{L^\infty} - c\right).
\end{aligned}$$

By invariance of T we have that

$$\int_{E_N} \sum_{k=0}^{m-1} \left[(\phi(T^k x) - c)\right] \, d\mu = m \int_{E_N} (\phi - c) \, d\mu$$

Now we integrate both sides of the estimation:

$$m \int_{E_N} (\phi - c)\, d\mu = m \int_X (\phi - c)\chi_{E_N}\, d\mu \geq N\left(-\|\phi\|_{L^\infty} - c\right).$$

Hence, dividing by m,

$$\int_{E_N} (\phi - c)\, d\mu \geq \frac{N}{m}\left(-\|\phi\|_{L^\infty} - c\right).$$

Letting $m \to \infty$ we get

$$\int_{E_N} (\phi - c)\, d\mu \geq 0.$$

E_N is not exactly the set $A_c(\phi)$ but letting $N \to \infty$ completes the proof. □

A.11. Ergodic Decomposition

In this section we will show that on compact metric spaces X any invariant measure can be decomposed into ergodic measures. This section relies on general results in Functional Analysis. Throughout this section we assume that X is a compact metric space. In this case, the space of continuous functions is separable and hence there is a sequence $(f_n)_{\mathbb{N}}$ of continuous functions that are dense in $B = \{f \text{ continuous } : \sup_X |f(x)| \leq 1\} \ni f_n$. We can define a distance between finite signed measures on X. For $\mu, \nu \in \mathcal{M}(X)$ we define

$$d(\mu, \nu) = \sum_{n=1}^{\infty} \frac{|\int_X f_n\, d\mu - \int_X f_n\, d\nu|}{2^n}.$$

This metric generates the weak* topology. $\mathcal{M}(X)$ is also separable (the measures giving rational values to the integrals for each f_n form a countable dense set). Moreover, the unit ball is convex and hence $\mathcal{M}(X)$ is locally convex (any two distinct measures can be separated by small convex neighbourhoods, e.g., disjoint balls of small diameter each containing exactly one of the two measures.) Then the Hahn-Banach Theorem[**Rud91**] (geometric version) implies that there are sufficiently many continuous affine functionals to separate measures, i.e., for $\mu \neq \nu$ there is a continuous affine function $f\colon X \to \mathbb{R}$ such that $f(\mu) \neq f(\nu)$. By separability, we can choose a countable collection of affine functions that separates measures.

First we study the space of invariant measures.

Lemma A.10. *The space $\mathcal{M}_{inv}(X)$ is compact and convex as a subset of the linear space $\mathcal{M}(X)$ of all signed finite Borel measures on X.*

Proof. We equip $\mathcal{M}(X)$ with the weak* topology, i.e., $\mu_n \to \mu$ if for any continuous $f\colon X \to \mathbb{R}$ we have $\int_X f\, d\mu_n \to \int_X f\, d\mu$, which is metrisable. $\mathcal{M}_{inv}(X)$ is a subset of the unit ball in $\mathcal{M}(X)$ which is compact by the theorem of Banach-Alaoglu [**Rud91**, 3.15, p.68]. Since any accumulation point of a sequence (μ_n), $\mu_n \in \mathcal{M}_{inv}(X)$ is in $\mathcal{M}_{inv}(X)$, the accumulation points of invariant probability measures are invariant probability measures:

$$\mu_n(A) \geq 0 \implies \mu(A) \geq 0,$$
$$\mu_n(X) = 1 \implies \mu(X) = 1$$

and

$$\mu_n(T^{-1}A) = \mu_n(A) \implies \mu(T^{-1}A) = \mu(A).$$

Moreover, for $0 \leq \lambda \leq 1$, $\mu_1, \mu_2 \in \mathcal{M}_{inv}(X)$

$$\lambda\mu_1 + (1-\lambda)\mu_2 \in \mathcal{M}_{inv}(X).$$

□

LEMMA A.11. *The set of extreme points of $\mathcal{M}_{inv}(X)$ are exactly the ergodic measures.*

PROOF. Assume that μ is not ergodic. Then there is a set $A = TA$ and $0 < \mu(A) < 1$. We define two distinct invariant (because A is invariant) probability measures μ_1, μ_2 by

$$\mu_1(B) = \frac{1}{\mu(A)}\mu(A \cap B)$$

and

$$\mu_2(B) = \frac{1}{\mu(X \setminus A)}\mu((X \setminus A) \cap B).$$

Then with $0 < \lambda = \mu(A) < 1$ we get for any measurable set B

$$\mu(B) = \lambda\mu_1(B) + (1-\lambda)\mu_2(B)$$

contradicting ergodicity. Hence, any extreme point is ergodic. For the reverse implication we restrict to the case when T is invertible. The general case is slightly more complicated. Let $\mu \in \mathcal{M}_{erg}(X)$ and $\nu << \mu$ an absolutely continuous probability measure, i.e. $\mu(A) = 0 \implies \nu(A) = 0$. Using Riesz Representation Theorem[**Rud91**] one can prove that there is a function $f \in L^1(X)$ such that $\nu(A) = \int_A f(x)\, d\mu(x)$ for all measurable A (Radon-Nikodym Theorem[**Rud91**]).

Now for $\nu \circ T^{-1} = \nu << \mu = \mu \circ T^{-1}$

$$\begin{aligned}\int_A d\nu \circ T^{-1} &= \int_X \mathbb{1}_A \, d\nu \circ T^{-1} = \int_X \mathbb{1}_A \circ T \, d\nu \\ &= \int_X \mathbb{1}_A \circ T \cdot f \, d\mu = \int_X \mathbb{1}_A \cdot f \circ T^{-1} \, d\mu \circ T^{-1} \\ &= \int_A f \circ T^{-1} \, d\mu \circ T^{-1}.\end{aligned}$$

This shows that the density f is constant μ-a.e.

Suppose that μ is not an extreme point, i.e. there are two different invariant probability measures μ_1 and μ_2 and a number $0 < \lambda < 1$ such that for any measurable A

$$\mu(A) = \lambda\mu_1(A) + (1-\lambda)\mu_2(A).$$

In particular this implies that $\mu_i << \mu$ and $d\mu_i = f_i d\mu$ with μ-a.e. invariant f_i. But $\mu_i(X) = 1 = \int_X f_i \, d\mu$ we have for $\mu \neq \mu_i$ that f_i cannot be constant. This contradicts ergodicity of μ. $\square$

In the following we will follow [**Phe01**] where a comprehensive study of Choquet theory is given.

Let h be a continuous affine function on $\mathcal{M}(X)$, i.e. for $0 \leq \lambda \leq 1$ we have $h(\lambda\mu + (1-\lambda)\nu) = \lambda h(\mu) + (1-\lambda)h(\nu)$. The set A of all those functions is a linear space that contains the constant functions and functions of the type $h(\mu) = f(\mu) + r$ where f is a linear function and $r \in \mathbb{R}$. Therefore A contains sufficiently many functions to separate distinct measures, i.e. for $\mu \neq \nu$ there is an $h \in A$ such that $h(\mu) \neq h(\nu)$.

For a bounded function $f\colon \mathcal{M}(X) \to \mathbb{R}$ we define

$$\overline{f}(\mu) := \inf\{h(\mu) \ : \ h \in A, \ h \geq f\}.$$

We leave the following properties to the reader.

i $\overline{f}$ is concave, bounded and upper semi-continuous

ii $f \leq \overline{f}$ and if f is concave and upper semi-continuous then $f = \overline{f}$.

iii If f, g are bounded then $\overline{f+g} \leq \overline{f} + \overline{g}$ and $\|\overline{f} - \overline{g}\| \leq \|f - g\|$

iv For $g \in A$ we have $\overline{f+g} = \overline{f} + g$

v For $r > 0$ we have $\overline{rf} = r\overline{f}$

Since $\mathcal{M}(X)$ is metrisable there is a dense set $(h_n)_n$ in A with $\|h_n\| = 1$ and hence, separates distinct measures. Let

$$f := \sum_{n=1}^{\infty} \frac{h_n^2}{2^n}.$$

This sum converges uniformly and its limit is a continuous function. If $\mu \neq \nu$ then there is an n such that $h_n(\mu) \neq h_n(\nu)$ and h_n^2 as a square of a non-constant affine function is strictly convex on $[\mu, \nu]$. Hence, f is strictly convex.

Let B denote the subspace $A + \mathcal{R}f$ of continuous functions.

By (iv) and (v) we get for $p(g) := \overline{g}(\mu_0)$ that $p(g+h) \leq p(g) + h(g)$ and $p(rg) = rp(g)$, $r > 0$.

We define a linear functional on B via $l(h + rf) := h(\mu_0) + r\overline{f}(\mu_0)$. If $r \geq 0$ then $l(h + rf) = \overline{h + rf}$ while for $r < 0$ we have $l(h + rf) = h+rf \geq h+r\overline{f}$ since in this case $h+rf$ is concave. Hence, l is dominated by p. The Hahn-Banach Theorem then gives a linear functional m on the space of continuous functions on $\mathcal{M}(X)$ that coincides with l on B

and is dominated by p. If $g \leq 0$ then $0 \geq \overline{g} \geq m(g)$. Therefore it is non-positive on non-positive functions and therefore continuous.

By the Riesz Representation Theorem[**KH96**, A.2.6 p.713] there exists a non-negative finite Borel measure $\mathbb{P}$ on $\mathcal{M}(X)$ with $\int_{\mathcal{M}(X)}(\int_X \nu(g)\, d\nu)\, d\mathbb{P}(\nu) = m(g)$.

Since $1 \in A$ we get $1 = m(1) = \int_{\mathcal{M}(X)} d\mathbb{P}$. So $\mathbb{P}$ is a probability measure.

Moreover,

$$\int_{\mathcal{M}(X)} f\, d\mathbb{P} = m(f) = \overline{f}(\mu_0).$$

Since $f \leq \overline{f}$ we see

$$\overline{f}(\mu_0) = \int_{\mathcal{M}(X)} f\, d\mathbb{P} \leq \int_{\mathcal{M}(X)} \overline{f}\, d\mathbb{P}.$$

If on the other hand $h \in A$ and $h \geq f$ then $h \geq \overline{f}$ and

$$h(\mu_0) = m(h) = \int_{\mathcal{M}(X)} h\, d\mathbb{P} \geq \int_{\mathcal{M}(X)} \overline{f}\, d\mathbb{P}.$$

This together with the definition of $\overline{f}$ implies that $\overline{f}(\mu_0) \geq \int_{\mathcal{M}(X)} \overline{f}\, d\mathbb{P}$. Hence,

$$\int_{\mathcal{M}(X)} \overline{f}\, d\mathbb{P} = \int_{\mathcal{M}(X)} f\, d\mathbb{P}$$

or

$$\mathbb{P}(\{\mu \,:\, f(\mu) \neq \overline{f}(\mu)\}) = 0.$$

Lets assume that μ is not ergodic, i.e. $\mu = \lambda\mu_1 + (1-\lambda)\mu_2$, $0 < \lambda < 1$, $\mu_1 \neq \mu_2$. Then by strict convexity of f and the concavity of $\overline{f}$

$$f(\mu) < \lambda f(\mu_1) + (1-\lambda) f(\mu_2) \leq \lambda\overline{f}(\mu_1) + (1-\lambda)\overline{f}(\mu_2) \leq \overline{f}(\mu)$$

and such a μ is not in the (Borel) support of $\mathbb{P}$. This means that $\mathbb{P}$ is concentrated on the set of ergodic measures.

For an alternative proof based purely on probability theory see [**Sar20**].

APPENDIX B

Notes and Answers to the Exercises

These notes are intended to be read *in articulo mortis.* Rather, you should try the exercises on your own, regardless of the opinions, approaches or suggestions written here. If you do not succeed at the first try, insist. When everything else is lost, come to these notes. They may help some of the readers. Answers to a few exercises are added at the end.

Notes

1.1 This is explicit from the text.

1.2 Having the flow explicitly, we can explicitly compute F by setting $t = 1$. Plot typical orbits in both systems.

2.1 Use that if $v = \dot{x}$, then $\dot{v} = \ddot{x}$.

2.2 (b) Rewrite $t = \beta\tau$ and find suitable β such that the original equation is independent of ω_0.

2.3 Standard exercise in the theory of matrices. Compute $exp(tA)$ and put $x(t) = exp(tA)x_0$. Note that A_2 is non-diagonalisable.

2.4 After changing variables to $z = x/t$ the equation is separable.

2.5 Change variables to $u = exp(x)$ and set $u(t) = \frac{\dot{a}(t)}{C+a(t)}$. The equation for $\dot{a}$ is separable.

2.6 Derive the equation with respect to x and eliminate C.

2.7 If $v(x) = 0$ then x is constant. This is a possible solution of the ODE. Other solutions $x(t)$ do not cross this solution, so they never come to an x such that $v(x) = 0$. Use then continuity.

2.8 Use that the basis vectors of the space are linearly independent and they all participate in the solution.

2.9 Prove continuity by the definition. Check the proof of Picard's Theorem for a hint.

2.10 Don't solve the equations. Just perform the Picard iterates.

2.11 Same as above, after recasting the problem as a first-order equation system.

3.1 Move around x and t until they land in different sides of the equality, then integrate.

3.2 Apply the transformation formally. You will find that certain a and b highly simplify the remaining calculations.

3.3 Same as above.

3.4 The solution is C^k in μ.

3.5 (a): For all times. (b) Yes, see Index theory in the Appendix.

3.6 The condition on the vector field allows to consider an open set as in the definition of attracting set lying inside the trapping region.

3.7 Note here that the ω-limit of *any* point is compact but the set $\mathfrak{O}$ (the union of all ω-limits) consists of three isolated points.

3.8 Use polar coordinates. Calculations are lengthy.

3.9 There is a fixed point at the origin and in addition a pair of fixed points appear for $r > 1$, their location depending on r.

3.10 We want a closed connected set encircling some ball containing the origin, which we know is a fixed point. Let us try with an ellipsoid. If the vector field points inwards on the surface of the ellipsoid, it will be a trapping region. The ellipsoid looks like $H(z, y, z) = Ax^2 + By^2 + Cz^2 = D$ for some positive A, B, C, D. We want the vector field to point inwards, which means that $\nabla H \cdot v < 0$ on the surface of the ellipsoid. Putting $C = B$ facilitates the computations. Verify that for $0 < r < 1$, *any* such ellipsoid as proposed (based on the origin) will be a trapping region by choosing e.g., $B = C = 1$, $A = 1/\sigma$.

3.11 Here we want a trapping region that is good for a large region in r, e.g., $0 < r < 35$ in order to encompass all types of interesting behaviour of the system (this is to be understood after Chapter 7). The approach with an ellipsoid with $B = C$ still holds, but now we need a much larger region, D will have to be comparatively large. Computations are largely simplified if the ellipsoid is not centered at the origin, but rather at some point $(0, 0, z_0)$, with $z_0 = r + A\sigma/B$. On the surface of the ellipsoid $H(z, y, z) = Ax^2 + By^2 + B(z - z_0)^2 = D$, one has to verify that the function $\frac{\nabla H \cdot v}{2B} = -y^2 - \frac{A\sigma}{B}x^2 - b(z - \frac{z_0}{2})^2 + b\frac{z_0^2}{4}$ is negative. Using the usual parameters of the Lorenz system, namely $\sigma = 10$ and $b = 8/3$, you may verify that there are a few critical points on the surface of the ellipsoid and that on all of them the vector field points inwards for $\frac{D}{B} > \frac{16}{15}z_0^2$. The critical points (following this approach) are: $(y = 0 = x, z = z_0 \pm \sqrt{\frac{D}{B}})$, $(y = 0, z = \frac{z_0}{2}(\frac{b-2\sigma}{b-\sigma}), Ax^2 = D - B(z - z_0)^2)$ and $(x = 0, z = \frac{z_0}{2}(\frac{b-2}{b-1}), y^2 = \frac{D}{B} - (z - z_0)^2)$.

4.1 Write the first few iterations and try to guess the general rule. x is multiplied by powers of a and b is summed up times powers of a. Finally, $x_k = a^k X + b(1 + a + \cdots + a^{k-1})$, $k > 0$.

4.2 Change variables to $y = \ln x$ and try as in the previous Exercise.

4.3 Rewrite the dynamics for z and choose γ such that it becomes linear.

4.4 Population growth requires $r^n > 1$ and hence $r > 1$. Population decrease correspondingly demands $r < 1$.

4.5 For point 4 it is necessary to check when f has Lipschitz constant smaller than unity. The derivative of f may help.

4.6 Rewrite the problem in polar coordinates. Verify that the dynamics for the polar angle is decoupled and take as Poincaré section the line $\phi = \phi_0$.

4.7 Try with polar coordinates again for the phase variables y and $p = y'$.

4.8 The x-coordinate is decoupled from the y-coordinate. Take arbitrary open sets U, V and points (x_0, y_0), $({x_0}', {y_0}')$ in their interiors. Consider k large enough so that the square region $S_U = [x_0, x_0 + 1/k] \times [y_0, y_0 + 1/k] \subset U$ and similarly for the corresponding square S_V but now with side-length $2/k$. Let α be irrational. The forward iterates of the first square will eventually stretch along the whole y-coordinate of the torus (before taking mod 1). Show that it is enough to iterate j times, with $j > k$. Take then $j_0 > k$ such that ${x_0}' < F^{j_0}(x_0) < {x_0}' + 1/k$ (this is possible because the projection of the orbit on the first coordinate is dense). Show now that $F^{j_0}(S_U) \cap S_V \neq \varnothing$. Further, the map cannot be mixing, since this property fails for the x-coordinate, as discussed in Example 3.3. If α is rational, then the x-coordinate is periodic (of period, say, p), and the p-power of the map is a rational shift in y, hence periodic. See [**Rob08**].

4.9 A similar Exercise is worked out in detail in Section 23.3 of [**HW79**, pp. 378, 4th. ed.].

4.10 This is a special case of a more general statement discussed by Akin and Carlson in [**AC12**].

4.11 Writing each number in decimal notation, f shifts the decimal point to the right. The preimages of 0 are all points with finite decimal expression.

4.12 Use the hint. In binary expression each number in $[0, 1)$ is written as $x = 0.a_1a_2\cdots$ where the $a_k = 0$ or 1. The map shifts the dot to the right.

4.13 Use the hint and rewrite the map as a map for θ. Writing each number in $[0, 1)$ in base 3, the map shifts the dot to the right, cf the previous Exercise.

5.1 A quadratic, positive definite function will do.

5.2 First rewrite the equations as a dynamical system. The usual trick of trying a quadratic, positive definite function of the coordinates does not work. For (a) try $f(x, y) = \alpha x^4 + \beta x^2 + \gamma y^2$. The second problem is more tricky, you have to add to the previous trial a term βxy.

5.3 Apply Lyapunov's second theorem, i.e., study the eigenvalues of the linearisation.

5.4 We have h invertible and differentiable such that $h(x(t, h^{-1}(y_0) = y(t, y_0)$ or $h(e^{tA}h^{-1}(y_0) = e^{tB}y_0$. Differentiate with respect to y_0 at $y_0 = 0$: $(dh/dx)|_{x=0} e^{tA} (dh^{-1}/dy)|_{y=0} = e^{tB}$. The h-derivatives are matrices which are inverse of each other. Hence, the matrices A, B are similar (they have the same Jordan normal form and eigenvalues).

5.5 This result is related to a theorem by Chetaev. The *Chetaev function* is analogous to the Lyapunov function for unstable fixed points. The function f is increasing along trajectories because of the positive derivative. An arbitrary neighbourhood of p will contain a point x_n of the sequence. But $f(x(t, x_n)) > 0$ and the trajectory cannot approach zero.

5.6 We have $v(p) = 0 = -\nabla f(p) \Leftrightarrow df = (p) = \sum_i (\partial f/\partial x_i) dx_i = 0$. Also $D(v(p)) = \partial^2 f/\partial x_i \partial x_j$ is a symmetric matrix and $d^2 f(p)$ is a bilinear form. Non-degenerate means no zero eigenvalues and hence the fixed point is hyperbolic.

5.7 This is a classical exercise in electrodynamics. Check a book on Classical Electromagnetic Theory. The electrostatic potential inside a circle of radius R is given by

$$V(r, \theta) = \frac{1}{2\pi} \int_0^{2\pi} V(R, \phi) \frac{R^2 - r^2}{R^2 + r^2 - 2Rr \cos(\phi - \theta)} d\phi.$$

It turns out that the integral equals unity for $V(R, \phi) = const.$

5.8 This is now a classical exercise in Dynamical Systems. Solutions for the first part can be found somewhere in the Internet. The period-3 is more involved. Solve it numerically using the suggestion and plotting the polynomial that should give the roots. For analytic computations, see [**Bec96**].

5.9 Same as above.

5.10 The systems can be solved in closed form by direct integration.

5.11 Apply Lyapunov's Definitions and Theorems.

6.1 Straightforward application of the Theorems in this Chapter.

6.2 Same as above. The dynamical equation is now one–dimensional.

6.3 Disregard the fact that one eigenvalue of the linearisation at the origin is larger than unity and focus on the shape and dynamics of the Centre Manifold.

6.4 Recast μ as a third coordinate with dynamics $\dot{\mu} = 0$. This system has a two-dimensional centre manifold. Compute the manifold up to second order and write the dynamics on the manifold. Verify that in a small μ-interval around zero, the system has two fixed points of opposite stability, that collide exactly at $\mu = 0$.

6.5 Same ideas as the previous one, now in maps. There is only one fixed point along with an invariant set of the form $x^2 + y^2 = C$, constant and $\dot{\phi} = 1$ (the polar angle).

7.1 Your first Hopf calculation? Check that it fits in the previous theory.

7.2 First compute the eigenvalues of the linearisation, then check for what value(s) of a the eigenvalues are $+1$, -1 or complex with modulus one. All three cases actually occur.

7.3 Now we need pure imaginary eigenvalues ($a = -1$) or having one of the eigenvalues equal to zero, which occurs for $a = -3, 0$.

8.1 The logistic map revisited. Computations are straightforward and some of them have been done in previous exercises. Use a computer algebra programme.

8.2 Same as above. Color-coding is not essential but might give a beautiful picture to hang on your office. This is the famous logistic bifurcation diagram that most books display. See `http://en.wikipedia.org/wiki/Logistic_map`.

8.3 More colorful computations that should be straightforward using a computer algebra programme and/or a numerical computations programme.

8.4 All symbolic sequences correspond to trajectories. Build a sequence with "half" the sequence of $W_s(p)$ and half of $W_u(q)$. Assume there exists a stable orbit. Since there exists a dense orbit, any neighbourhood of the assumed stable orbit will have points (of the dense orbit) wandering away from the neighbourhood, which is a contradiction.

8.5 Try to adapt the technique used in Example 8.1. The idea is to find the fixed points/periodic points of the return map and use the linearisation to analyse their stability properties.

8.6 Let S be the unit square. Periodic points satisfy $A^n(x, y) = (x, y) \mod 1$ or $(A^n - Id)(x, y) = (0, 0) \mod 1$. The number of such points as well as the number of preimages of a point in the unit square is equal to the number of times $(A^n - Id)S$ covers the unit square, which in turn equals the area of the image. This is given by $\det(A^n - Id) = (\lambda_1^n - 1)(\lambda_2^n - 1)$ where λ_i are the eigenvalues of A. There are many webpages with examples and analyses of this problem.

8.7 See Appendix A.2.1 first. You can "construct" homoclinic symbolic sequences matching together portions of $W_s(p)$ and $W_u(p)$.

9.1 Check the definition inspiring the exercise. If the product of two nonnegative numbers is zero, then one of them is zero.

9.2 Express the characteristic function of an interval with its Fourier series: $\chi_{(a,b)}(x) = \sum_k c_k e^{2\pi i k x}$. Then, $\chi_{(a,b)}(R_\alpha^n(x)) = \chi_{(a,b)}(x + n\alpha \mod 1) = \sum_k c_k \left(e^{2\pi i k n\alpha}\right) e^{2\pi i k x}$. Hence,

$$\lim_{N\to\infty} \frac{1}{N} \#\{0 \le n < N \,:\, R_\alpha^n(x) \in (a, b)\} =$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \sum_k c_k \left(e^{2\pi i k n\alpha}\right) e^{2\pi i k x} = c_0 = \int_{[0,1)} \chi_{(a,b)}(x)\, dx,$$

since characteristic functions take the values 1 or 0. The second equality follows making the geometric sum on n and taking the limit (only $k = 0$ survives the N-limit). The last equality is just the definition of c_0 in Fourier series. We can extend this result from characteristic functions to continuous functions as well, hence:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \phi(R_\alpha^n(x)) = \int_{[0,1)} \phi(x)\, dx,$$

so the only invariant measure is the Lebesgue measure (the only functional that fits in the RHS of the last equation). If $\alpha = p/q \in \mathbb{Q}$ then there are many invariant ergodic measures: Take any periodic orbit $A = \{x_1, \ldots, x_q\}$ and define a measure with values $1/q$ if $x_i \in A$ and zero otherwise. Recall the previous Exercise.

9.3 We have two *different* ergodic measures. Hence, there must be some continuous function such that

$$\int_X \phi(x)\, d\mu \neq \int_X \phi(x)\, d\nu.$$

We use the Ergodic Theorem again:

$$\begin{aligned} \int_X \phi(x)\, d\mu &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \phi(f^n(x)), x \in A_\mu \\ \int_X \phi(x)\, d\nu &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \phi(f^n(x)), x \in A_\nu \end{aligned}$$

A_μ is a set of unit μ-measure where the averages make sense. Since the measures are different, $A_\mu \cap A_\nu = \varnothing$ and the measures are singular.

9.4 Suppose μ is invariant. Then $\mu(f^{-1}(\operatorname{supp}\mu)) = \mu(\operatorname{supp}\mu) = 1$ and also $\mu(f^{-1}(\operatorname{supp}\mu) \cap \mu(\operatorname{supp}\mu)) = 1$. Suppose $\operatorname{supp}\mu$ is not invariant. Then there exists a point $x \in \operatorname{supp}\mu \backslash f^{-1}(\operatorname{supp}\mu)$ or $x \in f^{-1}(\operatorname{supp}\mu) \backslash \operatorname{supp}\mu$. Since $\operatorname{supp}\mu$ is closed, there must be a whole ball around x such that

$$B(x) \subset \operatorname{supp}\mu \backslash f^{-1}(\operatorname{supp}\mu) \cup f^{-1}(\operatorname{supp}\mu) \backslash \operatorname{supp}\mu.$$

This ball should have zero measure, otherwise it is not true that $\mu(f^{-1}(\operatorname{supp}\mu) \cap \mu(\operatorname{supp}\mu)) = 1$. But in such a case either $x \notin \operatorname{supp}\mu$ or $f(x) \notin \operatorname{supp}\mu$ which is a contradiction. Hence $\operatorname{supp}\mu$ is invariant.

9.5 Assume the measure of this set is $a > 0$. Then there exists a compact set $K \subset X \setminus \bigcup_{x\in X} \omega(x)$ such that $\mu(K) > a/2$. However, by Poincaré Recurrence Theorem, almost any point

$x \in K$ returns infinitely often to K and hence, since K is compact, $\omega(x) \cap K \neq \varnothing$, which is a contradiction.

A.1 The complement of the Cantor set consists of a disjoint union of open intervals. Its measure is then the sum of the individual measures. At step $n \geq 0$ we eliminate 2^n subintervals of width $1/3^{n+1}$ so the total measure of the eliminated intervals is $m = (1/3)\sum_{n\geq 0}(2/3)^n = 1$. Hence, the Cantor set has zero measure.

A.2 The measure of the remaining set is still zero. After a more involved calculation we recover $m = 1$ for the deleted intervals, for any $0 < p < 1/3$.

A.3 We still eliminate 2^n intervals of length $1/k^{n+1}$. The total area of the subtracted part is $m = 1/(k-2)$, positive but smaller than unity for $k > 3$.

A.4 For $\mathbb{R}$ notice that the function $\arctan x$ maps the real line onto a subset of itself. $\mathbb{Z}$ can be mapped to $\mathbb{N}$ through the function $f(0) = 0$ and $f(k \neq 0) = 2|k| + (1 - sg(k))/2$. Positive numbers map to even numbers and negative map to odd. The second proof requires writing the rationals in the form of an infinite matrix. The last statement is the famous *diagonal proof* of Cantor. Assume the set of reals in $[0, 1]$ to be equipotent to $\mathbb{N}$. Express the numbers in decimal form and write each element in the set in a list that by assumption is exhaustive. Consider the number having as k-th decimal figure $x_k^k + 1$ where x_k^k is the k-th decimal figure of the k-th element in the list. If $x_k^k = 9$ we take $x_k^k - 1$ instead. The number so constructed is not in the list, hence the list was not complete.

A.5 Write the elements in C in base-3 and map each 2 in the expansion to a 1. The image of the map consists of all possible strings of 0's and 1's, i.e., all real numbers in $[0, 1]$, now written in binary form.

A.6 If the hint is not enough, see [**USM95**].

A.7 To compute a fixed point it is enough to solve $x_{n+1} = x_n$. To verify that it is unique, we need to show that the map is a contraction for a suitable interval and then use Banach's Theorem.

A.8 Same as the previous Exercise.

Answers to Selected Exercises

The solution to some selected exercises is given here. One may better profit of this Section if read together with the Notes in the previous Section.

1.2 $F(x) = xe$ and $x_n = x_0 e^n$, $n \geq 0$.

2.2 (a) $\dot{x} = p$; $\dot{p} = -w_0^2 x$; $x(t) = x_0 \cos w_0 t + p_0 \sin w_0 t$. (b) Rescale as $x = y/w_0$ and $t = T/w_0$. (c) $\dot{x} = p$; $\dot{p} = -\beta p - w_0^2 x$. (d) $w_0^2 = k/m$.

2.3 (1) $x(t) = \dfrac{7}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix} e^{-t} - \dfrac{1}{2}\begin{pmatrix} 1 \\ -1 \end{pmatrix} e^{-3t}$.

(2) $x(t) = \begin{pmatrix} 3+4t \\ 4 \end{pmatrix} e^{t}$.

(3) $x(t) = 4\begin{pmatrix} 2 \\ 1 \end{pmatrix} e^{t} - 5\begin{pmatrix} 1 \\ 0 \end{pmatrix}$.

2.5 $u = e^x$ and $u(t) = \dfrac{e^{t^2/2}}{\int_0^t e^{s^2/2} ds}$.

2.6 (a) $y' = yx/(x^2 - y)$, (b) $2xyy' = 2x^3 + y^2$.

2.10 (a) $x_0(t) = 0$, $x_1(t) = t^2/2$, $x_2(t) = t^2/2 - t^5/20$.

3.1 (a) $x(t) = \dfrac{x_0}{1+t_0}(1+t)e^{-(t-t_0)}$, (b) $x(t) = \dfrac{x_0}{1 + x_0 \ln \frac{t^2-1}{t_0^2-1}}$.

3.2 Use $x = z - 2t$ and obtain $x(t) = 1 - 2t + Ce^t$.

3.3 $x = -t$ and $\omega = -1$.

3.4 $x(t) = -1/t + \mu(1 - 1/t^2) + \mu^2(8/(3t^2) + t/3 - 2/t - 1/t^3)$.

3.7 Three fixed points at $y = 0$, $x = -1, 0, 1$. Cartesian coordinate axes through the fixed points are also part of the invariant sets. A large circle enclosing all fixed points serves as a trapping region. The ω-limit set is not connected.

3.8 Polar coordinates. A fixed point at the origin and a periodic orbit at $r = \sqrt{a}$, which is the ω-limit of all points outside the origin.

3.9 The origin is a fixed point, always present. For $r > 1$ a pair of fixed points appear at $x = y = \pm\sqrt{b(r-1)}$, $z = r - 1$.

3.10 For $r < 1$, the ellipsoid $x^2/\sigma + (y^2 + z^2)/r = D^2$, $D \geq 0$ is a trapping region.

4.1 For $0 \neq a \neq 1$, $x_n = a^n x_0 + b\displaystyle\sum_{k=0}^{n-1} a^k$; for $a = 0$, $x_n = b$ $(n \geq 1)$; while for $a = 1$, $x_n = x_0 + nb$ $(n \geq 0)$.

4.2 $x_n = x_0^{\alpha^n} \beta^{\frac{\alpha^n - 1}{\alpha - 1}}$.

4.3 For $\gamma = b/(a-1)$, $x_n = \dfrac{a^n x_0}{1 + \gamma x_0 (a^n - 1)}$.

4.4 (1) $x_n = x_0 r^n$; (2) Exponential growth for $r > 1$; (3) Exponential decrease for $r < 1$; (4) Valid for small-sized populations that do not influence the environement.

4.5 (1) When population densities increase towards their maximum $(= 1)$, birth-rate decreases; (2) The model admits real-valued solutions, however, “true” populations can only take rational values with denominator K; (3) A fixed point at $x = 0$ and another at $x = 1 - 1/r$, whenever $r \geq 1$; (4) For $r < 1$.

4.6 Return map: $P(r) = \dfrac{re^{2\pi}}{\sqrt{1 + r^2(e^{4\pi} - 1)}}$, with a fixed point at $r = 0$.

4.11 Writing each x in decimal form, $f(x)$ shifts one step to the right in the decimal expression of x: $f(0.a_1a_2a_3\cdots) = (0.a_2a_3\cdots)$.

5.1 The function $H(x,y) = x^2 + 2y^2$ is a Lyapunov function.

5.2 (a) E.g., $H(x,p) = p^2 + \omega^2 x^2(1 + x^2/2)$.
(b) E.g., $H(x,p) = p^2 + \omega^2 x^2(1 + x^2/2) + 2x(p + x)$.

5.3 $-2 < a < -1$.

5.8 (a) For fixed points see Ex 4.5. The origin is stable for $\mu < 1$ and the other fixed point is stable for $1 < \mu < 3$.

5.10 In the first case the origin is "stable from one side" and unstable from the other. In the second case the origin is unstable.

5.11 (a) The origin is center, $x = \pi$ is saddle. (b) Origin is saddle, points in $(\pm 1, 0)$ are stable for $\epsilon > 0$, unstable for $\epsilon < 0$. (c) Origin saddle, fixed point in $(1, -1)$, unstable.

6.2 In the coordinates diagonalising the linear part, $h(u) = \begin{pmatrix} -3u^3/968 \\ 3u^2/8 \end{pmatrix}$ and $\dot{u} = u^3 + O(4)$.

6.3 $h_4(x) = \frac{2}{3}x^2 - \frac{4}{3}x^3 - \frac{8}{9}x^4$ and $x_{n+1} = -x_n - \frac{4}{9}x_n^4 + O(5)$.

7.2 Fixed point at the origin, Hopf bifurcation for $a = 5/3$, period doubling for $a = 5/2$ and saddle-node for $a = 9/4$.

7.3 Fixed point at the origin, Hopf bifurcation for $a = -1$, a bifurcation of the saddle-node family for $a = -3$.

8.1 See Ex. 4.5 and 5.8. Same holds for 8.2 and 8.3.

8.4 (a) If $(a_0, \cdots, a_m)$, $(b_0, \cdots, b_n)$ are the periodic strings of p, q, then $z = \cdots(b_0, \cdots, b_n).(a_0, \cdots, a_m)\cdots \in W_s(p) \cap W_u(q)$.

8.5 Following ideas in Example 8.1, the return map is
$$\begin{array}{rcl} x & = & \lambda_-^m(X + ax + b(y - Y)) \\ y - Y & = & \lambda_+^m(\mu + cx + e(y - Y)^2) \end{array} .$$

A.1 The eliminated intervals add up to measure one. The Cantor set is measurable, with measure zero.

A.2 Same as above.

A.3 The eliminated intervals have measure $A_E = 1/(k-2)$ $(k > 3)$, hence the complement has positive measure.

A.7 (a) $x = (1 + \sqrt{5})/2$, (b) $x = \sqrt{3}$.

A.8 $x = 1/2$.

Bibliography

[AAA$^+$97] Dmitrij V Anosov, S Kh Aranson, Vladimir Igorevic Arnold, IU Bronshtein, Yu S Il'yashenko, and VZ Grines, *Ordinary differential equations and smooth dynamical systems*, Springer-Verlag, 1997.

[AC12] E. Akin and J. D. Carlson, *Conceptions of topological transitivity*, Topology and its Applications **159** (2012), no. 12, 2815–2830.

[AR94] H. Anton and C. Rorres, *Elementary linear algebra, applications version*, John Wiley & Sons, New York, 1994, 7th. edition 1994.

[Arn73] V. I. Arnold, *Ordinary differential equations*, M.I.T., Cambridge, MA, 1973.

[Arn89] ______, *Mathematical methods of classical mechanics*, second ed., Springer, New York, 1989, See pages 203– and 271–285.

[Ban22] S. Banach, *Sur les oprations dans les ensembles abstraits et leur application aux quations intgrales*, Fundamenta Mathematicae **3** (1922), no. 2, 133–181.

[Bec96] J. Bechhoefer, *The birth of period 3 revisited*, Mathematics Magazine **69** (1996), no. 2, 115–118.

[Ben01] Ivar Bendixson, *Sur les courbes dfinies par des quations diffrentielles*, Acta Math. **24** (1901), 1–88.

[Bir31] George D. Birkhoff, *Proof of the ergodic theorem*, Proceedings of the National Academy of Sciences **17** (1931), no. 12, 656–660.

[Bol71] Ludwig Eduard Boltzmann, *Einige allgemeine sätze über wärmegleichgewicht*, K. Akad. der Wissensch., 1871, Reprinted in Wissenschaftliche Abhandlungen, vol. 1, 259-287.

[Car81] J. Carr, *Applications of centre manifold theory*, Springer. New York, Heidelberg, 1981.

[CL55] Earl A. Coddington and Norman Levinson, *Theory of ordinary differential equations*, McGraw-Hill, New York, 1955.

[Dev86] R. L. Devaney, *An introduction to chaotic dynamical systems*, Benjamin/Cummings, 1986.

[EG79] Murray Eisenberg and Robert Guy, *A proof of the hairy ball theorem*, The American Mathematical Monthly **86** (1979), no. 7, 571–574.

[Emm80] Michele Emmer, *Visual art and mathematics: The moebius band*, Leonardo **13** (1980), no. 2, 108–111.

[FS03] Ai-Hua Fan and Jrg Schmeling, *On fast birkhoff averaging*, Mathematical Proceedings of the Cambridge Philosophical Society **135** (2003), no. 3, 443467.

[Gan59] F. R. Gantmacher, *The theory of matrices, vol. 1*, American Mathematical Society Chelsea Publishing, Providence, 1959, reprinted 2000.

[GH86] J. Guckenheimer and P. J. Holmes, *Nonlinear oscillators, dynamical systems and bifurcations of vector fields*, Springer, New York, 1986, 1st. printing 1983.

[GL87] P. M. Gruber and C. G. Lekkerkerker, *Geometry of numbers*, North-Holland Mathematical Library, vol. 37, Elsevier, 1987, 2nd. edition.

[Gro59] David M Grobman, *Homeomorphism of systems of differential equations*, Doklady Akademii Nauk SSSR **128** (1959), no. 5, 880–881.

[Hal69] J. K. Hale, *Ordinary differential equations*, Wiley, New York, 1969.

[Har60] Philip Hartman, *A lemma in the theory of structural stability of differential equations*, Proceedings of the American Mathematical Society **11** (1960), no. 4, 610–620.

[HP69] M. Hirsch and C. Pugh, *Stable manifolds for hyperbolic sets*, Bulletin of the American Mathematical Society **75** (1969), 149–152.

[HW79] G. H. Hardy and E. M. Wright, *An introduction to the theory of numbers*, fifth ed., Oxford Science Publications, Oxford, 1979.

[Jew70] Robert I. Jewett, *The prevalence of uniquely ergodic systems*, J. Math-Mech. **19** (1970), 717–729.

[KH96] A. Katok and B. Hasselblatt, *Introduction to the modern theory of dynamical systems*, Cambridge University Press, 1996.

[KM99] A. Karlsson and G. Margulis, *A multiplicative ergodic theorem and non-positively curved spaces*, Comm Math Phys **208** (1999), 107123.

[KP06] M. Keane and K. Petersen, *Easy and nearly simultaneous proofs of the ergodic theorem and maximal ergodic theorem*, IMS Lecture Notes–Monograph Series **48** (2006), no. 6, 248–251.

[Kri72] Wolfgang Krieger, *On unique ergodicity*, Proceedings of the 6th. Berkeley Symposium, 1970 **1** (1972), 327–346.

[Kro84] L. Kronecker, *Nherungsweise ganzzahlige auflsung linearer gleichungen*, Monatsber. Knigl. Preu. Akad. Wiss. Berlin (1884), 1179–1193, 1271–1299.

[Lad73] N. N. Ladis, *Topological equivalence of linear flows*, Differential Equations **9** (1973), 938.

[Lax02] P. D. Lax, *Functional analysis*, Pure and Applied Mathematics, John Wiley & Sons, 2002.

[Lor63] N. Lorenz, *Deterministic non-periodic flow*, J. Atmospheric Science **20** (1963), 130–141.

[MSS73] N. Metropolis, M. L. Stein, and P. R. Stein, *On finite limit sets for transformations on the unit interval*, J. Combinatorial Theory **15** (1973), no. A, 25.

[MvS93] Welington De Melo and Sebastian van Strien, *One-dimensional dynamics*, Springer, 1993.

[Nei59] Y. I. Neimark, *On some cases of periodic motions depending on parameters*, Dokl. Akad. Nauk SSSR **129** (1959), 736–739, in Russian.

[Ose68] Valeriĭ Oseledets, *A multiplicative ergodic theorem. lyapunov characteristic numbers for dynamical systems*, Transactions of the Moscow Mathematical Society **19** (1968), 197221.

[Pea86] G. Peano, *Sull'integrabilit delle equazioni differenziali del primo ordine*, Atti Accad. Sci. Torino **21** (1886), 437–445.

[Pea89] Giuseppe Peano, *Arithmetices principia: Nova methodo exposita*, Fratres Bocca, 1889.

[Phe01] Robert R. Phelps, *Lectures on choquets theorem*, 2nd ed. ed., Lecture Notes in Mathematics, vol. 1757, Springer Verlag, Berlin, 2001.

[Rob08] C. Robinson, *What is a chaotic attractor*, Qualitative Theory of Dynamical Systems **7** (2008), 227–236.

[Rot88] J. J. Rotman, *An introduction to algebraic topology*, Springer, New York, 1988.

[Rud91] Walter Rudin, *Functional analysis*, 2nd ed. ed., McGraw-Hill, Boston, 1991.

[Rue79] D. Ruelle, *Ergodic theory of differentiable dynamical systems*, Inst. Hautes Études Sci. Publ. Math. **50** (1979), 2758.

[Sac65] R. J Sacker, *A new approach to the perturbation theory of invariant surfaces*, Communications on Pure and Applied Mathematics **18** (1965), no. 4, 717–732.

[Sar20] Omri Sarig, *Lecture notes on ergodic theory*, Tech. report, Weizmann Institute of Science, 2020, in preparation.

[Sma67] S. Smale, *Differentiable dynamical systems*, Bull. Am. Math. Soc. **73** (1967), 747.

[SNM96] H. Solari, M. Natiello, and B. Mindlin, *Nonlinear dynamics: A two-way trip from physics to math*, Institute of Physics, Bristol, 1996.

[Tay06] M. E. Taylor, *Measure theory and integration*, American Mathematical Society, 2006.

[Tes12] Gerald Teschl, *Ordinary differential equations and dynamical systems*, vol. 140, American Mathematical Society, 2012.

[Š73] A. N. Šošitaĭšvili, *Bifurcations of topological type of a vector field near a singular point*, Functional Anal. Applications **6** (1973), 97.

[USM95] K. Ueno, K. Shiga, and S. Morita, *A mathematical gift i, mathematical world, v. 19*, American Mathematical Society, Rhode-Island, 1995.

[Ver90] F. Verhulst, *Nonlinear differential equations and dynamical systems*, Springer, Berlin, 1990, 2nd. edition 1996.

[Wag85] Stan Wagon, *The banach-tarski paradox*, Cambridge University Press, 1985.

[Wey16] H. Weyl, *ber die gleichverteilung von zahlen mod. eins*, Math. Ann. **77** (1916), no. 3, 313–352, On the Distribution of Numbers Modulo One.

[Wig90] S Wiggins, *Introduction to applied nonlinear dynamical systems and chaos*, Springer, New York, 1990, 2nd. ed. 2003.

[Zer04] Ernst Zermelo, *Beweis, dass jede menge wohlgeordnet werden kann*, Matematische Annalen **59** (1904), no. 4, 514–516.

Index